

THE ROUTLEDGE COMPANION TO TRANSNATIONAL WEB ARCHIVE STUDIES

The Routledge Companion to Transnational Web Archive Studies explores the untapped potential of web archives for researching transnational digital history and communication. It covers cross-border, cross-collection, and cross-institutional examination of web archives on a global scale.

This comprehensive collaborative work, emerging from the WARCnet research network, presents an exploration of the ways web archive research can transcend technological and legal challenges to allow for new comparative, transnational studies of the web's pasts, and of global events. By combining interdisciplinary work and fostering collaboration between web archivists and researchers, the book provides readers with cutting-edge approaches to analyzing digital cultural heritage across countries. The book contains concrete examples on how to research national web domains through a transnational perspective; provides case studies with grounded explorations of the COVID-19 crisis as a distinctly transnational event captured by web archives; offers methodological considerations while unpacking techniques and skill sets for conducting transnational web archive research; and critically engages the politics and power dynamics inherent to web archives as institutionalized collections.

The Routledge Companion to Transnational Web Archive Studies is an essential read for graduate students and scholars from internet and media studies, cultural studies, history, and digital humanities. It will also appeal to web archiving practitioners, including librarians, web curators, and IT developers.

Susan Aasman is Professor in Digital Humanities at the Centre for Media and Journalism Studies at the University of Groningen (The Netherlands). Her expertise is in the field of media history with a focus on digital media and web historiography and the new emerging field of web archaeology. She is interested in private, common, and institutional digital archival practices and discourses.

Anat Ben-David is an associate professor of communication at the Open University of Israel. Her research focuses on internet histories, digital technologies, and the intersection of politics and knowledge. Her work in web archive studies critically examines how archival infrastructures and geopolitics shape the web's pasts and explores new methods for advancing critical web archive research.

Niels Brügger is Professor at Aarhus University, School of Communication and Culture. His research interests are web historiography, web archiving, and media theory. Within these fields, he has authored a number of publications, including *Web 25: Histories from the first 25 years of the World Wide Web* (Ed.; Peter Lang, 2017), and *The archived web: Doing history in the digital age* (MIT Press, 2018).

ROUTLEDGE COMPANIONS TO THE DIGITAL HUMANITIES

Series editors: Nirmala Menon, Isabel Galina Russell, Harold Short and Ray Siemens

The Routledge Companions to the Digital Humanities series will produce large, comprehensive edited volumes that provide a state-of-the-art overview and analysis of specialist areas in the Digital Humanities (DH).

Contributions from emerging scholars will be welcomed, as will those from the field's acknowledged leaders. The interdisciplinary nature of DH will be reflected in each volume and across the series as a whole. Addressing the many developments and challenges in DH – not only in research, but in pedagogy – the Companion volumes will discuss cutting-edge innovation, projects and best practice across a range of relevant topics. Aiming to be truly international in scope, the series will highlight the work of editors and authors from a diverse range of countries and cultures around the globe.

The Routledge Companions to the Digital Humanities series will appeal to the large and growing body of those engaged or interested in DH. The volumes will be of interest to academics, researchers, and postgraduate students, as well as professionals working in DH, whether in academic institutions or in galleries, libraries, archives and museums.

This series complements the well-established monograph series Digital Research in the Arts and Humanities: www.routledge.com/Digital-Research-in-the-Arts-and-Humanities/book-series/DRAH

Titles published in the series include:

THE ROUTLEDGE COMPANION TO LIBRARIES, ARCHIVES, AND THE DIGITAL HUMANITIES

Edited by Isabel Galina Russell and Glen Layne-Worthey

THE COMPANION TO DIGITAL HUMANITIES IN PRACTICE

Edited by Constance Crompton, Laura Estill, Richard Lane, and Ray Siemens

For more information about this series, please visit: Routledge Companions to the Digital Humanities – Book Series – Routledge & CRC Press

THE ROUTLEDGE COMPANION TO TRANSNATIONAL WEB ARCHIVE STUDIES

*Edited by Susan Aasman, Anat Ben-David, and
Niels Brügger*

 **Routledge**
Taylor & Francis Group
LONDON AND NEW YORK

Designed cover image: Courtesy of Anat Ben-David and Dall-e.

First published 2025
by Routledge
4 Park Square, Milton Park, Abingdon, Oxon OX14 4RN
and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2024 selection and editorial matter, Susan Aasman, Anat Ben-David, and Niels Brügger;
individual chapters, the contributors

The right of Susan Aasman, Anat Ben-David, and Niels Brügger to be identified as the authors
of the editorial material, and of the authors for their individual chapters, has been asserted in
accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised
in any form or by any electronic, mechanical, or other means, now known or
hereafter invented, including photocopying and recording, or in any information
storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks,
and are used only for identification and explanation without intent to infringe.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

A catalog record has been requested for this book

ISBN: 9781032497785 (hbk)

ISBN: 9781032505558 (pbk)

ISBN: 9781003398998 (ebk)

DOI: 10.4324/9781003398998

Typeset in Galliard
by Newgen Publishing UK

CONTENTS

<i>Series editors' introduction</i>	<i>ix</i>
<i>List of figures</i>	<i>xi</i>
<i>List of maps</i>	<i>xiii</i>
<i>List of tables</i>	<i>xiv</i>
<i>List of contributors</i>	<i>xv</i>
1 Introducing transnational web archive studies <i>Anat Ben-David, Susan Aasman, and Niels Brügger</i>	1
2 “History web,” “web history,” and “history of the web”: Three subfields and why (and why not) integrating them <i>Gabriele Balbi</i>	15
PART I	
Entire national web domains from a transnational perspective	27
3 Iconography in flux: A transnational exploration of the evolution of climate news imagery through the Wayback Machine <i>Anat Ben-David, Adam Amram, and Noa Turgeman</i>	29
4 Comparing the holdings of closed national web archives through summaries <i>Janne Nielsen, Yves Maurer, and Eld Zierau</i>	52

Contents

5	Exploring the evolution of .lu domain names through a transnational comparison: Similarities and differences between .lu and .dk <i>Carmen Noguera, Janne Nielsen, Yves Maurer, and Ben Els</i>	73
6	Comparing national web domains across national web archives: Methodological and practical challenges of doing transnational studies <i>Niels Brügger</i>	93
7	Conversation 1: Transnational	110
PART II		
	The COVID-19 crisis as a transnational event	119
8	Oral histories and scalable reading: Analysing born-digital collecting practices during the COVID-19 pandemic <i>Friedel Geeraert, Nicola Bingham, Helle Strandgaard Jensen, Jane Winters, and Frédéric Clavert</i>	121
9	Surveying the landscape of COVID-19 web collections in European GLAM institutions: An explorative analysis <i>Nicola Bingham, Karin de Wild, Friedel Geeraert, and Caroline Nyvang</i>	142
10	What can we learn from URLs?: Understanding the scope of COVID-19 web archive collections for transnational analyses <i>Friedel Geeraert, Karin de Wild, and Susan Aasman</i>	160
11	The challenges of searching for women in the COVID-19 web archive collections: Promises, achievements, and pitfalls <i>Valérie Schafer, Susan Aasman, Frédéric Clavert, Karin de Wild, and Joshgun Sirajzade</i>	177
12	Conversation 2: Events	192
PART III		
	Methods and skills in web archive studies	197
13	Information ecosystems through the lens of web archives <i>Matthew S. Weber</i>	199

Contents

14	History of virtual museums and web archives: Opportunities for rescaling research <i>Nadezhda Povroznik</i>	215
15	Exploring skills and training requirements for the web archiving community <i>Helena Byrne, Juan-José Boté-Vericad, and Sharon Healy</i>	231
16	Teaching web archiving in higher education: Best practices and future perspectives <i>Márton Németh</i>	246
17	Conversation 3: Communities	261
PART IV		
	Politics of web archives as collections	271
18	The trouble with community: Constructing, deconstructing and reconstructing transnational “community” micro-archives <i>Saskia Huc-Hepher and Xiao Ma</i>	273
19	An inclusive approach to web archiving: The case of the Middle East and North African websites in the IIPC Novel Coronavirus collection <i>Sophie Gebeil</i>	290
20	The many lives of WeChat: Curating histories of the web in museum environments <i>Simone Natale</i>	306
21	Participation, platforms, and cultural heritage: Web archiving challenges <i>Marta Severo</i>	316
22	Building an archive of historical web defacements <i>Michael Kurzmeier</i>	331
23	Conversation 4: Institutional challenges	352

Contents

PART V	
Institutional challenges	359
24 Screens in struggle: From archived web corpus to readable data for history research <i>Sophie Gebeil and Jérôme Thièvre</i>	361
25 Towards transnational research data management practices for web archives: Challenges and possibilities <i>Eld Zierau, Beatrice Cannelli, Sally Chambers, Olga Holownia, Sharon Healy, Ditte Laursen, Susan Aasman, and Ulrich Karstoft Have</i>	379
26 The importance of legal requirements for web archives studies in Belgian and French laws <i>Lise-Anne Denis</i>	396
27 Public policies, technological infrastructure, and uses of web archives by the Digital Humanities in Brazil <i>Moisés Rockembach</i>	409
28 Conversation 5: The future	422
<i>Glossary</i>	426
<i>Index</i>	430

SERIES EDITORS' INTRODUCTION

Nirmala Menon (Indian Institute of Technology, Indore)
Isabel Galina Russell (Universidad Nacional Autónoma de México)
Harold Short (Emeritus, King's College London)
Ray Siemens (University of Victoria, Canada)

Digital Humanities (hereafter DH) is a rapidly growing field of academic activity, and the breadth of the field arises from its concern with the application of advanced technical/digital methods in research and pedagogy across all arts and humanities (and indeed many social science) disciplines. Growth in breadth and depth of the field has outpaced the availability of foundational resources to serve those involved and those training to be involved in its many activities — especially, but not exclusively, in areas seeing most active expansion, such as, for example: the practice and praxis of DH; the relation of DH to traditional and newer academic disciplines and its increasingly important role within them; its relation to cross-disciplinary groups such as those in the GLAM sector (galleries, libraries, archives, and museums); the exploration of important intersectional areas of diversity and indigeneity; and consideration of the impact of wide scale adoption of large language models and what is more popularly called artificial intelligence (AI).

Given the rapidity with which the field has been growing, there is a continuing need for state-of-the-art publications that address its many important challenges and developments, not only in research but also pedagogy, application, and further areas both within and outside academe. The common themes underpinning the series are the innovative methodologies used and developed in the application of digital approaches in the arts and humanities. Its intended scope, therefore, is to describe and discuss cutting-edge developments, projects, and best practice across the range of relevant topics.

The series is committed to promoting diversity and inclusivity. The Series Editors and the Editorial and Advisory Board are globally, culturally, and linguistically diverse, spanning a wide range of interests and expertise. Each volume will be edited by established specialists, whose brief will be to solicit contributions from authors from around the world. The intention is that editors and authors will be drawn from as diverse a range of expertise, stage of career, country, and culture as possible.

Series editors' introduction

The Companion series is intended to be complementary to the Digital Research in the Arts and Humanities series (DRAH), published by Ashgate/Routledge since 2008 (www.routledge.com/Digital-Research-in-the-Arts-and-Humanities/book-series/DRAH). Whereas the latter is a monograph series, with each volume having a relatively narrow focus, Companion series volumes will be more substantial, each aiming to provide comprehensive coverage across its specialist area.

The target audience for the two series is very much the same: the large and growing body of those engaged or interested in Digital Humanities. These include people working in higher education (librarians, students, researchers, lecturers, and administrators), in the GLAM sector (galleries, libraries, archives, museums), and in industry; also technical, research, and development professionals, and citizen scientists.

The language of publication will be English, but we are conscious that of necessity this restricts the extent to which its authorship can be truly global. The Series Editors will work with volume editors and authors, with the assistance of the Editorial and Advisory Board, to identify sources of funding for translation as and when possible, so that at least some authors can be included who cannot or choose not to write in English. In such cases, where it is possible for the original non-English chapter to be provided in camera-ready format, this will also be included in the publication.

FIGURES

3.1	Number of images per year	37
3.2	Number of images per source and per year	38
3.3	The distribution of the top 25 object labels per source (normalised proportions) (colour version in Zenodo repository)	40
3.4	Object labels (predictions by OpenAI CLIP) grouped into visual frames (colour version in Zenodo repository)	41
3.5	The distribution of visual frames per year (normalised proportions) (colour version in Zenodo repository)	42
3.6	The distribution of visual frames per source (normalised proportions) (colour version in Zenodo repository)	43
4.1	Overview of the web archives' holdings	61
4.2	Each ccTLD across web archives	62
4.3	Top 20 TLDs in the national web archives	65
4.4	Overlap analyses	67
4.5	Overlaps 2021	68
5.1	Number of .lu domains registered per year (1992–2022).	77
5.2	Number of disappeared .lu domains per year (1992–2022)	79
5.3	.lu domain name lifetimes (2000–2016)	80
5.4	LWA, IA, CC .lu domain names overlapping	83
5.5	Completeness of web archives for the .lu domain in Luxembourg (1996–2020)	84
5.6	Comparison of total registered domain names .dk and .lu (1992–2022, left y-axis .dk, right y-axis .lu)	85
5.7	Registered vs. disappeared domain names .dk	87
5.8	Registered vs. disappeared domain names .lu	87
8.1	Timeline of the start of COVID-19 collecting (colour version in Zenodo repository)	126
8.2	Themes in the oral history corpus, generated using Iramuteq (colour version in Zenodo repository)	130

List of figures

8.3	Visualisation of the relationship between topics in the corpus, generated using Iramuteq (colour version in Zenodo repository)	131
8.4	Visualisation of the structure and emphasis of the oral history interviews, generated using Iramuteq (colour version in Zenodo repository)	132
8.5	Representation of the corpus as a network of words, generated using Iramuteq (colour version in Zenodo repository)	133
10.1	(Inter)national scope of COVID-19 special collections	165
10.2	Total number of websites in the COVID-19 special collections of national web archives and the percentage of those websites having the same ccTLD as the archive's coverage area	167
11.1	Top 20 domains in the IIPC collection from one topic (by Joshgun Sirajzade)	185
13.1	Breakdown of news articles by critical information need	209
19.1	(a and b) The 10 most represented countries and languages in the Novel Coronavirus collection (colour version in Zenodo repository)	294
19.2	Distribution of archived Arabic-language content by country domains	297
19.3	(a and b) Comparison of actor categories between the entire IIPC collection (top) and the sites collected from the MENA zone (bottom) (colour version in Zenodo repository)	299
21.1	Tweet dedicated to the Georges Caron entry (colour version in Zenodo repository)	321
22.1	Distribution of archived material in the attrition archive per month	337
22.2	Directed network of link targets used in defacements, with a minimum of five incoming links, available on defacementarchive.com	338
22.3	Email address used by a defacer	339
22.4	Search results for a defacer's email address	341
22.5	Results list view with available export options	342
22.6	Distribution over time for keywords 'pakistan', 'kashmir', 'India'	343
22.7	Visualisation of affected domains per month with selected domains highlighted	344
22.8	Time difference in days between defacement and nearest capture	346
24.1	Pipeline: from corpus creation in web archives to readable data export	366
25.1	A visualisation of the Research Data Lifecycle from JISC's Research Data Management Toolkit	381

MAPS

10.1	(Inter)national scope of the COVID-19 Special Collection of the IIPC in collaboration with Archive-It	166
13.1	Cumulative media information ecosystem in New Jersey	208

TABLES

3.1	The sources selected for analysis	35
8.1	Overview of types of web content collected by the initiatives	125
8.2	Overview of the ways in which selection of content was opened up	134
9.1	Organisation type	144
9.2	Level of the organisation	145
9.3	Job profile	145
9.4	Start date of COVID-19 collections	147
9.5	Overview of relative number of institutions who have (not) stopped collecting content at time of survey	148
9.6	Tools used to collect web content	149
9.7	Long-term preservation	149
9.8	Most important challenges for the COVID-19 special collections	151
10.1	Overlap between the IIPC collaborative collection and national collections (based on the number of URLs)	168
10.2	Overlap between the IIPC collaborative collection and national collections (based on domain names)	169
10.3	Overlap between two national collections (based on domain name)	170
10.4	Overview of the presence of social media in the different COVID-19 collections (based on count of domain names)	171
21.1	Sources of information for studying institutional participatory platforms	322
21.2	Sources of information for studying amateur participatory platforms	325
22.1	Sample WARC header with reconstructed metadata	335

CONTRIBUTORS

Susan Aasman is Professor in Digital Humanities in the Faculty of Arts at the University of Groningen (The Netherlands). She is Director of the Centre for Digital Humanities and one of the founders of both the Master-track and Minor program Digital Humanities at the Faculty of Arts. Aasman's expertise is the field of media history with a focus on digital media and web historiography and the new emerging field of web archaeology. She is interested in private, common, and institutional digital archival practices and discourses.

Adam Amram is a PhD candidate in Software and Information Systems Engineering at Ben-Gurion University. His dissertation explores large-scale code maintainability in GitHub Python repositories. Amram holds an MA in Information and Knowledge Management from Haifa University and a BA in Psychology from the Open University of Israel. He co-authored several articles and book chapters with Anat Ben-David on the political aspects of web archive research using computational methods.

Gabriele Balbi is Full Professor in Media Studies at the Institute of Media and Journalism, USI Università della Svizzera italiana (Switzerland). His research interests are in the field of media and communication history. In relation to the web, he is the principal investigator of a research project on the history of the web in collaboration with CERN. His last book is *The digital revolution: A short history of an ideology* (Oxford University Press, 2023).

Anat Ben-David is an associate professor of communication in the Department of Sociology, Political Science, and Communication at the Open University of Israel. Focusing on the intersection between Science and Technology Studies (STS) and New Media, her research explores the interplay between digital platforms, politics, and knowledge. Methodologically, she specializes in developing and applying digital and computational methods for internet research. She is co-editor of the journal *Internet histories: Digital technology, culture and society* (Taylor & Francis/Routledge). She has published extensively in the Archive Studies to advance the field toward a critical understanding of how the web's pasts are constructed by and through the archived web.

List of contributors

Nicola Bingham is the Lead Curator of Web Archiving at the British Library in the Department of Collections. Following BA and MA degrees in History from the University of Newcastle upon Tyne, Nicola began her archival career at Tyne and Wear Archives Service before joining the British Library in 2002. Nicola manages the Library's web archiving strategy and the ongoing road map for web archiving capability, ensuring that stakeholders across the Legal Deposit Libraries, and other partners, have the necessary tools for curating websites according to their own collection development policies. Nicola is Co-Chair of the International Internet Preservation Consortium Content Development Group.

Juan-José Boté-Vericad is a lecturer professor at the Faculty of Information and Media Studies at the University of Barcelona. He holds a BA in Computer Science, MA in Digital Content Management, and PhD in digital preservation. His main research interests are in web archives and digital preservation focusing on the research data lifecycle.

Niels Brügger is Professor at Aarhus University, the School of Communication and Culture. In 2000 he co-founded the Centre for Internet Studies, Aarhus University. Since 2014, he is Head of NetLab, a research infrastructure for the study of the archived web. His research interests are web historiography, web archiving, and media theory. Within these fields he has authored a number of publications, among others *Web 25: Histories from the first 25 years of the World Wide Web* (Ed., Peter Lang, 2017), *The web as history: Using web archives to understand the past and the present* (Ed. with Ralph Schroeder, UCL Press, 2017), and *The archived web: Doing history in the digital age* (MIT Press, 2018). He is co-founder (2017) and Managing Editor of the international journal *Internet histories: Digital technology, culture and society* (Taylor & Francis/Routledge).

Helena Byrne is the Curator of Web Archives at the British Library. Helena completed a Master's in Library and Information Studies at University College Dublin, Ireland in 2015. Helena is also an independent researcher who focuses on the history of women's football in Ireland. Her previous publications cover both web archives and sports history.

Beatrice Cannelli is a PhD researcher at School of Advanced Study, University of London, funded by the London Arts and Humanities Partnership (LAHP). She is an archivist by training and has worked with traditional and born-digital archives both in Italy and UK and is also a member of the WARCnet research network. Her research interests include web and social media archiving practices, digital preservation, and internet history.

Sally Chambers is Digital Humanities Research Coordinator at the Ghent Centre for Digital Humanities and coordinator of DATA-KBR-BE project at KBR, Royal Library of Belgium to facilitate data-level access to KBR's digitized and born-digital collections for digital humanities research. Sally has participated in the PROMISE project to develop a federal strategy for the establishment of a web archive for Belgium, the BE-SOCIAL project related to social-media archiving in Belgium and is also a member of the WARCnet research network. In March 2022, she was appointed to the DARIAH-EU Board of Directors.

Frédéric Clavert is Assistant Professor in Contemporary European History at C2DH, University of Luxembourg. He previously worked as a researcher in Strasbourg, at the Centre Virtuel de la Connaissance sur l'Europe (CVCE) in Luxembourg and at the University of

List of contributors

Lausanne. Frédéric led the #ww1 project around the Centenary of the Great War on Twitter and is also interested in the collective memory aspects of the current public health crisis. With Caroline Muller (Université Rennes 2), he is co-editor of the online book *Le Goût de l'archive à l'ère numérique* (*The allure of the archive in the digital era*). He is also Managing Editor of the *Journal of digital history*.

Lise-Anne Denis is a law researcher from the University of Namur (Belgium), evolving within the Research Centre in Information, Law and Society (CRIDS) since 2021. She is currently working on her PhD as part of an ASP research mandate from the Belgian Fund for Scientific Research (FNRS). As part of her work at CRIDS, she has participated in the BE-SOCIAL project, managed by the Royal Library of Belgium (KBR) and aiming to set up a sustainable strategy for archiving and preserving social media in Belgium. Her participation in the project as a legal expert led her to study the legal implications of archiving activities, mainly in relation with the protection of personal data and privacy, as well as copyright.

Ben Els has been the digital curator of the Luxembourg Web Archive at the National Library of Luxembourg since 2017. He has previously worked in the cultural sector, as a project coordinator for the Mierscher Kulturhaus and the Séibühn Ënsber asbl. Ben completed his Bachelor's degree in European cultures at the University of Luxembourg, which he followed up with a Master's degree in comparative literary and art studies at the University of Potsdam.

Sophie Gebeil has been an associate professor of contemporary history at Aix-Marseille University since 2016. Her research focuses on the study of memorial practices and online social mobilizations, using web archives as a source for the historian. She holds the position of director for the Visual Studies and Digital Humanities in the Mediterranean research group in her laboratory (TELEMMé). Additionally, she is a member of the Institut Universitaire de France.

Friedel Geeraert is an expert in web archiving at KBR, the Royal Library of Belgium. She previously worked as a researcher on the PROMISE project that focused on the development of a Belgian web archive at the federal level. She is also involved in the BE-SOCIAL project focusing on social media archiving, is a member of the WARCnet research network and represents KBR at the IIPC (International Internet Preservation Consortium). She holds a Master's degree in History (KU Leuven) and in Information and Communication Science and Technology (Université Libre de Bruxelles).

Ulrich Karstoft Have is a research IT developer at Aarhus University, Denmark, where he assists students and staff in digital methods and how to do research using the online web and web archives. He has been responsible for data analytics in various researcher-led projects at a national level and has been a member of the Committee on Research Data Management by The Danish e-Infrastructure Cooperation (DeiC). He is a member of the steering group of the WARCnet research network.

Sharon Healy is an independent researcher and archivist with strong interests in internet/web history, Research Data Management workflows for digital data, and the preservation of digitized, born digital, and reborn digital heritage. Sharon holds a BA (Hons) in Cultural

List of contributors

Studies, an MA in Digital Humanities, a PG Diploma in Historical Archives, and a PhD in Digital Humanities.

Olga Holownia, PhD, is the IIPC Senior Program Officer, based at Council on Library and Information Resources. Olga manages the communications and program of the International Internet Preservation Consortium (IIPC). She has been responsible for research engagement and outreach activities and co-chairs the IIPC Research Working Group. Olga has served on the Board of the Digital Humanities in the Nordic and Baltic Countries since 2017.

Saskia Huc-Hepher (she/her/elle) is a reader at the University of Westminster, where she researches French diasporic identity, belonging and homemaking in online and offline contexts. She has contributed to theorizing selective web archiving and highlighting the role of language and translanguaging in the (re)born-digital space. She is curator of the London French Special Collection in the UK Web Archive, co-editor of *Digital modern languages (modern languages open)* and author of *French London: A blended ethnography of a migrant city* (Manchester University Press).

Helle Strandgaard Jensen is Associate Professor of Contemporary Cultural History in the Department of History and Classical Studies, Aarhus University, and co-director of the Center for Digital History Aarhus (CEDHAR). Her work focuses on the contemporary history of mass media, childhood, and digitization in Scandinavia, Western Europe, and the US. Helle is currently working on two projects: A monograph on the transnational history of *Sesame Street* in the 1970s (with Oxford University Press); and *Sesame Street at 50: The shapes of digital memory* (SHAPES), which uses a combination of computer-assisted analysis and close readings to investigate how social media platforms shape remembrance cultures.

Michael Kurzmeier holds the position of Research Infrastructure Office at Austrian Centre for Digital Humanities and Cultural Heritage. His research focused on political communication within web defacements. His postdoctoral work emphasized digital scholarly editions for born-digital data. In his current role, Michael is part of the research unit DH Research & Infrastructure. He is currently working for the EU-funded H2020 project EOSC Future building up the European Open Science Cloud (EOSC). He is also involved in the projects Repository Landscape and SSH Open Marketplace.

Ditte Laursen, PhD, is Head of the Digital Cultural Heritage Department at the Royal Danish Library, where she is responsible for acquisition of digitally born cultural heritage materials, long-term preservation of digital heritage collections and for access to digital cultural heritage collections. Her areas of expertise include collection management, IT governance, and research and development. Special interests include digital cultural heritage, digital humanities, and digital research infrastructures.

Xiao Ma (she/her) is a doctoral researcher (2020–2024) at the University of Westminster. Her research interests include “race” and space, Chinese diasporas, and migration-related heritage studies. She is the curator of the “Chinese in the UK” Special Collection and the “East and Southeast Asians in the UK” Special Collection in the UK Web Archive. Outside

List of contributors

academic research, Xiao works at China Exchange, a UK-registered charity, and she has extensive experience in facilitating community-building through heritage-making projects.

Yves Maurer is the Head of the IT and Digital Innovation Division at the National Library of Luxembourg and has been the technical lead on the Luxembourg Web Archive since 2016. He has an active role in all things digital happening at the library, from digital preservation, digital legal deposit, AI methods for enhancing usability of digitized materials, open data, transitioning to a new ILS, etc. Previously he was responsible for the BnL's digitization program from 2007 onwards and the setting up of the portal of Luxembourg newspapers at eluxemburgensia.lu. In that period, he was a member of professional boards relating to digitization at IFLA and Igelu. Previously, he was Vice-President of Development at Atril Language Engineering in Madrid and responsible for the flagship DéjàVu Computer-Assisted Translation software. He holds an Msci in Mathematics and Computer Science from Imperial College London.

Simone Natale is Associate Professor in Media Theory and History at the University of Turin, Italy, and an editor of the journal *Media, culture & society*. His latest monograph is *Deceitful media: Artificial intelligence and social life after the Turing test* (Oxford University Press, 2021), and he is also the author of numerous articles in journals including *New media and society*, *Communication theory*, the *Journal of communication*, and *Convergence*. His research has been funded by international institutions including the AHRC, ESRC, the Humboldt Foundation, and Columbia University's Italian Academy.

Márton Németh is a digital archivist at the Blinken OSA Archivum (Hungary). Obtained his PhD at the University of Debrecen in the Doctoral School of Informatics. The thesis focused on the evolution of the Hungarian Web Archiving Project at the National Széchényi Library and offered a general overview of web archiving for Hungarian scholars. His work focuses on digital preservation, creating workflows, and managing processes in this field. He has published for 18 years about digital libraries, library management, and web archiving.

Janne Nielsen is Associate Professor of Digital Media History at the Department of Media and Journalism Studies, School of Communication and Culture, Aarhus University (AU). She is on the board of the Centre for Internet Studies (AU), and a member of The Centre for Digital Methods and Media (AU) and the Center for Digital History Aarhus (AU). Her research interests include media history, web historiography, web archaeology, web archiving, web tracking, and issues related to privacy and consent online. Her recent research projects have focused on large-scale historical studies of a national web domain and mixed-method studies of web tracking.

Carmen Noguera is a doctoral researcher at the University of Luxembourg and is working on a PhD thesis on digital cultures and their development in Luxembourg (the late 1990s to 2014). Carmen has a degree in journalism from the University of Sevilla, Spain, followed by a DEA (post-graduate diploma of advanced studies) in EU and International Relations from the University Complutense Madrid. With over 15 years of experience in digital marketing and communications in the private sector, Carmen Noguera has focused on developing marketing and communications strategies, corporate e-reputation, digital footprint, Social Media strategies, and digital mentorship.

List of contributors

Caroline Nyvang is a senior researcher with the Royal Danish Library. She received her PhD on the history of food in 2013 and has since published widely on how knowledge and ideals of food has been dispersed. Currently, her work focuses on archived web as a resource to food historians.

Nadezhda Povroznik is a research associate at the Institute for History (at the Chair of Humanities Data Science and Methodology), at Technical University of Darmstadt (Germany), with more than 15 years of experience in the digital humanities field, with a focus on digital history, digital heritage, and virtual museology, the author of the project on *Digital history of virtual museums*.

Moisés Rockembach is Professor of Information Science at University of Coimbra, Faculty of Arts and Humanities (Portugal), research leader of the Research Group in Digital Preservation and Web Archiving — Federal University of Rio Grande do Sul (Brazil), member of the Technical Advisory Chamber for the preservation of websites and social media of the National Council of The Archives of Brazil (CONARQ). Researcher at CEIS20 — Center for Interdisciplinary Studies (University of Coimbra) and CITCEM — Transdisciplinary Research Center for Culture, Space and Memory — R&D Unit (University of Porto), a researcher in the InterPARES Trust AI project — Artificial Intelligence (University of British Columbia/Canada), and a visiting professor at KU Leuven — Digital Society Institute (DigiSoc/Belgium).

Marta Severo is a full professor in information and communication sciences at the University of Paris Nanterre and a junior member of the *Institut Universitaire de France*. She is a member and deputy director of the Dicen-IDF laboratory. Participatory and contributory online practices, representations of places through web data, social media analysis, and information visualization are among the research themes she is developing. Her research has obtained several forms of funding (in particular the ANR COLLABORA project on cultural participatory platforms). In 2019, she joined the scientific board of Open Edition.

Valérie Schafer has been a professor in Contemporary European History at the C²DH (Centre for Contemporary and Digital History) at the University of Luxembourg since February 2018. She previously worked at the CNRS in France and is still an associate researcher at the Center for Internet and Society. She specializes in the history of computing, telecommunications, and data networks. Her main research interests are the history of the Internet and the Web, the history of European digital cultures and infrastructures, and born-digital heritage (especially Web archives). She is on the editorial board of the journals *Le Temps des Médias*, *Journal of Digital History*, *IEEE Annals of the History of Computing* and is a co-founder of the journal *Internet Histories Digital Technology, Culture and Society* (Taylor & Francis).

Joshgun Sirajzade was a post-doctoral researcher in Parallel Computing and Optimisation Group in Computer and Communication Science at the University of Luxembourg at the time of the project. His research is mostly focused on artificial intelligence and its application. It includes machine and deep learning, natural language processing, information retrieval, and text data mining. He is the main developer of LuNa corpus tools (Luxembourgish National

List of contributors

Corpus), a tool for processing and mining of Luxembourgish language with modules for tokenizing, POS-tagging, topic modeling, and sentiment analysis.

Jérôme Thièvre, Web Archive Team Manager, INA (National Audiovisual Institute), France. Jérôme has a PhD in Computer Science in the field of Information Visualization. He joined the INA's Web Archive team in 2008 to design and implement the full text search features of web archives access. He was also in charge of media and Twitter captures for several years.

Noa Turgeman is a PhD candidate in the Science, Technology, and Society graduate program at Bar Ilan University. Her doctoral thesis explores visual representations of nature and climate on Instagram. Noa holds a Bachelor's degree in French Literature and Geography from the Hebrew University of Jerusalem and two Master's degrees in Comparative Literature and East Asia Studies from Tel Aviv University.

Matthew S. Weber is an associate professor in the Department of Communication, School of Communication and Information, Rutgers University. Matthew studies organizational change, news media, and institutional theory. Recent work includes a longitudinal project mapping ecosystems of local news in communities and examining the impact of news and information on local policymaking processes. Additional research includes an examination of the role of technology in news organizations, and the effectiveness of adopting social media within organizations to improve external communication. Matthew received his PhD in 2010 from the Annenberg School of Journalism and Communication at the University of Southern California.

Karin de Wild is Assistant Professor in Contemporary Museum and Collection Studies at Leiden University (The Netherlands) with a particular research interest for digital collections. Before joining Leiden University, she was a digital fellow at the University of Leicester. She has a past life as curator and collaborated with a wide range of museums from Tate and de Young Museum to SFMOMA.

Jane Winters is Professor of Digital Humanities at the School of Advanced Study, University of London. She has led or co-directed a range of digital projects, including the Congruence Engine and Heritage Connector projects, the UK-Ireland Digital Humanities Association network, Big UK Domain Data for the Arts and Humanities, and Traces through Time: Prosopography in Practice across Big Data. Jane's research interests include digital history, born-digital archives, the use of social media by cultural heritage institutions, and open access publishing. She has published most recently on web archives and researcher access, non-print legal deposit, and born-digital archives and the problem of search.

Eld Zierau is a senior researcher at the Royal Danish Library focused on digital cultural heritage and preservation of it. She holds an Msci in Mathematics and Computer Science from 1989 and she holds a PhD from 2011 within digital preservation. She has a strong interest and involvement in work on web archives as basis for future research. She has been conducting a study in detection and inclusion of national web outside the national top-level domain, and involved in a study on persistent referencing to web archives, resulting in the suggested Persistent Web Identifier.

1

INTRODUCING TRANSNATIONAL WEB ARCHIVE STUDIES

Anat Ben-David, Susan Aasman, and Niels Brügger

Since the mid-1990s, the World Wide Web (or simply: the web) has played an increasingly important role in many societies, and it constitutes one of the main pillars in today's communicative infrastructures. Due to the fragility of born-digital media, the online web is disappearing at an unprecedented pace. Fortunately, since the mid-1990s, web archives have been established and have preserved a treasure trove of more than 25 years of digital sources. Nevertheless, the valuable digital sources collected by institutions, from the US-based Internet Archive to many national web archives, remain untapped resources that have not yet impacted research within the humanities and the social sciences to the degree that one would have expected. In addition, most of the emerging literature related to the research use of web archives focuses on national settings, whereas genuinely transnational studies are lacking. This book, *The Routledge Handbook of Transnational Web Archive Studies*, aims to draw attention to the many archived web collections and rethink the challenges and possibilities they constitute for transnational studies of past and present web communication.

The book focuses on two interrelated themes — transnational studies and events, particularly the Coronavirus pandemic (COVID-19) — that are mutually conditioned: Studies of entire national web domains provide knowledge about the background for studies of an agenda-setting transnational societal event, COVID-19. Furthermore, to perform these studies, it is necessary to reflect on how such studies can be done, what politics inform and enable the study of web archive data across borders, and what institutional challenges such analyses must consider. These themes are covered extensively in this book.

What is web archiving?

If we want to understand the development of the web as a public sphere where national and transnational events were articulated, created, and negotiated, we need the web of the past in an archived form, and therefore, it is important to understand what web archiving is.

In the broadest sense, web archiving is “any form of deliberate and purposive collection and preservation of web material” (Brügger, 2018, p. 79). This broad definition can be elaborated on and detailed in several ways.

Firstly, collection and preservation are imperative to distinguishing web archiving from merely opening a web page on one’s computer, whereby web material is collected on the computer. Still, when leaving the web page in question, the web material is not preserved (in most cases). Secondly, web archiving must be deliberate and purposive, which means that the person or institution performing the collection and preservation must have some sort of idea that this is the case and that it is done with a specific purpose in mind, be that building a collection for a research project, or preserving the cultural heritage. Thirdly, when looking at the actual forms of performing web archiving, this can be done in several ways, including at least the following seven: (1) making an image, (2) making a screen movie, (3) downloading individual files, (4) web crawling, (5) collecting web material from a database through an application programming interface (API), (6) collecting the web that has been taken off-line and preserved unchanged, and (7) collecting the web as presented in other media types (books, film, and television) (cf. Brügger, 2018, pp. 80–84). Web crawling is the most widespread form of web archiving, especially among large web archiving institutions such as international and national web archives. It is also, by and large, the way web archiving is approached in the following chapters. In web crawling, a list of web addresses (URLs) is inserted into dedicated web crawler software:

then the software contacts the web servers indicated in the list of URLs with a request for specific web pages and elements. It then retrieves and stores the files and continues to the next URL on the list. The result of web crawling is a collection of HTML files and the other files that compose a given web page, either stored as they were retrieved or aggregated in dedicated archiving file formats, such as ARC or WARC.

(Brügger, 2018, pp. 81–82)

Finally, web archiving has to take into account which strategies have to be used to have the online web collected in such a way that it can be used for its purpose. Strategies could imply having as much of the online web archived as possible, getting web material from a specific period, getting specific types of content or files (hyperlinks, images, etc.), or combinations.

Thus, the actual performing of the deliberate and purposive collection and preservation of web material takes place at the nexus between, on the one hand, an awareness that one is archiving and why and, on the other hand, the use of one or more concrete forms of archiving, as well as specific strategies for achieving the goal of the collecting and preserving.

Transnational web archive studies

The title of this book conveys a dual meaning of the “transnational” in web archive research. This duality encompasses the study of the web as a transnational medium — hence, “transnational web archive” studies — and the cross-border, cross-collection, and cross-institutional examination of web archives, embodying “transnational studies” of the archived web. This thematic bifurcation aligns with the concurrent evolution of the web and globalisation, marking a period where the expansive utilisation of the web paralleled the emergence of transnationalism as a contemporary human condition.

Introducing transnational web archive studies

Such intensified globalisation processes have also led to the emergence of transnational studies in various social and humanistic disciplines. Researchers started challenging the nation-state as the standard analytical unit, calling for a decentralised approach sensitive to diaspora networks, cross-border flows, and global–local nexuses (Appadurai, 1996; Chalaby, 2005). Post-colonial critiques of Eurocentrism have challenged established nationalist historiographies (Stam & Shohat, 1994). Furthermore, in anthropology, globalisation research highlighted multi-sited ethnographies and concepts of cultural hybridity that challenged binary thinking (Burawoy et al., 2000).

The transnational turn has also significantly impacted media and communication research and media history. Theories such as global media events (Dayan & Katz, 1992) and comparative media systems (Hallin & Mancini, 2011) have propelled research that transcends methodological nationalism in communication research. Studies in media history have embraced transnational approaches to study the history of specific media, such as Television (Fickers & Johnson, 2010) and Cinema (Higbee & Lim, 2010), and highlighted overlooked histories of non-Western narratives, global media objects, and lives (Shome, 2006).

However, the notion of transnational studies and web histories may seem superfluous, given the internet's transnational character. Unlike television or cinema, the web emerged as an inherently transnational medium, only belatedly “tamed” or circumscribed by national borders. In its nascent 1990s form, national regulation struggled to contend with the internet's boundary-defying transnationality — a perceived challenge to state sovereignty that contributed to its celebratory framing as a borderless utopia. However, nation-states swiftly reasserted control within their territories through legislation and regulation (Goldsmith & Wu, 2006), countering such euphoric borderless narratives. From the mid-2000s onward, the web's infrastructural localisation was further reified through corporate personalisation and localisation algorithms, whose default disposition was to take you home (Rogers, 2013), anchoring user experience within national-cultural parameters.

The evolution of web archiving has paralleled the internet's development, originating with the Internet Archive as a transnational repository for websites irrespective of origin. However, the early 2000s witnessed nation-states reasserting their prerogative by establishing national web archives housed within various countries' national libraries and memory institutions (primarily in the Global North). This “national turn” in web archiving signified not merely repatriating the web surfer through localised defaults but also delineating histories of this ostensibly transnational medium along national boundaries. Such boundaries were often the product of technical definitions codified in law and policy, demarcating discrete national digital heritage corpora to be preserved while relegating the archiving of the web's broader transnational dimensions to external actors (Schafer & Winters, 2021). The proliferation of national institutions archiving web materials deemed constituent of their digital cultural heritage prompted the establishment of the International Internet Preservation Consortium (IIPC) in 2003. The IIPC has since evolved into a platform facilitating the transnational exchange of knowledge, technologies, standards, and best practices across member institutions and countries (Baucom, 2019). Its formation reflects the transitional recognition that while nation-states may stake claims over discrete online “territories,” the long-term preservation of the web's amalgamated transnational corpus necessitates collaborative frameworks transcending fragmented national efforts.

The formative years of web archive studies were accordingly shaped by the practice of collecting archived web materials along national boundaries, giving rise to a predominant focus on “national webs” and “national web histories” (Weltevrede & Helmond, 2012;

Ben-David, 2016; Brügger & Laursen, 2019; Ogden & Meamura, 2021). As computational tools for exploring web archives emerged, a new wave of scholarship began probing the characteristics and temporal development of “entire national web corpuses” (Ben-David & Huurdeman, 2014; Nielsen, 2019; Brügger et al., 2020). Consequently, the availability and delineation of archived web holdings within national library institutions significantly circumscribed the scope of empirical inquiry, determining the research questions and methodological approaches feasible in studying the archived web (Winters, 2019).

This volume provides a comprehensive overview of approaches and methods endeavouring to transcend the national boundaries that have circumscribed web archive studies, thereby enabling new modes of conceptualising the archived web from a transnational perspective consonant with the medium’s origins. These attempts may be broadly categorised into three overarching strategies. The first, cross-collection analysis, entails applying uniform methods and inquiries across archived web holdings from disparate institutional repositories, accentuating emergent similarities and divergences through comparative analytics. The second, a cross-cultural approach, examines differential representations and articulations of (national) cultural identities and worldviews manifested within a given archived web corpus, whether constituted as a national or transnational collection. The third avenue attends to the regulatory and policy frameworks that delimit the transnational mobility of data, knowledge, and methodological praxis pertaining to the archived web across national jurisdictions and archival repositories.

The WARCnet network

This book is one of the outcomes of the work undertaken in the international research network WARCnet, “Web ARChive studies network researching web domains and events,” funded by the Independent Research Fund Denmark | Humanities (grant no. 9055-00005B). The aim of the WARCnet network was to promote national and transnational research that could help us understand the history of (trans)national web domains and transnational events on the web, drawing on the increasingly important digital cultural heritage held in national web archives. Initially, WARCnet was planned to run in 2020–21, but due to the COVID-19 pandemic, the project period was extended several times, until the project ended in August 2023. The network activities took place in four (later six) dedicated Working Groups, at several network meetings in different European countries, and Short-Term Network Stays, among others. At the project start the WARCnet network comprised 24 humanities scholars, IT developers, and web archivists, including leading national web archives and the IIPC. During the project period, the number of participants was expanded to include 37 members, which allowed the network to pay particular attention to including the next generation of early career researchers. This diversity in network members was a pivotal element of WARCnet’s intention to promote close collaboration between researchers and web archivists, and this composition of the network is also mirrored in the following chapters.

In addition to the edited volume, other tangible results of the network activities are the two series of publications, WARCnet Papers and WARCnet Special Reports. Since the project’s start, WARCnet Papers have published numerous keynotes, interviews, round table discussions, presentations, status reports, etc. Since 2022, WARCnet Special Reports have offered a publication outlet for longer reports. The 29 published WARCnet Papers and the three WARCnet Special Reports all underpin the insights and results of many of the chapters in the present book.

Introducing transnational web archive studies

The ideas behind the WARCnet network, as well as its prehistory, are introduced in the welcoming publication *Welcome to WARCnet* (Brügger, 2020), and all WARCnet activities have been documented in annual reports (Brügger, 2021, 2022a, 2023a, 2023b). During the project, the core communicative outlet was the website (warcnet.eu). Upon the project's closure, the website has been preserved in the US-based Internet Archive (<https://archive.org/>) as well as in the Portuguese web archive Arquivo.pt (<https://arquivo.pt/>), where it is freely available, including all WARCnet Papers and WARCnet Special Reports. Also, the UK Web Archive at the British Library downloaded a copy of the website, which is available in the Web Archives and Digital Preservation subsection of the IT Collection (www.webarchive.org.uk/en/ukwa/collection/520). All embedded videos and presentations from the project's network meetings and the closing conference have been uploaded to Zenodo, under the community "WARCnet" (<https://zenodo.org/communities/warcnet>).

**Transnational web archive studies with the COVID-19
pandemic as a transnational event**

As explained, a few months after WARCnet started, the COVID-19 pandemic hit the world in early 2020 and soon became a transnational (crisis) event. The World Health Organization (WHO) declared COVID-19 a public health emergency on 30 January 2020, escalating it to a pandemic on 11 March 2020, with the emergency status ending on 5 May 2023. Web archives responded quickly as several European web archiving institutions started creating special COVID-19 collections, including a transnational "Novel Coronavirus Collection" curated by the IIPC. Many archivists involved in these initiatives were also part of the WARCnet community. These archival initiatives allowed the network to transform the challenge related to "the shadow of COVID-19" (Brügger, 2021) into an opportunity to explore what archiving an unforeseen event crisis entails and, in addition, how to study such an archived transnational crisis event. This exploration became a substantial part of this book (see Part II). However, it also resonates in many other chapters as it questions record-keeping procedures, legal and political challenges, research data management practices, or working with (trans)national datasets. Such explorations fit the observations made by scholars such as Jessica Ogden, who explored the socio-technical and cultural dimensions of web archiving, including archival logics of "crisis collecting" (Ogden, 2022, 2023). She emphasises that "practices are fundamentally shaped by an assemblage of sociocultural factors and technical actors, as well as who is doing the archiving" (Ogden, 2023, p. 6). Although her focus is very much on web archival interventions with dying platforms, it is also possible to take this approach to think about these logics with web archival challenges of transnational events.

As mentioned, the idea of mediating historical events has been theorised and analysed as a specific genre of mass communication by Dayan and Katz (1992) in their ground-breaking work on global media events. They described media events as live broadcasts of historic events (contests, conquests, and coronations). As pre-planned and scripted "high holidays of mass communication," these events interrupt the routine of regular media broadcasting (Dayan & Katz, 1992, p. 1). In their view, media events were embedded in the emerging transnational media infrastructures of the late 20th century, still dominated by radio and television. Since then, much has been written about media events, including some revisions to accommodate the concept of unforeseen crisis events, such as war and disasters (Hepp & Couldry, 2010). Only more recently, the impact of digital media — and especially the web — on the continuation of media events has been considered (Brügger, 2022b).

With the web as a dominant medium in times of transnational (crisis) events in mind, it has become essential to consider the role of web archives in capturing and documenting such interruptive, or — in some cases — disruptive historical moments, whether they are foreseen (such as recurring sports events like the Olympics), unforeseen (natural disasters, terrorist attacks, pandemics), or partly planned/unforeseen (parliamentary elections). The COVID-19 pandemic as an unforeseen crisis event allowed us to look behind the scenes and examine which decisions were made (selection criteria, metadata, etc.) by whom and how the datasets would be shared with which users.

It remains to be seen how unique the COVID-19 collections will become. Still, the fact that we as a community could follow almost “live” the act of heritagisation of an unfolding event — through interviews with archivists, surveys with archival institutions, and analysis of URLs and metadata — can undoubtedly be labelled as unique (Aasman et al., 2021). Heritagisation is a concept used in archaeology to indicate a more “processual understanding of heritage making” (Thouki, 2022, p. 1036). Indeed, it is worthwhile considering the logic of heritagisation of an ongoing event on the live web as a process that adds a new dimension to the study of transnational events, as it deals no longer with collecting and archiving materials long after an event has ended. This might not be so self-evident, as Astrid Erll observed, as it might be challenging to understand how an event emerges “if people don’t know what the event is, where it started, and how it evolved?” (Erll, 2020, p. 867). Another level of complexity is related to how these potential web archival materials need to be understood within the broader transnational memory cultures. If, as Andrew Hoskins (2011) points out, we are living in an age of real-time and instantaneous memory-making on a global scale, “boundaries of space, time and social experience” become less dominant (Keightley & Schlesinger, 2014, p. 747). At the same time, this book showcases not only the opportunities but also the many obstacles in place for both web archivists and researchers, whether they are related to technical elements, national legal policies, or biases in archives’ selection criteria.

The book’s structure

This edited volume takes a multifaceted approach to exploring transnational perspectives in web archive studies. It is organised into five parts that collectively examine this central theme through complementary lenses. Part I grapples with studying entire national web domains through a transnational perspective. Part II provides a grounded exploration of the COVID-19 crisis as a distinctly transnational event captured by web archives. Part III turns toward methodological considerations, unpacking techniques, and skill sets for conducting transnational web archive research. Part IV critically engages the politics and power dynamics inherent to web archives as institutionalised collections. Finally, Part V considers institutional and legal challenges to transnational web archive studies.

In the opening chapter, Gabriele Balbi traces how the relations between the web and history have been studied from three different perspectives. He addresses the “History web,” where the web is mainly taken as a platform to help historians in their daily work because of digitised sources, databases, and the creation of websites dealing with history; “Web history” which mainly focuses on websites, the material and software dimension of the past web and the archived web, and finally, the “History of the web,” where the political, economic, technical, and cultural history of the web itself is at the centre.

Introducing transnational web archive studies

In Part I, Anat Ben-David, Adam Amram, and Noa Turgeman examine the crucial role of iconography in mediating the climate crisis over the past decades. Earlier studies identified the emergence of climate change aesthetics through the creation of several iconic images, such as the polar bear and melting icebergs. By analysing the evolution of images published in the climate section of news websites across 30 countries from 2010 to 2022, the authors of this chapter trace the historical significance of news media in shaping collective imaginaries of the climate crisis through visual framing. In parallel to these findings, the chapter also presents a comprehensive methodological approach utilising the Internet Archive Wayback Machine (IAWM) to conduct large-scale transnational visual web historical analyses.

Although the IAWM provides interesting opportunities for transnational research, the chapter by Janne Nielsen, Yves Maurer, and Eld Zierau emphasises two significant obstacles to transnational web archives studies: The closed nature of many national web archives and the different aims, scopes, and legal frameworks of those national archives. This chapter asks how we can better understand what the holdings of different archives contain, how we can compare holdings across archives, and what types of analyses can be done based on such comparisons. The authors suggest extracting summaries of web archives' holdings using a program they developed called *cdx-summarize* to get simple, comparable statistics (number and size of domains per mime type per year). Summary statistics can be used to understand the collection of an individual archive but also to compare holdings across archives or to analyse historical developments. To illustrate the method's potential, the chapter uses materials from the national web archives of France, Luxembourg, Denmark, and Hungary, as well as what has been preserved of these countries' national web domains in the collections of the Internet Archive and Common Crawl, to provide examples of possible comparative, transnational studies of archived national web domains.

Another approach that can be useful is to study the evolution of the national web and use this as a starting point for transnational studies is evident in the chapter by Carmen Noguera, Janne Nielsen, Yves Maurer, and Ben Els. They argue that the domain names list from a national registrar can be a relevant source to study similarities and differences between national web domains. While doing so, they assess if the existing methodology used in national web studies can be reproduced to establish some comparisons, trends, and lessons learnt to apply in transnational studies. As a case study, the authors studied the evolution of the Luxembourgian country code Top-Level Domain (ccTLD), .lu, the national registrar holdings with the national and international web archive institutions to assess the (in)completeness of the Luxembourg web archives, which showed the importance of national and international partnerships for national web preservation.

To what extent is it possible to make a comparative study of several countries' national web domains, based on the holdings in different national European web archives? Niels Brügger explores this question by interviewing staff from five European web archives. The chapter investigates themes such as: What does the web archive landscape look like for a researcher who wants to make a transnational study of different nations' web domains across different web archives? And how do the differences and similarities of the web collections affect how the researcher can navigate this landscape?

In Part II, the focus shifts from national web domains to web archival collections related to the COVID-19 crisis as a transnational event. Friedel Geeraert, Nicola Bingham, Helle Strandgaard Jensen, Jane Winters, and Frédéric Clavert discuss how special collections were created during the COVID-19 pandemic. They base their analysis on the findings of an oral

history project organised under the auspices of WARCnet. A series of 12 semi-structured interviews were conducted with librarians and archivists involved in the planning and implementing COVID-19 special collections.

In a follow-up chapter, Nicola Bingham, Karin de Wild, Friedel Geeraert, and Caroline Nyvang focus on how GLAM institutions (Galleries, Libraries, Archives, and Museums) in Europe documented the web during the pandemic. They examine the main similarities and differences between European countries in creating COVID-19 web archiving collections, and the impact of differences in the maturity of web archiving programmes. Maturity was interpreted as encompassing the duration of the web archiving programmes, how web archiving practices have evolved over time, and the scale on which web archiving is done.

The question is what we can learn from the URLs in those event-related special collections. To answer these questions, Friedel Geeraert, Karin de Wild, and Susan Aasman undertake a comparative study about COVID-19 web archive collections at the IIPC and the national libraries of Luxembourg, Hungary, France, the UK, Denmark, and the Netherlands. The study was based on a collaboration between web archivists, academic researchers, and data scientists in 2020 with the aim of assessing how to make use of seed lists and accompanying metadata provided by the archival institutions. This chapter considers the usefulness of the URLs provided in understanding the scope of COVID-19 web archive collections from the point of view of both researchers and web archivists.

The chapter by Valérie Schafer, Susan Aasman, Frédéric Clavert, Karin de Wild, and Joshgun Sirajzade explores a more research-specific theme by considering the presence, retrievability, and analysis relating to women, gender, and COVID-19 in the international “Novel Coronavirus IIPC Collection.” This collection raises challenges such as multilingualism, “big data” and scalable reading, access and searchability, silence and noise, duplicates, and loss of information. The authors explored ARCH, a research interface developed by the Archives Unleashed Team, using computational tools and algorithms for conducting text mining and topic modelling.

Part III focuses on methods and skills in web archive studies, while also pointing at the complexities of doing so. The first chapter of this part aims at understanding the web and thus web archives as modern information ecosystems in a transnational digital world. Information ecosystems are defined as broad spaces of information production and dissemination involving core stakeholders focused on a common topic. In this chapter, Mathew S. Weber examines contexts as varied as news media production, policymaking processes, and the dissemination of health-related information, to underscore the utility of web archives in this space. Weber further provides examples from various geographic contexts to underscore the international nature of information ecosystems. He points to the methodological challenges of integrating disparate datasets to map information ecosystems through web archives. The closing discussion offers the reader tangible guidelines for studying information production and dissemination using web archives, including the future utility of web archives as information ecosystems that increasingly expand beyond the boundaries of the web to incorporate mobile and app-based domains.

The virtual museum is one context in which to examine the transformation of web-based information ecosystems. Nadezhda Povroznik’s chapter explores how the emergence of the web has reshaped museum practices. As museums have adapted to the digital age, their approaches to curating, selecting, communicating, and publishing materials online have evolved in tandem with shifting visions of their institutional roles and priorities. This chapter considers the phenomena of museums’ multiplatform digital presence, analysing case studies

Introducing transnational web archive studies

to trace the evolution of museum websites through close reading methods. Additionally, the chapter discusses the potential for scaling up research by incorporating mixed qualitative and quantitative methods.

Povroznik also emphasises the importance of born-digital cultural heritage and the need to ensure its sustainability. This is very much the topic of the next chapter about exploring skills and training requirements for the web archiving community. Helena Byrne, Juan-José Boté-Vericad, and Sharon Healy examine current trends in the challenges for library and archival roles in a digital context and then take a look at the challenges, and skills and training requirements for librarians and archivists who participate in web archiving as part of their role in a cultural heritage organisation. They further examine the landscape of training materials and courses organised by various organisations and institutions across Britain, Ireland, and Spain. Overall, the chapter finds that web archiving presents some unique challenges requiring additional skills for librarians and archivists because of the characteristics of the web and continual advances in technology. It further finds that a collaborative transdisciplinary approach is needed when it comes to designing new curricula for web archiving education and this should be representative of the multiple stakeholders who participate in the activities of the web archiving lifecycle.

Márton Németh also addresses the need to teach web archiving courses in universities. He made an intensive inventory of the various disciplinary perspectives in contemporary web archiving education. The chapter further offers some best practices and views into the future of web archiving education.

Part IV shifts the attention to the politics of web archives as collections. Saskia Huc-Hepher and Xiao Ma focus on their experience of curating diasporic “French” and “Chinese” “community” collections within the UK Web Archive. They make a case that the normalisation of these collections as belonging to “a community” serves to invisibilise the underlying problems at play, particularly in the blurry in-between space of the transnational context. Drawing on French philosophical thought, specifically that of Jean-Luc Nancy and Jacques Derrida, alongside Benedict Anderson’s seminal notion of *imagined communities* and contemporary xenofeminist ideas from the international collective Laboria Cuboniks, the authors investigate the trouble with “community” in three dimensions: first, constructing community, as web curation process and principles; second, deconstructing community, as concept in the transnational archiving space; and third, reconstructing community as xenofeminist antidote to the archive’s endemic trouble.

Likewise, critically reflective is the chapter by Sophie Gebeil about the representation of the Middle East and North African (MENA) websites in the “Novel Coronavirus IIPC Collection.” Considering the political implications connected to web archiving, the place of the MENA web sphere in the IIPC collection serves as a case study, prompting reflection on strategies for researchers and archivists dealing with countries marked by geopolitical tensions and almost non-existent archiving initiatives. How did the IIPC manage to collect web content from the MENA related to COVID-19? What are the outcomes and limitations of this initiative? More broadly, what insights can be gleaned from this experiment regarding the challenge of designing a more inclusive transnational web archive? Drawing on the contents of the collected websites, the chapter analyses how the geopolitical specificities of the MENA zone impart a distinctive characteristic to the collected sites, highlighting how discussion centred on decolonising web archives could be initiated.

The challenges in curating histories take on another dimension when considering social media platforms. Simone Natale considers in his chapter the case of the acquisition of the

social media platform WeChat by the Victoria & Albert Museum through a theoretical perspective that is sensitive to the multiple dimensions or “lives” of the web and of digital objects in general. The practices and strategies that museums activate to preserve and exhibit elements of the web provide crucial insights for empowering not only ongoing efforts of museums to curate software web-based content but also wider approaches to web historiography and archiving.

In recent years, participatory platforms have gained significant popularity in the global cultural heritage domain. As Marta Severo explains, research typically distinguishes two types of platforms. Firstly, institutional platforms, often adopting the form of crowdsourcing tools, are established by libraries, archives, museums, and similar institutions to involve the public in specific tasks related to their collections. Secondly, informal or formal groups of enthusiasts utilise the web to create collections of cultural objects, typically on commercial platforms and social media. By analysing a sample of six transnational participatory platforms (comprising three institutional and three amateur platforms), Severo identifies and discusses the theoretical, technical, and ethical challenges associated with web archiving digital participation in the cultural heritage domain.

Michael Kurzmeier explores the process of creating an archive of political web defacements while also addressing the potential research opportunities presented by analysing them. The chapter reveals the challenges web defacements pose, such as their interconnectedness with the original page at the time of defacement. The chapter further investigates how cross-referencing with other web archives can help contextualise web defacements. Kurzmeier explores existing web archiving playback software and custom solutions in the context of creating a web archive from defaced pages. The chapter concludes by providing a publicly available data source for the research of web histories, social movements online, and the history of political hacking in the pre-2002 era.

Part V considers institutional challenges and starts with a chapter about the struggle to archive webTV. Based on a corpus of a substantial number of French activist webTV sites over a period of five years, from May 2010 to May 2015, Sophie Gebeil and Jérôme Thière developed a pipeline for extracting information from the files collected by the Institut National de l'Audiovisuel in France. The aim was to use the HTML files of the corpus to enable data mining to identify the main characteristics of the corpus and the changes observed during the five-year period. Focusing on methodological aspects, this chapter presents how a highly localised case study can contribute to discussing (trans)national web-archival research for historians, addressing the lack of international standards.

A specific need related to web archival research is the need for good data management practices. However, as web materials are very different from other historical sources, the generic research data management practices are not sufficient on the operational level. When it comes to transnational studies, it gets even more complicated. The chapter by Eld Zierau, Beatrice Cannelli, Sally Chambers, Olga Holownia, Sharon Healy, Ditte Laursen, Susan Aasman, and Ulrich Karstoft examines evidence about current research data management practices for web materials. The authors describe the practices and challenges compared to general research management practices and specifically look at how this can support the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles to archived web data. On this basis, the chapter discusses specific gaps, including what can be used as recommendations and what needs further investigation.

Lise-Anne Denis takes on a legal perspective as she maps several European directives and regulations related to web archiving and de facto their implementation in national law.

Introducing transnational web archive studies

Researchers prefer for the archives to be as open as possible, but this tends to collide with the legal framework, which often rather focuses on restricting access. As Denis stresses, it is important to know these legal requirements in advance. Her chapter focuses on the two main legal areas applicable to web archives studies at the European level — copyright, and the General Data Protection Regulation (or GDPR) — to show how it can actually translate at the national level and what it practically implies for researchers and archiving institutions of these two states.

So far, the focus has been mainly on European and North American web archival practices. In the last chapter of the book, Moisés Rockembach addresses legal and research challenges through the analysis of relevant literature, documents, and case studies within the context of Brazil. In Brazil, there are specific laws and resources put in place to ensure the preservation of government documents, and efforts have been made to establish a policy for digitally preserving websites and social media. The research aims to describe the progress made, as well as the obstacles and gaps that exist in web preservation in Brazil, while also exploring its connections to other countries in Latin America.

Interwoven between these thematic parts are “Conversations”-chapters facilitating cross-pollinating dialogues. A “Conversation” takes up one or more topics or discussions addressed in the preceding part, and/or points forward to issues discussed in the following part. The editors formulated the questions of each conversation, based on suggestions from the authors, and all chapter authors were invited to contribute. The conversations took place in a shared online document.

Conversation I reflects on the very notion of “the transnational” and its implications for web archives. Conversation II examines how “events” are construed and delimited within transnational web archive contexts. Conversation III grapples with defining and representing “communities” through the transnational prism of web archives. Conversation IV bridges methodological and political dimensions by addressing institutional challenges and charting future trajectories for transnational web archive studies. Lastly, in Conversation V the authors chart future trajectories for transnational web archive studies.

Readers of this book will notice that in line with the main topic of the book, all references to online sources point to archived versions in the Internet Archive. Additional references to colour figures lead to WARCnet’s community repository at Zenodo (<https://zenodo.org/communities>). All chapters have undergone a round of single-blind review, in addition to editorial review. Finally, readers of this book may refer to the glossary, which explains core concepts, organisations, and events.

Acknowledgements

First and foremost, this book is indebted to the participants of the international research network WARCnet, including all invited keynote speakers and presenters, because they stayed in the network during troubled COVID-19 times, and because they kept their initial drive to the very end of the project. They are hereby thanked.

This book and the WARCnet network on which it builds would not have been made possible without the support of the Independent Research Fund Denmark | Humanities; hence we owe the funder many thanks.

We would also like to offer our thanks to Sophie Gebeil, Associate Professor of Contemporary History at Aix-Marseille University for having hosted the network meeting at the Museum of Civilizations of Europe and the Mediterranean in Marseille in June 2023,

where authors discussed the first drafts of their chapters, and to Nathalie Fridzema, the minutes keeper who provided extensive minutes from this meeting.

The book has also benefited from the financial support of the following funders: Aarhus University, Department of Media and Journalism Studies, Université de Luxembourg, Centre for Contemporary and Digital History, Université Paris Nanterre, Laboratoire Dicen, and Grup de Recerca Consolidat “Cultura, Biblioteques i Continguts Digitals (CBCD)” (grant no. SGR-2021-00282/Generalitat de Catalunya).

In addition, we would like to express many thanks to our publisher, Taylor & Francis/Routledge, in particular Editorial Assistant Manas Roy, for having shepherded the final steps of the book so smoothly, and Editor Heidi Lowther for having supported the idea of this book from the very beginning, or, in fact, before there was a real beginning.

Finally, in the spirit of transnational collaboration that underpins this volume, our editorial journey has also travelled across Denmark, The Netherlands, and Israel. Although most of the editorial work has been done remotely, our face-to-face meetings have been marked not only by scholarly exchange but also by culinary camaraderie. These meetings were punctuated with delightful lunches, generously provided under the humorous guise of “thanking the Danish taxpayer.” This inside joke, evolving into our shared mantra of “you know who you should thank,” serves as a whimsical reminder of the invaluable support and contributions that have made this book a reality and the pleasure of co-editing this volume.

References

- Aasman, S., Bingham, N., Brügger, N., De Wild, K., Gebeil, S., & Schafer, V. (2021). Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections. https://web.archive.org/web/20230924000852/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Aasman_et_al_Chicken_and_Egg.pdf
- Appadurai, A. (1996). *Modernity at Large: Cultural Dimensions of Globalization*. University of Minnesota Press.
- Baucom, E. (2019). A brief history of digital preservation. In *Digital Preservation in Libraries: Preparing for a Sustainable Future* (pp. 3–19). American Library Association.
- Ben-David, A. (2016). What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. *New Media & Society*, 18(7), 1103–1119. <https://doi.org/10.1177/1461444816643790>.
- Ben-David, A., & Huurdeman, H. (2014). Web archive search as research: Methodological and theoretical implications. *Alexandria*, 25(1–2), 93–111. <https://doi.org/10.7227/ALX.0022>.
- Brügger, N. (2018). *The Archived Web: Doing History in the Digital Age*. The MIT Press.
- Brügger, N. (2020). *Welcome to WARCnet*. WARC Papers. https://web.archive.org/web/20230921193419/https://cc.au.dk/fileadmin/user_upload/WARCnet/1.Bru_gger_Welcome_to_WARCnet.pdf.
- Brügger, N. (2021). *The WARCnet Network: The First Year*. WARCnet Papers. https://web.archive.org/web/20230902214749/https://cc.au.dk/fileadmin/user_upload/WARCnet/Bru_gger_The_first_year.pdf.
- Brügger, N. (2022a). *The WARCnet Network: The Second Year*. WARCnet Papers. https://web.archive.org/web/20230902214718/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Bru_gger_The_second_year.pdf.
- Brügger, N. (2022b). Media events in an age of the Web and television: Dayan and Katz revisited. *Nordic Journal of Media Studies*, 4(1), 37–55.
- Brügger, N. (2023a). *The WARCnet Network: The Third Year*. WARCnet Papers. https://web.archive.org/web/20230902214751/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Bru_gger_The_third_year.pdf.

Introducing transnational web archive studies

- Brügger, N. (2023b). *The WARCnet Network: The Fourth Year*. WARCnet Papers. https://web.archive.org/web/20230908030806/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Bru__ger_The_fourth_year.pdf.
- Brügger, N., & Laursen, D. (Eds.). (2019). *The Historical Web and Digital Humanities: The Case of National Web Domains*. Routledge.
- Brügger, N., Nielsen, J., & Laursen, D. (2020). Big data experiments with the archived Web: Methodological reflections on studying the development of a nation's Web. *First Monday*, 25(3). <https://doi.org/10.5210/fm.v25i3.10384>.
- Burawoy, M., Blum, J. A., George, S., Gille, Z., & Thayer, M. (2000). *Global Ethnography: Forces, Connections, and Imaginations in a Postmodern World*. University of California Press.
- Chalaby, J. K. (2005). From internationalization to transnationalization. *Global Media and Communication*, 1(1), 28–33. <http://doi.org/10.1177/174276650500100107>.
- Dayan, D., & Katz, E. (1992). *Media Events: The Live Broadcasting of History*. Harvard University Press.
- Erl, A. (2020). Afterword: Memory worlds in times of Corona. *Memory Studies*, 13(5), 861–874. <https://doi.org/10.1177/1750698020943014>.
- Fickers, A., & Johnson, C. (2010). Transnational television history: A comparative approach: Introduction. *Media History*, 16(1), 1–11. <https://doi.org/10.1080/13688800903395411>.
- Goldsmith, J., & Wu, T. (2006). *Who Controls the Internet? Illusions of a Borderless World*. Oxford University Press.
- Hallin, D. C., & Mancini, P. (Eds.). (2011). *Comparing Media Systems Beyond the Western World*. Cambridge University Press.
- Hepp, A., & Couldry, N. (2009). Introduction: Media events in globalized media cultures. In *Media Events in a Global Age* (pp. 1–20). Routledge.
- Higbee, W., & Lim, S. H. (2010). Concepts of transnational cinema: Towards a critical transnationalism in film studies. *Transnational Cinemas*, 1(1), 7–21. <https://doi.org/10.1386/trac.1.1.7/1>.
- Hoskins, A. (2011). Media, memory, metaphor: Remembering and the connective turn. *Parallax*, 17(4), 19–31. <https://doi.org/10.1080/13534645.2011.605573>.
- Keightley, E., & Schlesinger, P. (2014). Digital media — social memory: Remembering in digitally networked times. *Media, Culture & Society*, 36(6), 745–747. <https://doi.org/10.1177/0163443714532985>.
- Nielsen, J. (2019). Experimenting with computational methods for large-scale studies of tracking technologies in web archives. *Internet Histories*, 3(3–4), 293–315. <https://doi.org/10.1080/24701475.2019.1671074>.
- Ogden, J. (2020). Saving the web: Facets of web archiving in everyday practice. Doctoral dissertation, University of Southampton.
- Ogden, J. (2022). “Everything on the internet can be saved”: Archive Team, Tumblr and the cultural significance of web archiving. *Internet Histories*, 6(1–2), 113–132.
- Ogden, J. (2023). ‘Crisis Collection’: Archival logics in the face of dead and dying platforms. In Paßmann, J., Gerzen, L., Schories, M., Ogden, J., Maemura, E., & MacKinnon, K. (Eds.), *Web Histories in the Making: Web Archives & the Logics of Practice*. Panel presented at AoIR2023: The 24th Annual Conference of the Association of Internet Researchers. AoIR, 6–8. <https://spir.aoir.org/ojs/index.php/spir/article/view/13535/11483>.
- Ogden, J., & Maemura, E. (2021). ‘Go fish’: Conceptualising the challenges of engaging national web archives for digital research. *International Journal of Digital Humanities*, 2(1), 43–63. <https://doi.org/10.1007/s42803-021-00032-5>.
- Rogers, R. (2013). *Digital Methods*. MIT Press.
- Schafer, V., & Winters, J. (2021). The values of web archives. *International Journal of Digital Humanities*, 2(1), 129–144. <https://doi.org/10.1007/s42803-021-00037-0>.
- Shome, R. (2006). Transnational feminism and communication studies. *The Communication Review*, 9(4), 255–267.
- Stam, R., & Shohat, E. (1994). Contested histories: Eurocentrism, multiculturalism, and the media. In D. T. Goldberg (Ed.), *Multiculturalism: A Critical Reader*. Blackwell.

The Routledge Companion to Transnational Web Archive Studies

- Thouki, A. (2022). Heritagization of religious sites: In search of visitor agency and the dialectics underlying heritage planning assemblages. *International Journal of Heritage Studies*, 28(9), 1036–1065. <https://doi.org/10.1080/13527258.2022.2122535>.
- Weltevrede, E., & Helmond, A. (2012). Where do bloggers blog? Platform transitions within the historical Dutch blogosphere. *First Monday*, 17(2). <https://doi.org/10.5210/fm.v17i2.3775>.
- Winters, J. (2019). Giving with one hand, taking with the other: E-legal deposit, web archives and researcher access. In P. Gooding & M. Terras (Eds.), *Electronic Legal Deposit: Shaping the Library Collections of the Future* (pp. 159–178). Facet Publishing.

2

“HISTORY WEB,” “WEB HISTORY,” AND “HISTORY OF THE WEB”

Three subfields and why (and why not)
integrating them

Gabriele Balbi

Introduction

The history of media and communication has boomed in recent decades, with university courses, book series, conferences, sections/divisions of associations, and special issues of journals emerging over time. Likewise, the history of digital and digitization has also grown rapidly and, more in general, digital communications have begun to be increasingly analyzed from a historical perspective (see Balbi & Magaudo, 2018, for an analysis on this topic).

Under the impetus of international scholars and organizations such as the Web ARChive studies network researching web domains and events (WARCnet), which is the protagonist of this book, the web has also been increasingly analyzed from a historical perspective. However, when we say that the web is analyzed from a historical perspective, are we certain of our meaning? Or are there different and potentially conflicting ways of interpreting and studying the relations between the web and history? This chapter aims to highlight the various definitions and subfields that emerge from the relationship between historiography and the web. It is an introduction, or rather a preface, to this book because it makes an upstream theoretical and historiographical argument, which then the various scholars joining the WARCnet community (many of whom are cited in this chapter) have analyzed in their writings each in a different way.

Three subfields of the relationship between history and the web can be identified: the “History Web,” the “Web History,” and the “History of the Web.” At the intersection of these three subfields, some research has already been conducted. In his book *The Archived Web*, Niels Brügger (2018) did something similar to what I attempted in this chapter, identifying different cases of web history (actually naming “Web History” something I have placed in other categories) and claiming that there are “multiple-strata cases” in which different typologies of historical studies of the web are combined. Apparently, the “History Web,” the “Web History,” and the “History of the Web” all play differently with the words history and

web, but small changes in word placement result in significant changes in the way the historical study of the web is approached and in the posture that researchers take when looking at the object-web.

In the second section of the chapter, I discuss the reasons why to keep the subfields separated or instead combine them. The hope is that this theoretical approach may be useful for the WARCnet community to reflect on their work, but more generally for all media and communication scholars who approach the web from a historical perspective, would like to have a clear overview and clear definitions of different sectors, and want to situate themselves in one of the three subfields or to make them interact with one another.

History Web

The first subfield considered is the so-called “History Web.” This term was notably introduced by Cohen and Rosenzweig (2006, p. 13) to indicate the fact that already in the early 2000s the web was becoming a platform to help historians in their day-to-day work. Indeed, contemporary historians began to use the web for their daily research practices very early, sensing its potential in connecting databases, providing access to digitized sources, but also in disseminating history. Over time, archives accessible via the web have been created and they are so powerful and useful to historians precisely because they are online. Among many possible examples, textual databases for newspapers and magazines, such as Factiva and Lexis Nexis, have come to the fore, various textual or audio–visual archives managed by international and national institutions or private individuals have made their catalogs accessible online or at least allow for subject search via the web.

Along a different trajectory, but still in the realm of the “History Web,” there has been a proliferation of websites dealing with history: similarweb.com, for example, indicates that as of April 2023, the most visited websites on history were archive.org (which hosts the Wayback Machine, which will be discussed later in the chapter) worldhistory.org and biografiasyvidas.com.

The ways in which the web has impacted historical research have produced many contributions and it is almost impossible to do an exhaustive review. Perhaps the most encompassing text on the topic was written by Ian Milligan (2019), but as noted many of the authors, who are part of WARCnet and this book, have participated over the years with various contributions and topics of discussion. A substantial amount of historiographical research and reflection focuses on the web as a source and its consequent archiving: there is research on the ways, problems, and difficulties of archiving the web (see, for example, Ben-David, 2016; Gomes et al., 2021; Schafer & Winters, 2021); on the substantial difference between web archives and traditional paper archives or to audio–visual archives (Ankerson, 2012; Musiani et al., 2019); on the methodologies employed for web archiving and web historiography (see the special issue edited by Aasman, de Haan, & Teszelszky, 2019); and on the computational methods that have invested historical research in recent years and that seem destined to change it in toto (see, for example, the roundtable Brügger et al., 2019; Blaney et al., 2021). This review is not exhaustive at all, but historians may have a hint on how the web has changed their own discipline in a few decades.

The “History Web” has emerged not only in academia and there are several interesting examples that can be highlighted at different levels. For instance, a multitude of public and private institutions have chosen to digitize their archival collections and make them accessible online, either in part or in their entirety. These available sources can be used

“History web,” “web history,” and “history of the web”

by historians, but also and specifically by the people who approach them maybe with a degree of nostalgia: typically, public service broadcastings around the globe have digitization projects and have created specific repositories or they use social media to connect with their audiences and to enhance their memories. A transnational example of “History Web” can be considered Europeana, a web portal created by the European Union with digitized cultural heritage collections of more than 3,000 institutions across Europe, such as libraries, archives, museums, and galleries (Petras et al., 2017).¹ The efforts of different European institutions in several countries, the number of digitized sources, and the possibility to search in so many different ways made Europeana a successful example of “History Web” which has become a reference for historical research today.

These archives, at various levels, have digitization and source preservation programs for which they generally get national and international funding or, those who do not have such programs are now perceived as falling behind in preserving and promoting their sources to the historical community. Let’s consider for example the public-funded *Conseil Européen pour la Recherche Nucléaire* (European Organization for Nuclear Research, better known as CERN), the international institution where the web itself was conceived, which has digitized and made accessible some of the sources precisely on the origins of the World Wide Web (WWW). They can be found on the page devoted to the WWW collection² and, among them, the well-known proposal written by Tim Berners Lee in 1989, an object of veneration and source of fetishism at various levels and which can be downloaded as a pdf³ (on the WWW collection preserved at the CERN archive, see Fomasi et al., 2023).

Historiographic work is now substantially based on web resources. If one thinks step by step about typical historiographic research, there is a web component in all stages: one often starts with websites like Google Scholar or Elicit, library or publisher databases to scan or acquire books and articles for literature review; then archivists are typically contacted via email found on reference websites; an archival visit is then planned or if sources are digitized and available one accesses them via the web; sources are then often analyzed with digital tools available on the web or web-based; articles or books are published which in turn enter databases and are searchable, and they are given prominence on personal social media by historians themselves or by publishers or, again, in web-hosted newsletters. The job of historians would be very different without the WWW.

Web History

The second disciplinary subfield that academically connects history and the web has been defined, as already done in the past, as “Web History” (Brügger, 2010; Rogers, 2017). It mainly focuses on the analysis of past websites at any level, their preservation, and their attribute of source for investigating past history, but also as autonomous and investigable documents with their own historical legitimacy.

There are at least three ways in which “Web History” can be considered. First, past websites are historical sources to study contemporary history: indeed, historians can analyze past events through web pages concerning that event. For example, websites can be used to study migration movements in North Africa in the early 2000s as done by Sophie Gebeil (2019). Another example, in the narrower field of media history, shows that past websites can be historical sources to study specific media sectors: for example, in a book devoted to the multiple relationships between public service broadcasting and the WWW, Burns and Brügger (2021) studied how public service broadcasters described themselves and are in

contact with their audience through their institutional websites. In this sense, the website becomes a powerful self-narrative mechanism, and studying its evolution helps precisely to understand how broadcasters have represented themselves over time.

Websites or social media can thus be useful sources, specifically, for analyzing the history of public or private institutions, in their modes of self-representation and in their attempt to identify long or short-run narratives. The past website thus helps to understand what is cut or modified from the institution's memory and what, instead, newly becomes part of the standard narrative over time. It seems to me, however, that the use of websites as a source is still limited among contemporary historians and other scholars, in favor of more classic sources that have been used for decades such as archival documents and audio-visual sources. This may seem odd, as websites are often textual (and of course also visual) and thus are in line with many of the sources most used by historians, but they also have an unstable nature and they have been institutionally preserved since relatively few years.

A second mode of historical study of websites that composes "Web History" is more self-referential: websites can themselves be objects of historical analysis. One can, in short, focus on the website not so much for its relevance in narrating a historical event, but as an autonomous and legitimate object of analysis. To draw a parallel with media history, my own field, we are talking about doing history of photography through photographs, the history of cinema through films, history of radio and television as a succession of the most relevant programs. Among various research on the subject, we can mention the study on the websites of American political campaigns (Foot & Schneider, 2006), or on the websites of crucial media events like the Olympic games from 1996 to 2016 (Brügger, 2022), aiming to integrate the web sphere into the famous 1990s model of Dayan and Katz on media events.

I would expect in the future to have anthologies and canons of websites of the past, in which on par with the history of film or literature we point to websites that have been exemplars, that have become "classic" references for the construction of other websites. Of course, in this sense, the languages and tools for creating websites are also decisive: for example, the standard of the HTML language and all its variants has been accepted for several decades now, likewise, the idea that a link should be underlined and have a different color, or even WordPress as a website building tool has become the standard tool when one wants to open a website. All of these are historically determined factors and affordances in the ways of imagining, constructing, and enjoying websites that help explain their structure, usability, and even historical dimensions (we immediately recognize a website from the 1990s not only by its use of Comic Sans, but also by its navigation methods, colors, mouse arrow, navigation browsers, and more).

Preserving and making available the sites and web pages of the past is thus a fundamental dimension for "Web History," which not surprisingly has unified historians, archivists, and preservation specialists into its conferences and methodological discussions. Under this perspective, WARCnet represents a perfect melting pot for "Web History," perhaps one of the places where the hybrid historical-technical nature of this sub-discipline is most evident, where skills straddling the humanities, social sciences, and computer science are needed.

But, as mentioned previously in the case of "History Web," also "Web History" is not only an academic discipline and non-academic actors have contributed greatly to preserving and restoring the past web. The Wayback Machine, for example, is a project launched in 1996 by the nonprofit organization Internet Archive with the aim of archiving pages and websites at the transnational level and then made available in 2001, when those websites

“History web,” “web history,” and “history of the web”

became freely available to any users (Mohr, 2007). Because of this interface, in many cases, it is now possible to trace back to web pages that have since been deleted or to earlier versions of current web pages: in short, the Wayback Machine is a useful archive for understanding the dynamic, diachronic and constantly moving dimension of a source that varies incessantly as the web is (see, for example, Schafer, 2019).

There are also initiatives of reenactment or outright restoration of websites, as CERN in Geneva has done, for example, in its restoration project where it has put the first (or one of the first?) web pages back online.⁴ The interest in these kinds of initiatives is also aesthetic in nature as the design of websites has changed and is changing profoundly over time. Keeping this in mind, the Web Design Museum has built a “Web design history timeline” in which it has mapped out the precisely aesthetic evolution of the most relevant websites from 1991 to 2006. According to this institution, the main goal is “to preserve the creative legacy of web designers from the turn of the millennium for future generations, since internet users in 2030 would hardly guess how unique the websites in 2003 were in terms of their design.”⁵ A final example of preservation projects related to “Web History” is instead promoted by the Computer History Museum, which focuses instead on surfing practices and the usability of websites. How did the experience of surfing the web, according to a well-known metaphor of the 1990s, look like? The California museum offers some materials and experiences to answer this question.⁶

History of the Web

The third and final subfield dealing with the relationship between the web and history can be defined as “History of the Web.” What I mean here by “History of the Web” has actually already been called in other ways by other scholars: “histories about the web in general” (Brügger, 2010, p. 9), “web as such” (Brügger, 2012, p. 317; Brügger & Milligan, 2019, p. 21), “The web as a whole” (Brügger, 2011, p. 26; Brügger, 2018, p. 35), “Web history in context” (Schafer & Thierry, 2018), and many others. However, I prefer to use the phrase “History of the Web” also, especially because it seems to me (and to the ear of a non-native speaker like myself) to clearly indicate the central *object of study*: the web, precisely. The “History of the Web” is in fact a subfield that has at the center of its analysis *the web but websites* and that considers the genitive of the web narrowly. The studies that belong to this field put the web at the center, not as an aid to historical research, nor as a source, nor in its archival, aesthetic, or graphic dimension, but precisely by making the political, economic, technological, and cultural history *of the web* the center of interest — with the consequent risk of making a web-centric history of the web, to paraphrase David Morley (2009).

The “History of the Web” considers the web as a medium besides its content dimension: like studying newspapers but articles or journalistic styles, photography without pictures, broadcasting without considering programs, anchorwomen/men, or TV genres. Rooted in a McLuhanian perspective, the “History of the Web” actually applies to the WWW approach and methods quite common to other media and communication technologies. One thinks of the various histories of social or cultural construction of newspapers, broadcasting, and telecommunications with an STS, ANT, or classical social history approach. There are also various studies with a political economy perspective that aim to identify power structures in media, from Hollywood cinema to large digital platforms. Or even scholarly contributions close to or in the field of business history that consider the corporate histories of music or

record labels or brands that produce digital devices. There are many other disciplines and forms of histories that could be mentioned, but this is enough to give the idea. In this view, the web becomes a social construction where politics, institutions, corporate dimensions, technology, users, and many other factors have an influence. But the web, in itself, or in the companies and spaces that animate it, can also act as an actant that plays a role in structuring everyday actions, leisure and work time, and socialization practices.

There are various examples of “History of the Web,” and many of these focus mainly on the beginnings of the history of the web, which is problematic because they study this technology when it was used by a minority of nerds or early adopters — this is a problem typical of many early histories of technology as David Edgerton (2007) mentioned in his brilliant book *The Shock of the Old*.

The first typology of academic research in the subfield “History of the Web” can be retraced in the self-narratives of the founding fathers, who told how the web was born at CERN in Geneva and how it developed in the early years of its introduction. These narratives are varied, even outside the academia with interviews and celebratory events, but they essentially originated in two books authored or co-authored by Tim Berners-Lee and Robert Cailliau (Berners-Lee, 1999; Gillies & Cailliau, 2000).

A second field in which the “History of the Web” has seen various studies flourish are narratives, imaginaries, myths, and discourses around the web over time (see, for example, Bory et al., 2016; Natale & Bory, 2017; Bory, 2020). These studies have often analyzed available sources (until recently extremely scarce or inaccessible also due to the 30 years closure of archives) on how the history and evolution of the web has been told. One of the most interesting ways of approaching this dimension is through metaphors: the internet in general and the web in particular have long produced metaphors that have circulated in the social fabric: think of the frontier, the virtual plaza, information gathering by surfing, the global library, the walled gardens to describe the more recent platformization of the web, and several others (Wyatt, 2021). All these metaphors have, of course, a historical and a cultural dimension: that is, they change meaning either over time, retaining or losing some elements, or they live off national or local peculiarities and imageries. On the topic, I believe there is ample room for various comparative or transnational analyses in the future, as already claimed by Goggin and McLelland (2017).

Just as seen for other media, the business histories of companies that have become central to the web is a third dimension of the “History of the Web” that has produced various studies of companies such as Amazon, Facebook (and later Meta), Google, Alibaba, Wikipedia, and many others. Some of these histories are accurate and based on corporate sources (although they struggle with the atavistic scarcity of internal documents that these companies make available to historians), others are hagiographic and tend to extol the genius of founders, innovators, engineers, etc. These companies have not only an individual dimension but also a collective one as the acronym GAFA/GAFAM (Google–Amazon–Facebook–Apple–Microsoft) shows (Galloway, 2017): the main goal is emphasizing the ability of all these players to concentrate large amounts of capital and expand throughout the world or nearly so, as, in some countries, such as China and Russia, these companies are banned from entering the market. For this chapter, it is interesting to note that such a block of companies is also variously referred to as internet or web companies, web giants, web-based companies, and many other designations containing the term web. In other terms, there is a way to study the “History of the Web” which is starting and focusing on the companies and, more in general, on the corporate perspective.

“History web,” “web history,” and “history of the web”

Especially with a political economy slant, there is an interesting national/local and international/global dialectic that has emerged in the “History of the Web.” On the one hand, there are various national histories of the web, in which peculiar and distinctive aspects for each country have been emphasized, but also the common aspiration of individual countries or individual web cultures to emulate, for example, the United States or to create digital heroes shaped on the American model (Schafer & Thierry, 2016; Goggin & McLelland, 2017). On the other hand, there has been work that has sought to identify a more global and collective spirit about the political consequences of the web: an example for historians of the web is José van Dijck’s (2013) successful book on the culture of connectivity, which aims to advance a “critical history of social media,” according to the subtitle. The same can be told for the histories of platformization, a macrophenomenon that changed profoundly the social media as we knew them according, for example, to Anne Helmond (2015) in the case of Facebook. Are histories of social media, taken as a macro category or in their evolution as digital platforms, part of the “History of the Web” subfield? I believe so.

A final, typical historiographic strand of the “History of the Web” relates to periodization. In the case of the web, there are at least two types of periodization. The first and most popular borrows the nomenclature from software release cycles (web 1.0, 2.0, 3.0, and many others), to indicate how in each “era” a true or alleged radical change of the web would have occurred (see, for example, Vossen & Hagemann, 2007; Barassi & Treré, 2012; Allen, 2008; Kreps & Kimppa, 2015). This periodization is commonly accepted, so much so that even on Wikipedia there are clear definitions of Web 2.0 (“also known as participative or participatory web and social web refers to websites that emphasize user-generated content, ease of use, participatory culture and interoperability, i.e., compatibility with other products, systems, and devices for end users”) or web 3.0 (“The Semantic Web, sometimes known as Web 3.0 [not to be confused with Web3], is an extension of the World Wide Web through standards set by the World Wide Web Consortium. The goal of the Semantic Web is to make internet data machine-readable”). A second way of periodizing the web, destined to grow as the web itself ages, is by decades: Schafer and Thierry (2018), for example, edited a special issue on the 1990s as the turning decade for the internet and the web.

Periodization is a common practice for historians, but it is not the aim of this chapter to point out its strengths or weaknesses, even if it is interesting to note that even outside the academia there is a veritable craze for making timelines that bring order to the evolution of the web. While it is impossible to enumerate all timelines relating to the history of the web, I will provide some examples in addition to the “Web design history timeline” already mentioned. There are web timelines designed by world-renowned research centers such as the Pew Research Center,⁷ by global magazines such as *The Guardian*,⁸ or by private individuals.⁹ CERN, too, has proposed its own timeline, which can have at least a twofold value: it can indeed be seen as a way to understand how CERN self-reports itself as the protagonist of this story, but it is also a useful source to understand which version of the birth and evolution of the web seeks to be proposed and imposed as “official.”¹⁰ Comparing these timelines, and consequently, the stories they convey and tell could be a beautiful study of conflicting memories and the subjective nature of timelines that often appear neutral.

Reasons *not* to combine the subfields

After attempting to define and analyze the three subfields, some of the main academic work, and even the non-academic popularity of the historicized web, in the final part of this chapter

I focus on the reasons for combining or not combining the three ways of approaching web and history.

We start with the opposing reasons. “History Web,” “Web History,” and “History of the Web” are three different subfields with different research and analytical interests: the theoretical backgrounds are different, as are the methodologies and research questions that aim to investigate different issues that cannot be reconciled. But the difference that seems most significant to me is the fact that they can be categorized into different disciplines or fields. “History Web” is classifiable more in the field of historiography, particularly contemporary historiography, but I would say more generally it can have value in terms of historical research methodology. “Web History” can have a close placement with digital history, but it can also be on the borderline with design or even some hard sciences such as computer science. Among historians employed in this subfield, the interest goes as said not on the web per se, except for websites, but as a source for political, economic, and socio-cultural history. Finally, “History of the Web” can be considered a subfield of media and communication history because historians who practice it have interest in the web as media or as communication technologies. These three fields are then different from each other, each has specific research interests and questions, methodologies, journals, conferences, and even specific academic or technical jobs and positions that are not easily reconciled in most cases. The professional and academic figures afferent to these three subfields are also different: they are historians, media scholars, web archivists, software engineers, data analysts, and many other figures who interact with each other with difficulty because of different backgrounds and mindsets.

Reasons to combine the subfields

However, these three subfields share several common traits, so there are strong arguments in favor of combining them.

A basic commonality, which should not be underestimated, is the interest in the historical dimension of the web. Frequently, digital technologies, including the web, are analyzed from sociological, psychological, anthropological, political, and other perspectives, and the long-term dimension is frequently overlooked: in short, the three subfields mentioned above have in common the centrality of the past and the diachronic dimension, the analysis of changes, and the recognition of continuity.

A second common starting point, or better a mantra, reads as the *web is not the internet* (Brügger, 2013): thus, although these historians are interested in analyzing the web from different points of view, the web remains the sole focus of their research.

Although it may seem odd, there is a third similarity: all three subfields are essentially neglected or marginalized in their respective main fields. Web historians are interested in self-reflection on their discipline, in how sources are digitized and then made available through and because of the web. They are mostly methodologists and historiographers, in a discipline like history which has not shown passion to establish and discuss methods specifically.

Web historians use sources such as websites, which contemporary historians may find “strange” because they are visual, unstable, and have not been preserved by official institutions or archives for the past few years. They can also look more interested in the preservation of sources than in using them to tell stories (the main business of historians). Historians of the web, like media and communication historians, are neglected or at least considered quite marginal both by media and communication scholars and by contemporary historians as

“History web,” “web history,” and “history of the web”

well — and let me add Historians of the web are still marginal in media and communication history conferences and journals, a kind of neglect at route square.

However, I would like to conclude this contribution by citing at least three examples of possible areas of research that consider history and the web and could benefit from the combination of the three approaches I have identified.

Contemporary historians working on the 1990s (for example, the post-Cold War era, the Yugoslavia war, neoliberalism) can use digitized sources, websites, and interviews with web founders, and discuss the social role of the web in shaping those events and developments. Or, the W3C website (www.w3.org) could be studied as a multiple-layer source, discovering tensions and reasons for the W3C's formation, power dynamics between European (CERN) and American (MIT) academia, which of course predate the website itself, but the same website can be seen as a place of negotiations among different actors. In other terms, websites can be framed as repositories of political, social, and economic tensions as Andreas Fickers (2012) did in his seminal paper on the radio dial as a mediating interface on different political, cultural, and technical forms of imaginaries. Or, again, we could study the invention of the web at CERN archives, putting ourselves in the subfield of “History of the Web.” But then, to better understand what the collectors of the archive had in mind, questioning the reasons why they decided to print some web pages and preserve them (more in the field of an ironic analog “Web History”). And, finally, discussing with CERN on re-digitizing those websites and putting them on their website as a source for historians in the realm of the “History Web.”

Those are just three random examples where I hope it is clear that it is possible to start from one of the subfields and then expand it toward the others to enrich the main research questions, theoretical background, and sources to be used. Frankly, I think that creating a group of interdisciplinary scholars and combining three subfields into a unique field (called History *and* the Web?) could be a way to de-marginalize the historical study of the web. I am confident that this book will also contribute to this direction, so I hope that scholars and practitioners in all three subfields, as well as others, will read it.

Notes

- 1 Europeana (n.d.). *Europeana*. Web Archive. https://web.archive.org/web/20240000000000*/www.europeana.eu/en.
- 2 CERN Document Server (n.d.). *The World Wide Web (Archives)*. Web Archive. <https://web.archive.org/web/20230322092039/http://cds.cern.ch/collection/The%20World%20Wide%20Web%20%28Archives%29?ln=en>.
- 3 CERN Document Server (2011). *ARCH-WWW-4-010.pdf*. Web Archive. <https://web.archive.org/web/20240211210403/http://cds.cern.ch/record/1405411/files/ARCH-WWW-4-010.pdf>.
- 4 CERN (n.d.). *The World Wide Web project*. Web Archive. <https://web.archive.org/web/20240222051128/http://info.cern.ch/hypertext/WWW/TheProject.html>.
- 5 Web Design Museum (n.d.). *About us*. Web Archive. <https://web.archive.org/web/20240215001252/www.webdesignmuseum.org/about-us>.
- 6 Computer History Museum (n.d.). *The web revolution*. Web Archive. <https://web.archive.org/web/20231003093550/www.computerhistory.org/revolution/the-web/20/408>.
- 7 Pew Research Center (2014, March 11). *World Wide Web timeline*. Web Archive. <https://web.archive.org/web/20240214144626/www.pewresearch.org/internet/2014/03/11/world-wide-web-timeline/>.
- 8 The Guardian (2014, March 12). *The history of the World Wide Web: A timeline*. Web Archive. <https://web.archive.org/web/20221113055008/www.theguardian.com/technology/2014/mar/12/history-world-wide-web-timeline>.

- 9 The History of the Web (n.d.). *Timeline*. Web Archive. <https://web.archive.org/web/20240215135002/https://thehistoryoftheweb.com/timeline/>.
- 10 CERN (n.d.). *Timeline*. Web Archive. <https://web.archive.org/web/20231129212052/https://timeline.web.cern.ch/timeline-header/90>.

References

- Aasman, S., de Haan, T., & Teszelszky, K. (2019). Web archaeology: An introduction. *TMG Journal for Media History*, 22(1), 1–5. <https://doi.org/10.18146/tmg.433>.
- Allen, M. (2008). Web 2.0: An argument against convergence. *First Monday*, 13(3). <https://doi.org/10.5210/fm.v13i3.2139>.
- Ankerson, M. S. (2012). Writing web histories with an eye on the analog past. *New Media & Society*, 14(3), 384–400. <https://doi.org/10.1177/1461444811414834>.
- Balbi, G., & Magaudda, P. (2018). *A History of Digital Media. An Intermedia and Global Perspective*. Routledge.
- Barassi, V., & Treré, E. (2012). Does Web 3.0 come after Web 2.0? Deconstructing theoretical assumptions through practice. *New Media & Society*, 14(8), 1269–1285. <https://doi.org/10.1177/1461444812445878>.
- Ben-David, A. (2016). What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. *New Media & Society*, 18(7), 1103–1119. <https://doi.org/10.1177/1461444816643790>.
- Berners-Lee, T. (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. HarperSanFrancisco.
- Blaney, J., Winters, J., Milligan, S., & Steer, M. (2021). *Doing Digital History: A Beginner's Guide to Working with Text as Data*. Manchester University Press.
- Bory, P. (2020). *The Internet Myth. From the Internet Imaginary to Network Ideologies*. University of Westminster Press.
- Bory, P., Benecchi, E., & Balbi, G. (2016). How the Web was told: Continuity and change in the founding fathers' narratives on the origins of the World Wide Web. *New Media & Society*, 18(7), 1066–1087. <https://doi.org/10.1177/1461444816643788>.
- Brügger, N. (Ed.) (2010). *Web History*. Peter Lang.
- Brügger, N. (2011). Web archiving — between past, present, and future. In M. Consalvo and C. Ess (Eds.), *The Handbook of Internet Studies* (pp. 24–42). Blackwell Publishing.
- Brügger, N. (2012). Web history and the Web as a historical source. *Zeithistorische Forschungen — Studies in Contemporary History*, 9(2), 316–325. <https://doi.org/10.14765/zzf.dok-1588>.
- Brügger, N. (2013). Web historiography and internet studies: Challenges and perspectives. *New Media & Society*, 15(5), 752–764. <https://doi.org/10.1177/1461444812462852>.
- Brügger, N. (2018). *The Archived Web: Doing History in the Digital Age*. The MIT Press.
- Brügger, N. (2022). Media events in an age of the Web and television: Dayan and Katz revisited. *Nordic Journal of Media Studies*, 4(1), 37–55. <https://doi.org/10.2478/njms-2022-0003>.
- Brügger, N., & Milligan, I. (2019). *The SAGE Handbook of Web History*. SAGE.
- Brügger, N., Milligan, I., Ben-David, A., Gebeil, S., Nanni, F., Rogers, R., Turkel, W. J., Weber, M. S., & Webster, P. (2019). Internet histories and computational methods: A “round-doc” discussion. *Internet Histories*, 3(3–4), 202–222. <https://doi.org/10.1080/24701475.2019.1639352>.
- Burns, M., & Brügger, N. (Eds.) (2021). *Histories of Public Service Broadcasters on the Web*. Peter Lang.
- Cohen, D. J., & Rosenzweig, R. (2006). *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*. University of Pennsylvania Press.
- Edgerton, D. (2007). *The Shock of the Old: Technology and Global History Since 1900*. Oxford University Press.
- Fickers, A. (2012). Visibly audible: The radio dial as mediating interface. In T. Pinch & K. Bijsterveld (Eds.), *The Oxford Handbook of Sound Studies* (pp. 411–439). Oxford University Press.
- Fomasi, M., Barcella, D., Benecchi, E., & Balbi, G. (2023). Genealogy of an archive. The birth, construction, and development of the World Wide Web collection at CERN. *Internet Histories*, 7(3), 277–294. <https://doi.org/10.1080/24701475.2023.2238254>.
- Foot, K. A., & Schneider, S. M. (2006). *Web Campaigning*. The MIT Press.

“History web,” “web history,” and “history of the web”

- Galloway, S. (2017). *The Four: The Hidden DNA of Amazon, Apple, Facebook, and Google*. Portfolio, Penguin.
- Gebeil, S. (2019, April 1–4). Web archives as sources for the uses of the past during the 2000s. For a social, micro-historical and qualitative approach. In *Audiovisuals and Internet Archives: Histories of Healthy Bodies in the 21st Century*. ERC Body Capital, Spring School.
- Gillies, J., & Cailliau, R. (2000). *How the Web Was Born: The Story of the World Wide Web*. Oxford University Press.
- Goggin, G., & McLelland, M. (Eds.) (2017). *Routledge Companion to Global Internet Histories*. Routledge.
- Gomes, D., Demidova, E., Winters, J., & Risse, T. (2021). *The Past Web: Exploring Web Archives*. Springer.
- Helmond, A. (2015). The platformization of the web: Making web data platform ready. *Social Media+ Society*, 1(2). <https://doi.org/10.1177/2056305115603080>.
- Kreps, D., & Kimppa, K. (2015). Theorising Web 3.0: ICTs in a changing society. *Information Technology & People*, 28(4), 726–741. <https://doi.org/10.1108/ITP-09-2015-0223>.
- Milligan, I. (2019). *History in the Age of Abundance? How the Web Is Transforming Historical Research*. McGill-Queen's University Press.
- Mohr, G. (2007). Archival tools to match the web: Open, international, comprehensive. In D. H. L. Goh, T. H. Cao, I. T. Sølberg, & E. Rasmussen (Eds.), *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*. ICADL 2007. Lecture Notes in Computer Science (vol. 4822). Springer. https://doi.org/10.1007/978-3-540-77094-7_3.
- Morley, D. (2009). For a materialist, non-media-centric media studies. *Television & New Media*, 10(1), 114–116. <https://doi.org/10.1177/1527476408327173>.
- Musiani, F., Paloque-Bergès, C., Schafer, V., & Thierry, B. (2019). *Qu'est-ce qu'une archive du Web?* Open Editions Book.
- Natale, S., & Bory, P. (2017). Constructing the biographies of the web: An examination of the narratives and myths around the web's history. In N. Bruggen (Ed.), *Web 25: Histories from the First 25 Years of the World Wide Web* (pp. 29–42). Peter Lang.
- Petras, V., Hill, T., Stiller, J., & Gäde, M. (2017). Europeana — a search engine for digitised cultural heritage material. *Datenbank Spektrum* 17, 41–46. <https://doi.org/10.1007/s13222-016-0238-1>.
- Rogers, R. (2017). Doing Web history with the Internet Archive: Screencast documentaries. *Internet Histories*, 1(1–2), 160–172. <https://doi.org/10.1080/24701475.2017.1307542>.
- Schafer, V. (2019). De la Wayback Machine à la bibliothèque: les différentes saveurs des archives du Web. *Gazette des archives*, 253(1), 95–107.
- Schafer, V., & Thierry, B. G. (2016). The “Web of pros” in the 1990s: The professional acclimation of the World Wide Web in France. *New Media & Society*, 18(7), 1143–1158. <https://doi.org/10.1177/1461444816643792>.
- Schafer, V., & Thierry, B. G. (2018). The 90s as a turning decade for Internet and the Web. *Internet Histories*, 2(3–4), 225–229. <https://doi.org/10.1080/24701475.2018.1521060>.
- Schafer, V., & Winters, J. (2021). The values of web archives. *International Journal of Digital Humanities*, 2, 129–144. <https://doi.org/10.1007/s42803-021-00037-0>.
- van Dijck, J. (2013). *The Culture of Connectivity: A Critical History of Social Media*. Oxford University Press.
- Vossen, G., & Hagemann, S. (2007). From version 1.0 to version 2.0: A brief history of the web. In J. Becker, K. Backhaus, H. L. Grob, T. Hoeren, S. Klein, H. Kuchen, U. Müller-Funk, U. W. Thonemann, & G. Vossen (Eds.), *European Research Center for Information Systems (ERCIS)*. ERCIS Working Papers 4. University of Münster.
- Wyatt, S. (2021). Metaphors in critical Internet and digital media studies. *New Media & Society*, 23(2), 406–416. <https://doi.org/10.1177/1461444820929324>.

PART I

Entire national web domains from a transnational perspective

3

ICONOGRAPHY IN FLUX

A transnational exploration of the evolution of climate news imagery through the Wayback Machine

Anat Ben-David, Adam Amram, and Noa Turgeman

Introduction

Web archives are increasingly being recognised as invaluable information repositories, offering insights into the evolving dynamics of global communication. As the study of transnational web histories gains momentum, scholars are increasingly tasked with devising methodological approaches that effectively capture the multifaceted nature of the online phenomena, looking simultaneously at the evolution of the web as a medium, the evolution of the mediation of social–historical processes, and how such processes evolve differently across the globe.

This chapter contributes to such efforts by proposing a methodological framework for conducting large-scale transnational visual historiography using the Internet Archive, a relatively understudied area of web archive research (Amram et al., 2018).

Building upon previous research, we advocate for an inclusive approach that transcends traditional geographic boundaries and carefully incorporates primary sources from the global North and South (Goggin & McLelland, 2017). Our approach further contemplates questions related to critical comparative web archive research design (Ben-David, 2021), given that comparing archived web materials across different countries and regions requires being mindful of the divergent historical contexts, cultural nuances, and linguistic variations inherent in these sources.

We demonstrate our approach by narrating a transnational visual history of online news images of the global climate crisis. Utilising the Internet Archive’s Wayback Machine as a primary source that allows for conducting comparative historical research across countries and cultures, we endeavour to unravel the intricate transnational narratives that underpin the visual representation of climate news, shedding light on how news images have evolved to shape the transnational and local imaginaries of the current climate crisis. In so doing, we engage in multifaceted historiography, looking simultaneously at the evolution of online news outlets (the web as a medium), the evolution of the mediation of a social process (the editorial choices of news websites about the climate crisis), and how such processes evolve in different countries and regions (How did visual frames of the climate crisis evolve? How do

climate news images in various countries’ travel? Which images are iconised transnationally, and what visual frames remain local?).

Accordingly, the chapter begins by reviewing iconography’s important role in mediating the climate crisis of past decades. Next, we delve into the existing literature concerning the historical role of news media in shaping collective perceptions of the climate crisis through visual representations. Subsequently, we introduce our method for utilising the Internet Archive Wayback Machine to conduct a comprehensive transnational visual analysis of the progression of images featured in the climate sections of news websites across 30 countries from 2010 to 2022. We discuss our findings by asking how longitudinal and transnational analyses of the climate crisis help unravel competing and hidden global and local imaginaries.

Climate change iconography

In November 2022, the *Guardian* published an article featuring 20 journalistic photographs that became iconic in representing the current climate crisis (Schwarz, 2022). Most images displayed natural disasters or extreme weather conditions, such as droughts, floodings, melting icebergs, hurricanes, and wildfires.

What makes a climate image iconic? From the *Guardian*’s journalistic perspective, the images depict dramatic and highly aestheticised scenes, such as a bird’s-eye view of a circle of carcasses of giraffes that perished in a drought in Kenya or a devastated Greek woman, clenching her heart and gasping in agony in the front of the picture, while gigantic flames of a forest fire are raging in the background, threatening to burn down her house. The images capture a newsworthy moment, but the frame tells a powerful story whose symbolism extends beyond what the eye can see. The Greek woman image, for example, is iconic not only because of the concrete damage and danger it depicts but also because it can capture the collective imagination by becoming a symbol of the danger of forest fires. The giraffe carcasses — so perfectly arranged in a circle on brown, cracked soil — symbolise the dangers of global warming to biodiversity.

Over recent years, a growing scholarly focus has been on scrutinising and critically assessing the social and political implications communicated through visual representations (Purgar, 2014). Such scholarly focus has led to the “iconic turn” in cultural studies (Bachmann-Medick, 2016), signifying a heightened epistemological awareness of images in culture in ways that investigate images’ inherent meaning, power, and influence on shaping collective imagination and knowledge. Smith states that iconisation is the process of images becoming widely recognisable, raising proactive emotions, and acting as “operative forces” in a socio-political reality (Smith, 2021, p. 181). Smith further argues that digital images exhibit accelerated repetition as they are constantly duplicated, repeated, reproduced, related, and networked with similar images. Such “networkedness of images” (Smith, 2021, p. 172) creates a “visual order which organizes seeing in a particular way” (Smith, 2021, p. 175).

The first visual icon that evoked a connection to the environment and the emerging climate crisis is the “blue marble,” the image of planet Earth taken from space by the Apollo 17 space shuttle in 1972. Sheila Jasanoff (2001) showed how the “blue marble” symbol transformed global human consciousness, serving both as an environmental symbol revealing the planet’s vulnerability and as a capitalist symbol used by corporations to maintain existing hegemonic forces.

Iconography in flux

Since then, iconic images have always accompanied climate-change discourse. Schneider and Nocke (2014) describe the emergence of climate change aesthetics by creating several iconic images, such as the polar bear and melting icebergs. These images joined an assemblage of a collective visual “cultural memory,” indicating the tactics and politics of climate change, thereby adding to the epistemological status of the images. These “socio-semiotic systems” are widely used to communicate and comprehend different meanings and impacts of climate change. Schneider and Nocke further describe how “similar images start to migrate from one sphere to another. They might keep their basic gestalt (immutable), but as picture agents, they can serve various interest groups, trigger different associations, and offer new perspectives” (2014, p. 17). One widely researched example of visual iconisation of climate change is the polar bear (Schneider & Nocke, 2014; Born, 2019; O’Neill, 2022). According to Born (2019), the image of the polar bear standing on a melting iceberg established its iconic status and sustained its popularity throughout the years because of its inherent ability to enable a wide discourse and links between nature, culture, science, and politics. The bear “fosters an individualized and localized framing of climate change based on emotions and the anthropomorphisation of nature” (2019, p. 650). The persistent iconic status of the image is assigned to it becoming normalised as a climate imaginary (O’Neill, 2022). However, such imaginaries are contested, may carry multiple messages, and also compete with one another (Grevsmühl, 2016). Levy and Spicer (2013) further describe the political economy of the tactical use of climate images as a competition on imaginaries around the production, consumption, and prioritisation of environmental and cultural values. Similarly, Smith (2022) coined the term iconomy — the economy of icons — to describe the reduction of issues, conflicts, and wars to the confrontation and competition of icons.

In this chapter, we focus on the iconomy of climate images as they are mediated through online news websites worldwide. The iconic status of news images, such as those published by the *Guardian* and described at the outset of this chapter relies on a repetitive and longitudinal circulation of images and their sustained recognition among large publics over time. Iconic news images, therefore, gain power as frameworks of collective visual memory (Boudana et al., 2023).

Globalisation plays a pivotal role in shaping the perspectives of news consumers. In his genealogy of world pictures, Mitchell (2007) provides a valuable framework for understanding the globalisation of news images, their formation, and their dissemination on a global scale. Mitchell’s notion of “region” and “regionalism” become relevant in analysing the use of climate images based on nationality or geographical space. It highlights that regionality encompasses geographical boundaries and cultural, social, and ethnographic regions, such as “the Arab World.” Additionally, the economy of similarities of news images is constructed and sustained by the ownership of global news agencies and the availability of global image resources, such as image banks, where images are bought cheaply and used repeatedly (Machin, 2004; O’Neill & Smith, 2014).

Building on the literature on the political economy of the construction and contestation of climate imaginaries through iconomy, the empirical section of this chapter traces the temporality and spatial spread of climate images, looking to identify how certain icons gain collective awareness as they travel across countries and regions over time, while other climate images remain bounded in local contexts. Accordingly, the following section of the chapter reviews the literature on the history of visual climate images in news coverage.

The history of visual climate news coverage

The development of shared cultural memories is often seen as an enduring process, heavily influenced by the media's role in crafting and shaping them (Rigney, 2005). Rigney views cultural memory as “working memory,” formed and reshaped through public actions and adapting through cultural mechanisms, with the media playing a pivotal role in this evolution. Similarly, recent studies pointed to the visual communication of climate change in news media in shaping cultural memories and perceptions of the climate crisis. The literature highlights the power of images in creating familiarity, establishing a connection to collective memory, attracting attention, making climate change visible, and influencing public opinion and engagement (Wozniak et al., 2015; Schäfer, 2020; Rebich-Hespanha & Rice, 2016; O'Neill & Smith, 2014).

Moreover, the dominance of certain visual representations and their ability to shape public agenda and involvement are crucial in understanding climate change news reporting (Rebich-Hespanha et al., 2015). Climate visual frames determine what is visible to the public and what remains invisible or marginalised (O'Neill & Smith, 2014). O'Neill (2014) emphasises the importance of investigating media ownership and interests to understand how visual climate discourse is constructed and shaped in a particular way.

Previous studies analysed extensive corpora of climate news images by extracting images from (scanned) print and online news websites (Rebich-Hespanha et al., 2015; Wozniak et al., 2015; O'Neill & Smith, 2014) in different time ranges and geographies. For example, DiFrancesco and Young (2011) focused on Canadian print newspapers, where they carefully examined images for denotative representations of humans, nature, industry, and technology. Employing intertextual discourse analysis, they aimed to uncover hidden meanings within recurring themes such as politicians, the oil industry, polar bears, and refugees.

Similarly, O'Neill (2013) conducted a comparative study of 1600 images from Australia, the U.K., and the U.S. and analysed the political and cultural contexts surrounding the images. The dominant themes that emerged included depictions of people (particularly politicians), climate change's geographical or personal impact, protests, technological solutions, science, and weather-related visuals.

Wozniak et al. (2015) undertook a broader comparative approach by analysing images and texts from newspaper articles that covered the U.N. Climate Change Conferences in Germany, India, and South Africa. The researchers identified four main themes consistently present across the countries: global warming victims, civil society demands, political negotiations, and promoting sustainable energy. Interestingly, they noted a strong cross-national similarity in image frames, attributed to the global nature of the U.N. summits and the utilisation of global news image agencies like A.P., Reuters, and Getty Images.

Rebich-Hespanha and colleagues (2015) conducted an extensive longitudinal analysis of climate change news images in the U.S., spanning 40 years from 1969 to 2009. The researcher identified dominant frames, such as depictions of government and politics, the role of science and scientists, and the impact on the environment, such as melting ice, polar bears, and penguins. Comparative analysis also revealed a correlation between representations of geographic regions and dominant image frames. The study found that climate change was reported inclusively, focusing on images primarily from the U.S., particularly California for solar-related imagery, and Europe. However, Asia, Africa, and Latin America were largely absent or minimally represented.

Iconography in flux

Building upon these analyses, O'Neill and Smith (2014) identified time, truth, and power as the three main themes that emerged from the encoding/decoding communication cycle of climate change's visually mediated interactions. Schäfer (2020) found that news visualisation of climate change in different countries has strong similarities, while images of the consequences of climate change and well-known people are more prominent than others, and some types of imagery are more prominent than others. Additionally, frequently used images are infographics (such as temperature curves), representations of the causes of climate change, and untouched nature such as landscapes, glaciers, or icebergs. Untouched nature images complement a written text describing possible threats to nature by climate change (Rebich-Hespanha et al., 2015). Images of options for action, such as energy-saving products, are less frequently used. Similarities are found locally and globally between different countries on different continents. Such homogenisation in visual framing is explained by the hegemony of U.N. summits and organisations in dictating the climate discourse, as well as by global corporate ownership of news media companies and the news agencies' sources of information (Schäfer, 2020). Schäfer also highlights the increasing use of images in the news coverage of climate change as well as the power and ability of the image and its profound influence in this media, concluding that "it is surprising that only a few studies have tried to assess the extent of visualisation in climate change reporting" (2020, p 147).

Case study: A transnational analysis of archived climate news images, 2010–2022

One of the motivations guiding this research is to set a research agenda for conducting large-scale comparative (transnational and longitudinal) analysis of images using the Internet Archive Wayback Machine. As such, our case study aims to provide a general framework for selecting units for analysis and comparison, along with guidelines for the technical aspects involved in accessing, processing, and analysing images from the Internet Archive Wayback Machine. Previous research has delved into the methodological and epistemic challenges of large-scale historiographical research utilising the Wayback Machine. These challenges encompass various aspects that necessitate careful consideration for accurate and reliable analysis.

One significant concern in previous studies is the presence of known biases in the archival coverage of certain country-code top-level domains (ccTLDs) (Thelwall & Vaughan, 2004). Another significant challenge pertains to temporal inconsistencies in the replay of archived web pages, specifically regarding the temporal metadata of embedded elements within a web page, such as images. These embedded elements may exhibit timestamps that are months or even years earlier or later than the date and time when the web page itself was captured (Ainsworth et al., 2014). Additionally, the complexity of tracing the provenance of archived snapshots poses another challenge for large-scale historiographical research. A previous study showed that historical facts on the Wayback Machine are constructed through an entangled temporal and geographical assemblage of actors and technologies (Ben-David & Amram, 2018). Such entanglements make it challenging to determine the exact origin or context of individual snapshots and, accordingly, to evaluate their evidential value as primary sources.

Despite these challenges, the accessibility and global scale of archival content on the Wayback Machine make it a valuable source for comparative historiographical research. Our case study, accordingly, combines large-scale image mining from the Wayback Machine with an analysis that is mindful of the known epistemic limits raised above.

To simplify the complexity of the archived web as an object of study, we follow Brügger's (2009) typology of various web elements as units of analysis for conducting historical web research and focusing solely on archived images. On the one hand, extracting images from various news websites strips them of their textual context — the article in which they were initially embedded. Alternatively, the systematic collection and analysis of all images from websites that share a similar organising principle (in our case, the climate sections of news websites) allow for various comparative and temporal analyses of the cultural shaping of climate images across countries and over time. Such analyses are informed by the “cultural analytics” approach that identifies large-scale cultural trends through visual analysis of Instagram, claiming these are mass archives of global digital cultures (Manovich, 2009). A series of studies demonstrated the application of the cultural analytics approach on large-scale transnational social media data, such as the visual signatures of cities around the globe (Hochman & Manovich, 2013) and global cultural rhythms (Hochman & Schwartz, 2012). “Big cultural data” can be studied using computational analysis and methods (Manovich, 2020), empirical, large-scale data mining and analyses (Hochman & Manovich, 2013; Hochman & Schwartz, 2012), and qualitative analyses of manually selected data that look into the visual and textual elements of images. Rogers (2021) coined the term “meta pictures” — sets of online image collections that are created through querying techniques (used on digital platforms) as a methodology for studying images in groups (Colombo, 2019). Rogers developed the study of “style spaces” created through grouping images based on their similarity, either manually by visually analysing and categorising images that share common visual characteristics or themes or by using computer vision tools to cluster automatically and group similar images based on various visual features, such as colour, composition, or content.

Furthermore, previous research on visual methodologies has sought to develop new (often mixed-methods) approaches for analysing and interpreting the socio-political meaning of networked digital images (Pearce et al., 2018). Wevers and Smits (2020) assessed the effectiveness of machine vision algorithms, particularly convolutional neural networks (CNNs), in analysing historical visual sources. They emphasised that visual analysis may benefit historical research on visual materials, expanding the scope of digital humanities methods. By employing computer vision techniques on images from Dutch print news media, the study allowed for the tracking and analysis of specific visual representations within the archive across extensive timeframes.

Our case study follows these approaches. In addition to clustering images based on similarity, we cluster and analyse images based on the source, looking to identify “native” local and global, transnational, and transcultural similarities and differences. The comparative analysis includes a scan of “collages” of images clustered together, a close read, and a qualitative semiotic analysis of single images. In the following, we outline the methodological steps for our case study.

Method

Data extraction

To create our corpus, we identified online news websites with a designated climate section (for example, www.lemonde.fr/climat/). We systematically queried search engines for the term [“Climate News Section” + “Country Name”] in each country's local language, attempting to locate as many sources as possible while focusing on balancing sources from

Iconography in flux

Table 3.1 The sources selected for analysis

<i>News Website</i>	<i>Country/Region</i>	<i>Number of Images</i>	<i>Oldest Snapshot</i>
<i>Al Jazeera</i>	Qatar/“Arab World”	179	2021
<i>Asia News</i>	Asia	168	2017
<i>BBC</i>	UK/International	338	2021
<i>Bloomberg</i>	US/International	544	2019
<i>China Org</i>	China	329	2010
<i>El Pais</i>	Spain	381	2021
<i>El Tiempo</i>	Colombia	206	2014
<i>Folha</i>	Brazil	1298	2010
<i>Globe and Mail</i>	Canada	532	2020
<i>Haaretz</i>	Israel	268	2018
<i>Hindu Times</i>	India	450	2010
<i>Hungary Today</i>	Hungary	306	2019
<i>Japan Times</i>	Japan	799	2014
<i>Jeune Afrique</i>	“Pan African”	921	2015
<i>La Nacion</i>	Argentina	617	2019
<i>La Tercera Chile</i>	Chile	356	2016
<i>Le Monde</i>	France	3522	2012
<i>Mail And Guardian</i>	South Africa	103	2010
<i>Naharnet</i>	Lebanon	462	2015
<i>New York Times</i>	U.S./International	1719	2017
<i>Süddeutsche Zeitung</i>	Germany	3113	2010
<i>Sydney Morning Herald</i>	Australia	777	2010
<i>TASS</i>	Russia	1265	2014
<i>Wyborcza</i>	Poland	463	2021
Total		19116	

both the global North and South. We filtered the search results based on the following criteria to select sources for the corpus: (1) the news site has a designated climate section URL within a one-hop distance from the homepage; (2) the news source is published in the local language, except in countries where only English language sources were available. The final corpus includes 24 sources (see Table 3.1).

Our corpus includes established and internationally read news sources such as the *New York Times*, The *BBC*, or *Al Jazeera*, and news websites that attract local and regional audiences, such as *Folha* in Brazil and the Pan-African *Jeune Afrique*. Although our corpus does not cover all countries, it represents a balanced collection of sources from all continents in ways that allow transnational comparison.

Using Python, we accessed the Wayback Machine’s CDX server and extracted the URLs of each source’s snapshots, filtering for one snapshot per month. Subsequently, we extracted all image URLs from these snapshots to a database. Overall, we collected 29,674 URLs from Wayback Machine, of which 4670 URLs did not lead to images. 4532 images were not archived. The remaining 138 images returned other error messages and could not be read. To facilitate analysis, we decided to include only one source per country. In cases where there were two sources in one country, we did not include the source that featured fewer images (*Onet*, Poland, *The Age*, Australia, *Clarín*, Argentina, *The Times of India*, and *Deutsche Welle*, Germany).

Additionally, we removed sources with significantly fewer images (*Kathemarini* from Greece, *Jordan News*, and *Republica* from Peru). The start years of climate sections varied widely among sources, with some having dedicated URLs for climate issues as early as 2000 while most began these sections after 2010, we excluded snapshots obtained before 2010 to ensure consistency. The remaining dataset contains 19,116 images.

The images in the dataset are not distributed equally across sources and years. The number of images from 2019 to 2022 accounts for almost half of the images in the dataset (see Figure 3.1). We cannot explain at this point whether this is a “web historical” phenomenon — that is, that online news websites have gradually become more “visual,” or rather an “archival artifact,” indicating the improvement of image archiving by the Internet Archive Wayback Machine.

As is evident in Figure 3.2, some veteran sources, such as the *Sydney Morning Herald* and *Le Monde*, contributed relatively more images than sources that created designated climate sections in later years, such as the *El País*, *BBC*, and *Reuters*. Given the skewed distribution, the analysis below normalises the number of images relative to the total number of images per year and per source.

Image clustering

We further used Python to create collages of images, grouping them by year and source. Further, we automatically classified the images based on their visual qualities using the deep CNN VGG-16 (Simonyan & Zisserman, 2014), which was previously found effective in detecting species in images (Hridayami et al., 2019). We used VGG-16 to predict (extract) image features. Thereafter, we used principal component analysis, a statistical method for simplifying complex data by finding patterns and relationships among variables to reduce the dimensions of the image features. Subsequently, we used K-means clustering, a statistical method that identifies similar data points and groups them into clusters, to group the images into 10 clusters per year (Amram et al., 2018).

Image coding and labelling

Two researchers qualitatively examined the image clusters in the yearly collages. Per image, we coded its object (e.g., Wind Turbine) and subject (e.g., Clean Energy). We coded 2844 images from all years, approximately 11% of the dataset. We then joined the lists of coded objects and removed duplicates to create a list of 244 labels. We used automatic label prediction based on OpenAI Contrastive Language-Image Pretraining (CLIP) to code the remaining data. This publicly available neural network understands both images and natural language, allowing it to automatically assign a list of pre-given labels to a set of images (Radford et al., 2021).¹ Using our list of 244 labels, we used CLIP to predict the image objects in the remaining dataset. The results returned label probabilities for each object. We only kept the top label for each image after qualitatively inspecting the predictions and evaluating their accuracy against our coding.

Findings

Our findings indicate dynamic shifts in the iconography that depict the climate crisis over time. By 2015, the early iconic frames depicting CO₂ pollution, climate change (satellite

Iconography in flux

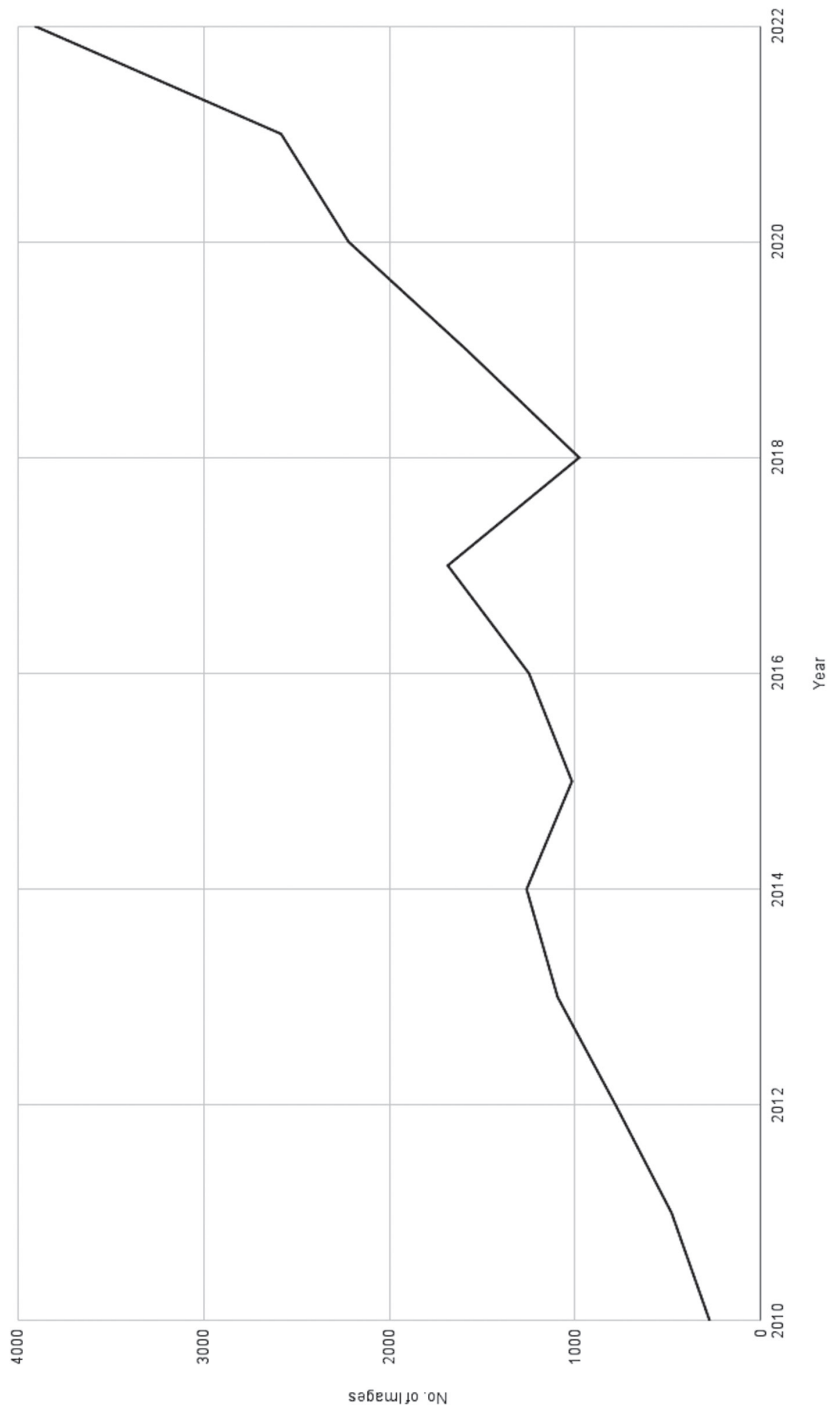


Figure 3.1 Number of images per year.

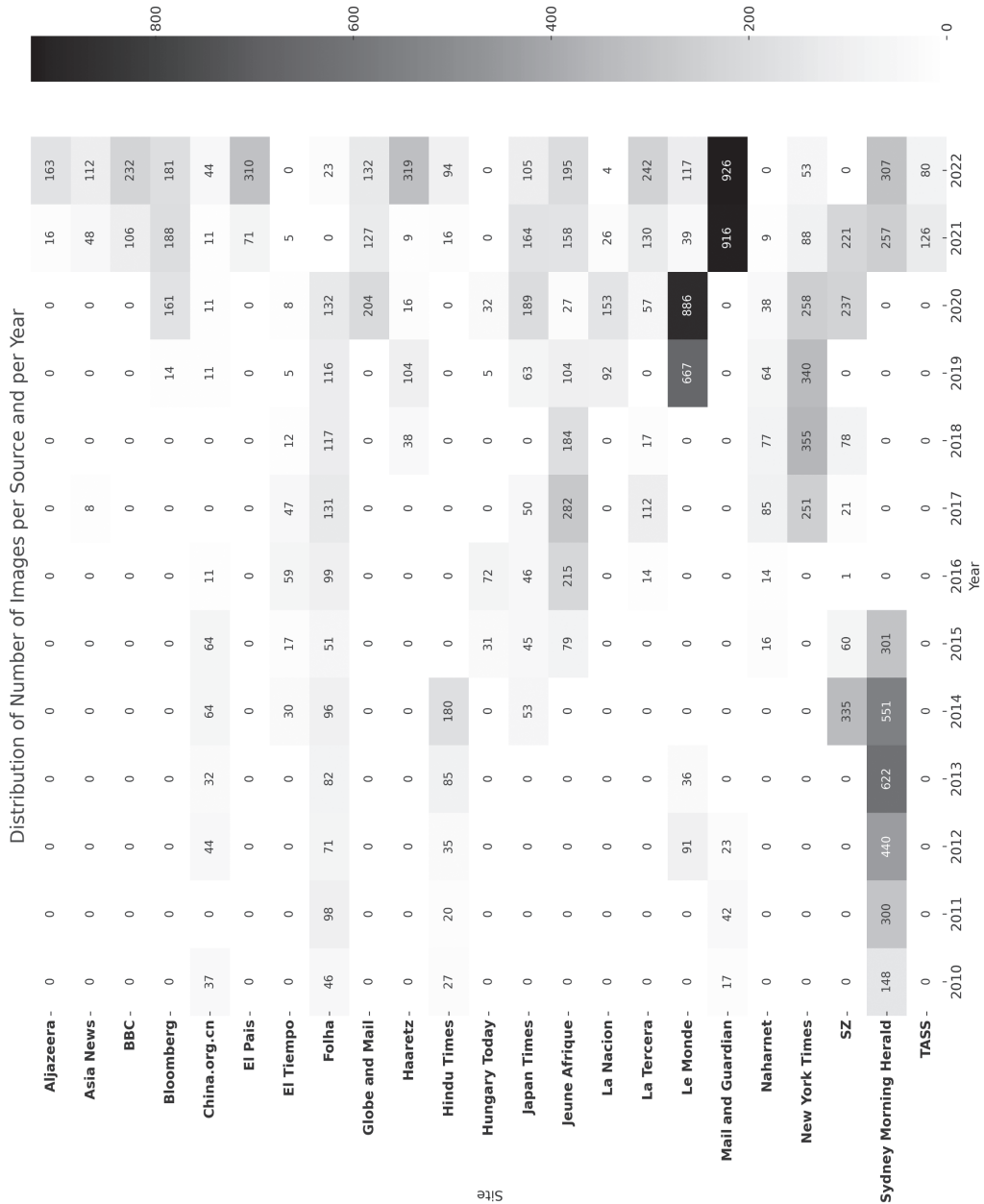


Figure 3.2 Number of images per source and per year.

Iconography in flux

images, melting icebergs, droughts), and climate solutions (solar panels) gradually declined, while new emphasis was placed on human-centric images of politicians, protests, refugees, workers, and speeches. At the same time, we notice an increase in the prevalence of images of concrete climate disasters such as floods, deforestation, and oil and gas rigs. Except for elephants, a popular image object between 2010 and 2012, animals are not included in the top 25 image labels, a point to which we return.

Looking at the distribution of the top 25 image labels of the corpus per source, we notice differences in topical focus across countries and regions throughout the examined period. For example, *Bloomberg* accounts for 26% of the satellite images, *Hungary Today* accounts for 46% of the images of speeches, and 33% of the images about temperature originate from the *South African Mail* and *Globe* (see Figure 3.3).

To further identify topical patterns across countries and regions, we grouped the image labels predicted by Open AI CLIP into the 17 visual frames of climate coverage described in previous research by Rebich-Haspanha et al. (2015) (see Figure 3.4). Such grouping confirms that the visual frame of Nature at Risk (Landscapes and Animals) declines over time (from 27% of the images in 2010 to 14% in 2022), compared to a sharp increase in the visual frame of disastrous and vulnerable landscape (from 6% in 2010 to 13% in 2022), indicating perhaps the transition from metaphorical concerns about possible implications of the climate crisis to concrete coverage of the actualisation of the crisis, such as extreme weather conditions, fires and floods. We further note the doubling of the visual frame of industry and pollution (from 7% in 2010 to 15% in 2022). It is worthwhile noting that the iconic frame of global warming through images of melting ice, glaciers, polar bears, and penguins is steady but relatively rare (ranging between 2% and 7% of the images annually). Similarly, images of the Earth, or the iconic “Blue Marble,” described at the outset of this chapter, are rarely found (ranging between 1% and 4% of the images each year) (see Figure 3.5).

Such a broad overview of the dynamic shifts in the visual frames of the climate crisis by online newspapers invites further reflection on the transnational aspects of the iconography of climate news images. Indeed, examining the distribution of the visual frames per source, we notice framing differences across countries and regions. For example, 40% of the images categorised in the Nature at Risk, Landscape, or Animals frame originate from India, compared to 10% from Germany and 2% from *Al Jazeera*. Hungary contributes the most images to the Politicians’ frame, while Anglo-Saxon countries (Australia, the UK, Canada, and the U.S.) use the Clean Energy frame more frequently. Similarly, although all sources dedicate a significant portion of their climate images to the frame of Industry and Pollution, the countries that stand out in this category are highly industrialised or invested in the Oil industry, such as Germany, Chile, Russia, and Japan (23%, 21%, 20%, and 18%, respectively) (see Figure 3.6).

We previously noted a decline in animal images for covering the climate crisis. From a documentary perspective, such a decline corresponds with the global decline in biodiversity (69%, according to the Living Planet Index). Some sources, including *Le Monde*, *Folha*, and the *Sydney Morning Herald*, use animal images more frequently. Tracing the placement of the iconic polar bear in the data is of particular interest. Although compared to other animals, the number of polar bear images is the highest (in terms of absolute numbers, before normalisation), we notice differences in the geographic distribution of the sources that use it. Sources repeatedly using the polar bear visual frame are *TASS*, Russia, the *SZ*, Germany, *Sydney Morning Herald*, Australia, and *La Nacion*, Columbia.

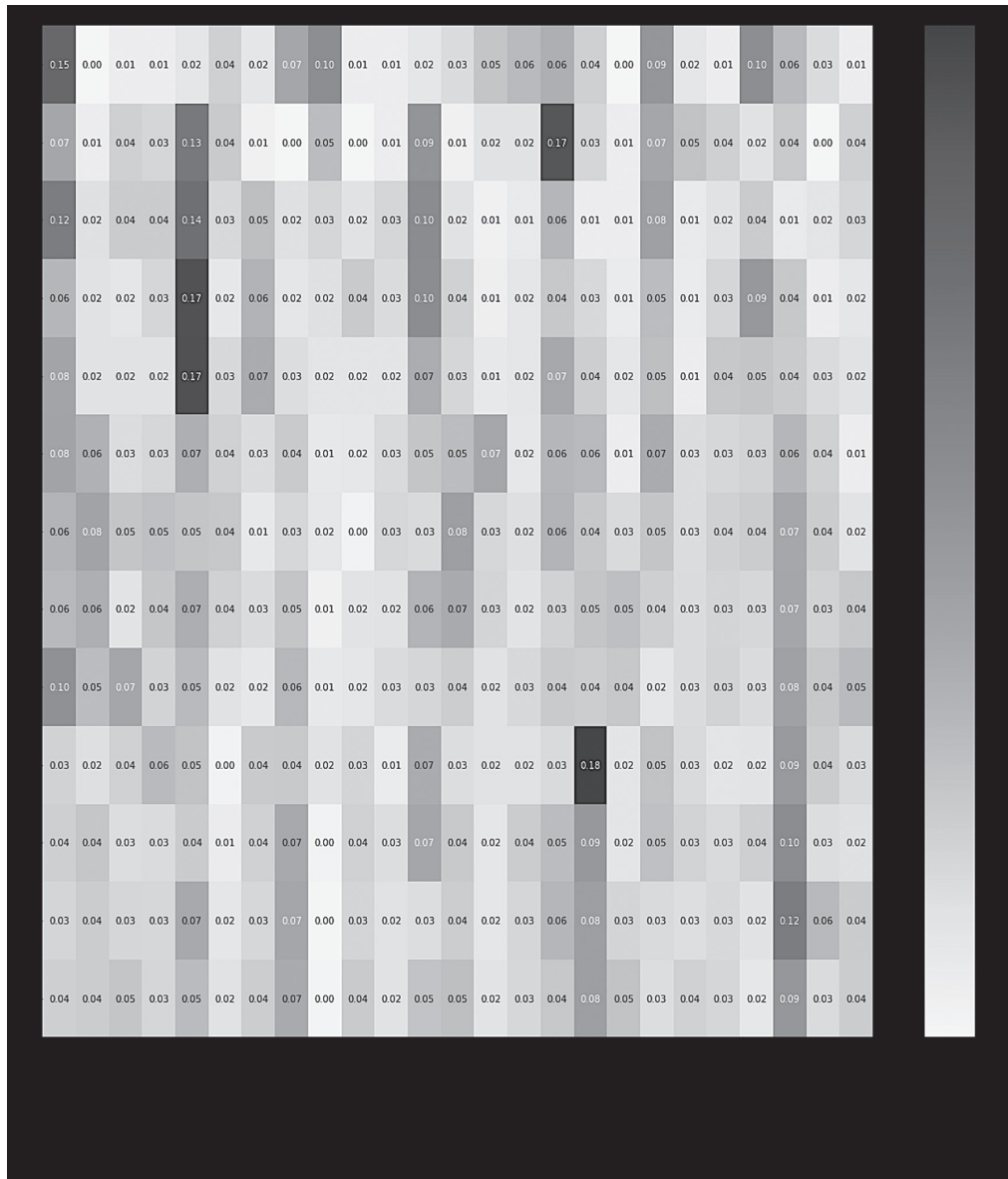


Figure 3.3 The distribution of the top 25 object labels per source (normalised proportions) (colour version in Zenodo repository).

Conversely, it is the least represented in countries such as China, Argentina, Brazil, and India. However, the analysis of the representation of images of animals in the dataset makes a case for the necessity of employing a transnational perspective when engaging in comparative web archive research. We report that the proportion of images of elephants (0.80) is higher than images of polar bears (0.77). However, unlike the shared imaginaries of the polar bears



Source: Adapted from Rebich-Hespanha et al. (2015).

It is interesting to note the frequent use of the visual frames of cows and livestock as bridging between the framing of the impact of the climate crisis on animals and between the consequences of livestock methane emissions that contribute to global warming. Thus, images of cows are found globally, and their frequency increases over time.

41

The Routledge Companion to Transnational Web Archive Studies

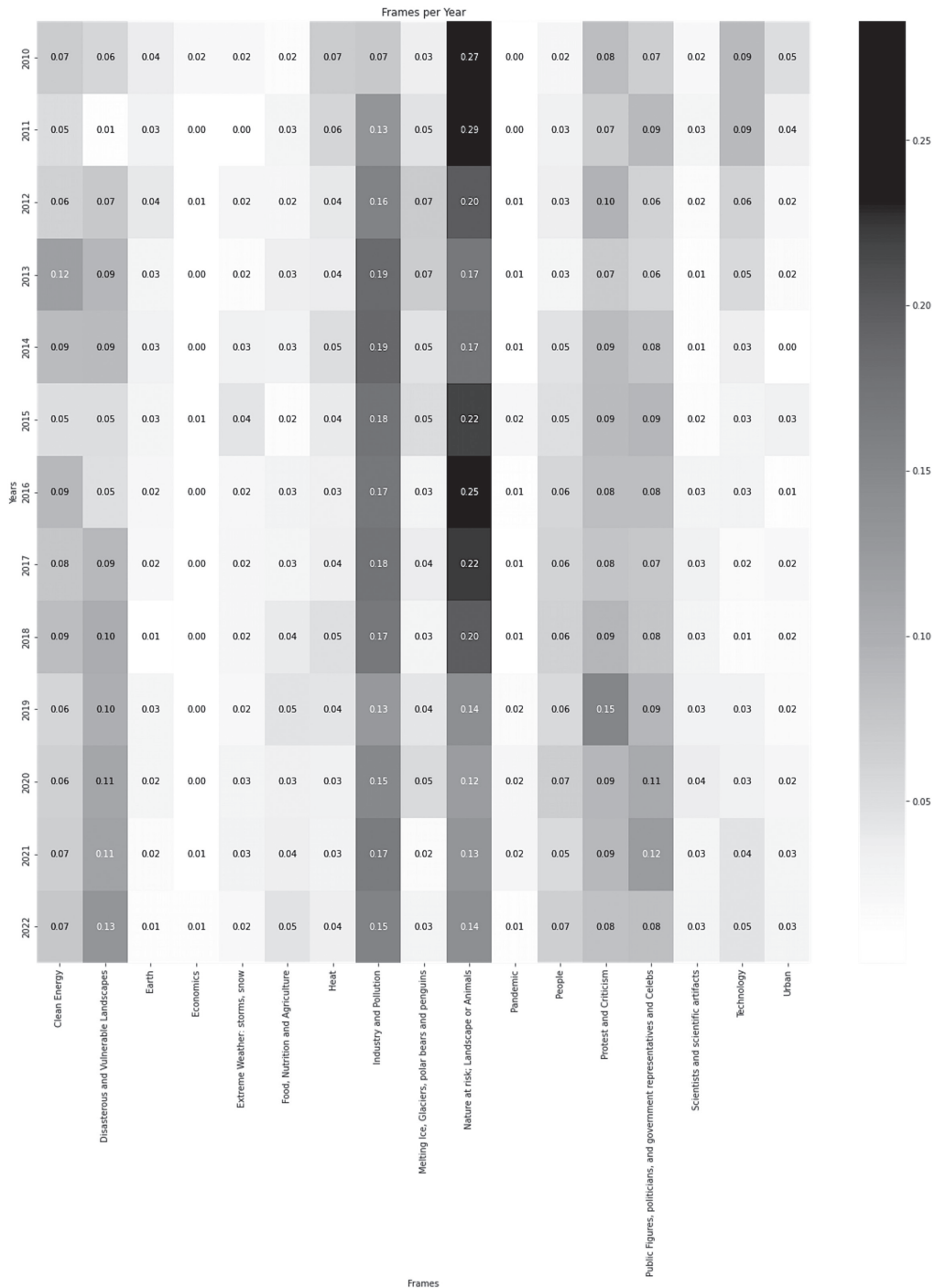


Figure 3.5 The distribution of visual frames per year (normalised proportions) (colour version in Zenodo repository).

Iconography in flux

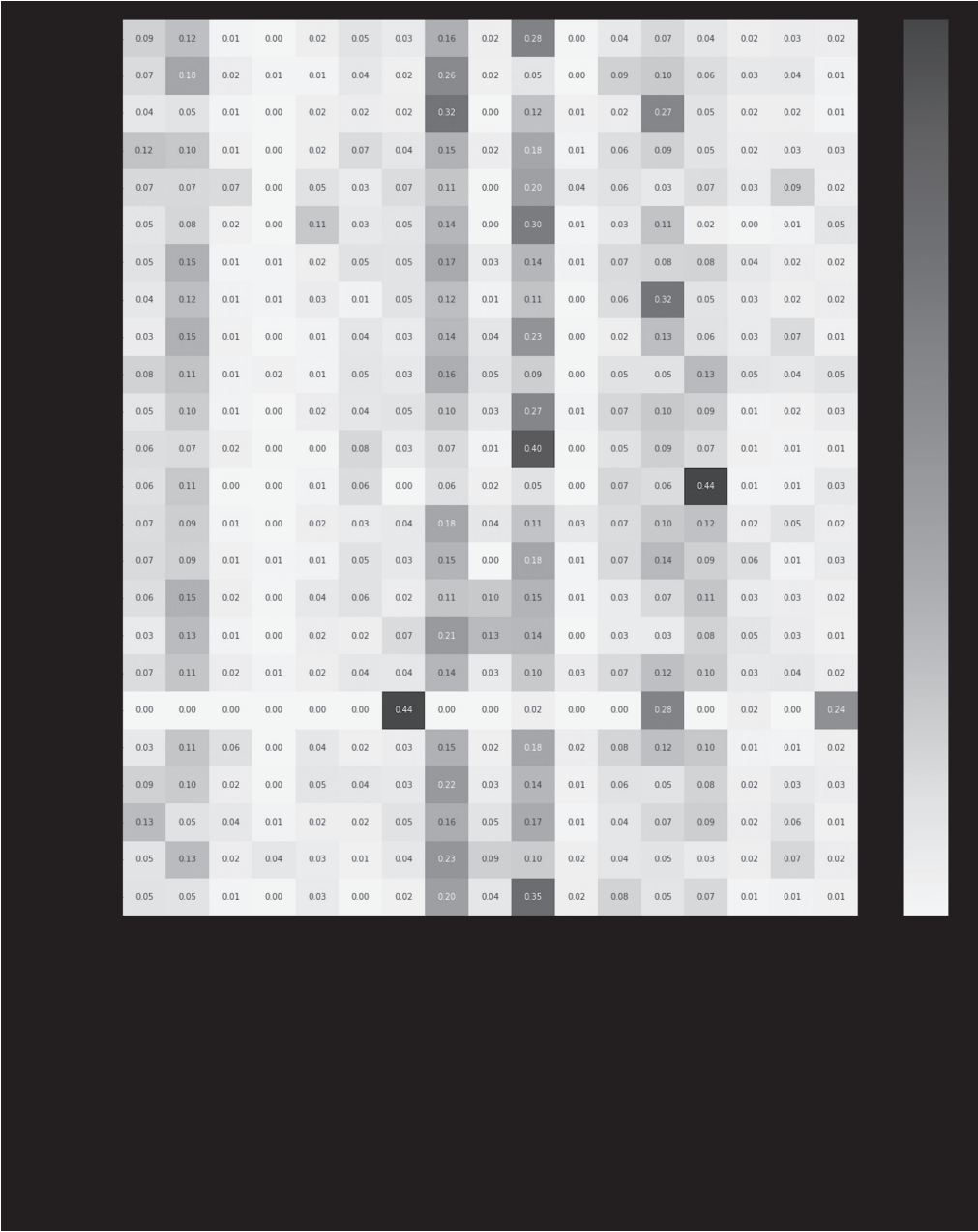


Figure 3.6 The distribution of visual frames per source (normalised proportions) (colour version in Zenodo repository).

Cluster analysis

How have the visual characteristics of the online news climate images changed over time? A zoomed-in analysis of the annual clustering of the images based on their visual similarity provides several answers to this question.

A cluster featuring saturated, orange-tinted images next to blue-tinted images is found repetitively in the dataset throughout the years. The dominant visual frame in this cluster is a setting sun, denoting global warming, rising temperatures, or heat. Characteristic, blue-tinted images in this cluster are wind turbines photographed against a background of bright blue skies and smoking factory chimneys photographed in daylight, melting icebergs, oceans, and oceanic animals such as whales and sharks. Despite the increase in the number of images in the corpus, the size of this cluster is shrinking over time until it dissolves in 2019.

Instead, the clusters depicting people — particularly politicians — are expanding. The cluster of politicians' headshots in the dataset increased from 117 images in 2011 (contributed mainly by the Australian *Sydney Morning Herald*) to 438 images divided into two clusters of headshots in 2022. The coverage of the 2019 and 2022 United Nations Climate Change Conferences partially explains the increase. The cluster of people in 2019 features different images of the climate activist Greta Thunberg during her iconic, damning speech at the 2019 Conference. In 2019 and 2022, the Conference's blue background and national flags behind the politicians are characteristic of the images in these clusters.

From a web history perspective, analysing the cluster grouping images of web advertisements is revealing. In the earlier years in the data, this cluster contained images of “older media,” such as cameras, GPS devices, and USB sticks. As time goes by, these images are replaced by images of smartphones and other handheld screens. However, after 2017, the proportion of ads in this cluster shrunk, giving way to logos of companies (that are most probably the topic of journalistic coverage), as well as to cartoons, designed gifs, and simplistic images of weather maps. One possible explanation for the decline of advertisement images in these clusters is the development of programmatic advertising and content recommendation platforms that generate ads from external servers that are not part of the HTML of the archived web pages (Ratner et al., 2023).

Finally, after analysing repeating clusters over the years, we also draw attention to the years wherein unique clusters emerged or disappeared. For example, 2014 features a unique cluster of “blue marble” images of planet Earth, contributed mainly by the *Sydney Morning Herald*. However, there is no significant clustering of this iconic image in other years in the dataset, indicating, perhaps, the decline of its iconic relevance to the current climate crisis.

2020, the year in which the climate crisis coincided with the outbreak of the COVID-19 pandemic, was also a year of thousands of wildfires in the Brazilian Amazon rainforest. Accordingly, one of the clusters in 2020 features devastating images of raging fires from most sources.

Additionally, we conducted a transnational visual comparative analysis, examining image clusters from Brazil, Argentina, China, India, the United States, Germany, and Russia, where we aimed to identify the climatic narrative of each source. First, we used the heatmap in Figure 3.6 to identify the dominant frames in each source. Then, we qualitatively identified visual patterns within each cluster. The qualitative visual analysis helped identify the unique “visual stamp” characterising each source, including the cultural uniqueness and any global references they may evoke. Our focus also involved recognising local aesthetics, unique images, and repeated compositions. By comparing these findings across countries, we aimed

Iconography in flux

to gain insights into the visual representations, cultural influences, and distinct characteristics of each country's image clusters. As describing the visual stamp of all countries in our data extends the scope of this chapter, we focus below on a comparison of the seven countries.

Brazil

The dominant visual frames of the website of the Brazilian daily newspaper *Folha* are Nature at Risk – Landscape and Animals (0.23), Disastrous and Vulnerable Landscapes (0.15), Industry and Pollution (0.14), and Public Figures and People (0.15). By contrast, this source features very few or almost none of the Clean Energy, Urban, and Extreme Weather frames. The image clusters from this source feature a prominent pattern of saturated, green-tinted images juxtaposed with white background-tinted images. The Brazilian cluster stood out for its high contrast and vibrant colours, with a prevalence of green, white, blue, and red hues. Zooming in, it became evident that many images captured various aspects of nature, focusing on landscapes, flora, and natural environments. These aesthetics are mixed with portraits of people and advertisements of “older media,” which explains the white hues. As for the frame, Disastrous and Vulnerable Landscapes, we notice denoted objects such as forest or jungle animals, forests, dry land, deforestation, flooding, and other natural disasters. The cluster features people, including local and international public figures, such as Barack Obama, Bill Gates, and Greta Thunberg. Additional identified objects denote pollution and waste. Although the distribution of objects and frames in the *Folha* cluster is dispersed, images of animals stand out compared to other countries. The diversity of animals includes frogs, monkeys, whales, jellyfish, sharks, tapirs, bears, bees, and more. Urban landscapes, streets, and buildings are not represented in this cluster. Protests are rare, especially compared to neighbouring Argentina.

Argentina

The climate section of the Argentinian news website *La Nacion* stands out by featuring images of protests and demonstrations (0.31). This source's other two dominant frames are Nature at Risk — Landscape and Animals (0.09) and Disastrous and Vulnerable landscapes (0.10).

The colour palette is diverse, and there is no distinctive colour pattern. Besides the various demonstrations, we identified dominant images of icebergs, animals, and other landscapes featuring a mix of blue, brown, and green colours. Like Brazil, there are barely any images of urban environments, buildings, and fires.

China

China.org.cn is a state-run web portal of the China Internet Information Center. As such, it somewhat deviates from the other news sources in the corpus. In its climate section, the dominant visual frames are Nature at Risk — Landscape and Animals (0.30), followed by Industry and Pollution (0.14).

The palette of images originating from this source is a combination of blue-tinted (water), grey-tinted (air pollution and disasters), and brown-green (floods) shades. Zooming in, there are repetitive images of panda bears, followed by very few animals, such as monkeys and tigers. Compared to the other images in the cluster that are usually in a more distant

composition, the panda images are composed of close-up shots of the bear and presented adorably, and many images feature water, which explains the blue-tinted hues. The cluster also features repetitive images of floods and other disasters, such as flooded villages, destruction of infrastructures, dams, overflowing streams and rivers, and people walking in standing water. There are few images of urbanity and buildings, although China has undergone massive urban development in the last 30 years. The visual analysis shows a focus on landscapes and animals, pollution, and disasters. There are very few images of people; most seem to be politicians. The overall aesthetics of the images seem faded with little contrast. Some images feature smog, adding to the cluster's faded, green-grey-tinted colouring.

India

The Environment section of *The Hindu* features a dominant visual coverage of Nature at Risk (0.40), whereas all other frames are evenly spread, with no additional distinctively dominant frame. A bird's eye view of the cluster of images shows a strong dominance of green-tinted and brown-tinted images. Zooming in, animals vary, with the strong presence of elephants and birds, tigers, monkeys, and frogs. The featured public figures are mostly Indian politicians, except for Barack Obama. Although most images depict nature and are almost faded, a closer look shows several images of clean energy, industry and pollution, and natural disasters. Urbanism, or any sign of construction and crowding, is not found in the cluster.

United States

Unlike the sources from Global South countries, the dominant image frames from the internationally read U.S.-based *The New York Times* are Industry and Pollution (0.22), Nature at Risk, Landscapes, and Animals (0.14), Scientists and Science (0.12), Protests and Demonstrations (0.10), and Disasters (0.10).

Accordingly, the cluster consists of a wide range of colours and hues, reflecting the diverse subject matter covered by *The New York Times*. White stands out as a relatively dominant colour, likely due to the presence of synoptic maps and charts, which often use a white background. Blue is also significant, possibly representing themes related to oceans, water, and global issues. Some green hues are also observed, which could be associated with nature, the environment, and sustainability. Upon closer examination, we identified trends connecting images based on objects and frames. The images related to industry and pollution feature smoking chimneys and oil rigs. Images of science feature scientists, data, and research. Another distinct cluster comprises synoptic maps and satellite images of the Earth, indicating a significant emphasis on global issues and events. The global coverage is also evident in the images depicting various public figures, including the American presidents during the studied period: Barack Obama, Donald Trump, and Joe Biden, along with international figures like the Pope, Angela Merkel, Prince William, and Greta Thunberg.

Germany

The German source in our corpus is *Süddeutsche Zeitung*, one of the largest daily newspapers in the country. The dominant visual frame in its online climate section is Industry and pollution (0.23), which suits a local narrative of a highly industrialised country. The following

Iconography in flux

frame is Disaster and vulnerable landscapes (0.13), suggesting global coverage and narrative. Additional frames are People (0.08) and public figures (0.07).

The visual analysis of the German cluster resembles that of the *New York Times*, displaying a diverse and colourful palette with no clear dominance of any specific colour or shade. Certain trends and themes become apparent through a closer visual analysis. There is a dominance of images depicting smoking chimneys, industrial plants, and pollution. These images are diverse and visually striking, often captured from interesting camera angles, which may help draw attention to environmental and industrial issues. The frame of Disasters and Vulnerable Landscapes provides a climate narrative that extends beyond local events, covering disasters worldwide. Images of dry lands, floods, and other natural disasters are represented through various colours and situations, reflecting the diversity of global climate-related disasters and damaged landscapes. The presence of public figures such as Kofi Annan, Leonardo DiCaprio, Harrison Ford, Barack Obama, and Angela Merkel is also global. It seems that images of protests feature local people. The narrative is varied both in colour and context. Ranging from blue-tinted and white-tinted images of glaciers and oceans — to green-tinted and brown-tinted images of forests and green landscapes, red-tinted and orange-tinted images depicting fires and heat, and brown-tinted and light green-tinted images for floods, grey for pollution, urbanism, and industry and smoke; and light brown-tinted images for dry land, and drought. We spotted clean energy, protests, and polar bears. There are very few images of animals, which starkly contrasts the sources from countries in the Global South reviewed above.

Russia

Like China, the Russian source in our corpus is state-owned. *TASS* is the largest information agency in the country and not an independent news source. This source uses images of Nature at Risk, Landscapes, and Animals (0.35), followed by Pollution and Industry (0.20). All other frames are less frequent and divided almost equally.

The visual analysis of the image cluster from *TASS* reveals a notable prevalence of images with a blue-grey-dirty-white tint, while red, orange, and green colours are either rare or completely absent. This colour pattern is closely related to the local environment in Russia, reflecting the natural dark lighting and the overall landscape. Zooming in, many images of snow and landscapes include glaciers, oceans, lakes, and rivers, suggesting coverage of local landscapes. Additionally, we identified images featuring pollution chimneys, industry, and urban landscapes, which explain the dominance of grey colour.

The animals featured in this cluster are mainly snow animals, such as polar bears, tigers, leopards, white seals, and whales, suggesting locale coverage. We note scarce usage of the orange colour, which is accounted for by images of the tiger and, in some cases, images of waste and plastic. Pollution is featured through smoking chimneys. There are almost no climate disasters like dry land or floods.

Discussion and conclusions

The climate crisis represents a profound global phenomenon that transcends geographical and political boundaries, impacting every corner of the planet. The consequences of the climate crisis are felt by communities across continents, affecting ecosystems, agriculture, water resources, and public health. Our analysis highlights the importance of applying a

transnational and longitudinal approach to understanding climate crisis imagery as socio-semiotic systems (Schneider & Nocke, 2014) in ways sensitive to time and local and global contexts. By analysing news images from sources from both the Global North and South, we attempt to set a research agenda for web archive research that considers the role of non-Western imaginaries in understanding the evolution of the climate crisis over time. Such a transnational approach has helped us understand why certain image frames travel across countries while others remain local or invisible.

Looking at the transnational climate news image space from an iconomy perspective (Smith, 2007) or that of competing frames (Levi & Spicer, 2013), our findings helped trace the decline of the dominant image frames that were previously documented in the literature (Rebich-Hespanha et al., 2015; Schäfer, 2020) and the rise of new ones, as well as to unravel hidden competing frames. There has been a terminological shift from “global warming” to “climate crisis” in the period spanning our research data (Rogers, 2019). The visual realm of online climate news corresponds to this shift: Iconic Western imaginaries of global warming depicting melting icebergs and polar bears gradually gave way to frames of crisis depicting disasters. The rise in the clusters of images depicting people — political leaders, protests, and refugees — also highlights the rise in the anthropogenic framing of a crisis of people caused by the people rather than a crisis of nature, which is also evident in the gradual disappearance of animals from the iconic space. From a semiotic perspective, climate news images undergo concretisation from alarming metaphors (a setting sun, a polar bear) to documentary coverage (a demonstration, a U.N. summit, refugees) and, accordingly, from climate concerns to climate action.

Although the themes identified in our analysis correspond with the visual frames described in earlier research on climate news imagery (Rebich-Hespanha et al., 2015; Wozniak et al., 2015; O'Neill & Smith, 2014), our attempt to draw comparisons between local and global imaginaries has helped us discover hidden patterns and local narratives. Among them are the different emphasis on nature in South America and India, compared to an emphasis on industrial pollution in Germany and the U.S. In particular, the analysis of the climate animals has surfaced the elephant as a dominant visual frame in African and Asian climate imaginaries, which escaped global attention.

The image clustering techniques have helped us identify similarities in the image space across countries and years. Dominant image objects like smoking chimneys or cracked soil look almost identical across sources. Such global frames might indicate the isomorphic formation of global climate imaginaries, whereby various debates are reduced to a limited number of similar visual frames, echoing previous findings by Wozniak et al. (2015). Disasters such as floods, oil leaks, fires, and droughts look similar across the globe. The cultural differences, instead, are found in the different proportions dedicated to such frames in each country.

This research has several limitations. Chief is the detachment of the images from the articles' context, which limits the ability to accurately estimate the ties between the visual framing and the article's framing. Further, although the analysis encompasses 30 countries, many others are missing, which limits our ability to conduct statistical analysis about differences in the visual elements of climate images that extend beyond descriptive statistics. Although we normalised the proportion of images and image objects across countries and years, we did not take potential country bias in the Internet Archive Wayback Machine in the archival coverage of different countries. Further, we cannot

Iconography in flux

determine the role of image banks in circling similar images across countries. Finally, our decision to focus only on the top label predicted by OpenAI Clip for each image may have lost some of the image's context, which was more apparent when looking at multiple labels per image. In future research, we will address these limitations. Firstly, we aim to broaden the corpus's scope to encompass all countries, facilitating a more extensive geographical and cultural analysis. Secondly, our future analysis aims to offer additional contextualisation regarding the semantic connections between images and texts within web news climate sections, encompassing article content, titles, and image captions. For instance, we may compare the outcomes from OpenAI Clip in this study with the labels generated from image captions. We anticipate that news websites will increasingly utilise generative AI tools in their editorial work and publications. Therefore, we consider such forthcoming research trends particularly crucial as they can offer significant historical insights into visual and textual editing practices in climate coverage.

Despite these limits, this exploratory research contributes to transnational web archive studies. First, we have demonstrated the utility of the Internet Archive Wayback Machine as a primary source for conducting a large-scale longitudinal analysis of the evolution of journalistic coverage of a global issue. Second, we advance methods for visual analysis of the archived web. Finally, we advance a critical investigation of the archived web in ways that are mindful of biases and other epistemic holes in web archives and especially mindful of extending our imaginaries — of both web history and the climate crisis — beyond Western eyes.

Note

- 1 Open AI Clip: <https://openai.com/research/clip>. Archived 14 July 2023: <https://web.archive.org/web/20230714023156/https://openai.com/research/clip>

References

- Ainsworth, S. G., Nelson, M. L., & Van de Sompel, H. (2014). A framework for evaluating memento temporal coherence. *arXiv:1402.0928* [cs]. <http://arxiv.org/abs/1402.0928> (accessed 23 July 2023).
- Amram, A., Ben-David, A., & Bekkerman, R. (2018). The colors of the national web: Visual data analysis of the historical Yugoslav Web domain. *International Journal on Digital Libraries*, 19(1), 95–106.
- Bachmann-Medick, D. (2016). *Cultural turns: New orientations in the study of culture*. De Gruyter.
- Ben-David, A. (2021). Critical web archive research. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The past web: Exploring web archives* (pp. 181–188). Springer.
- Ben-David, A., & Amram, A. (2018). The Internet Archive and the socio-technical construction of historical facts. *Internet Histories*, 2(1–2), 179–201.
- Born, D. (2019). Bearing witness? Polar bears as icons for climate change communication in National Geographic. *Environmental Communication*, 13(5), 649–663.
- Boudana, S., Cohen, A. A., & Frosh, P. (2023). How iconic news images travel: Republishing and reframing historic photographs in Israeli newspapers. *Journal of Communication*, 73(1), 49–59.
- Brügger, N. (2009). Website history and the website as an object of study. *New Media & Society*, 11(1–2), 115–132.
- Colombo, G. (2019). Studying digital images in groups: The folder of images. In L. Rampino & I. Mariani (Eds.), *Advancements in design research*. FrancoAngeli Design International.
- DiFrancesco, D. A., & Young, N. (2011). Seeing climate change: The visual construction of global warming in Canadian national print media. *Cultural Geographies*, 18(4), 517–536.
- Goggin, G., & McLelland, M. (Eds.) (2017). *The Routledge companion to global internet histories*. Taylor & Francis.

- Grevsmühl, S. V. (2016). Images, imagination and the global environment: Towards an interdisciplinary research agenda on global environmental images. *Geo: Geography and Environment*, 3(2), e00020.
- Hochman, N., & Manovich, L. (2013). Zooming into an Instagram City: Reading the local through social media. *First Monday*, 18(7), n.p., www.firstmonday.org/ojs/index.php/fm/article/download/4711/3698.
- Hochman, N., & Schwartz, R. (2012, May). Visualizing Instagram: Tracing cultural visual rhythms. In *Proceedings of the International AAAI Conference on Web and Social Media* 6(1).
- Hridayami, P., Putra, I. K. G. D., & Wibawa, K. S. (2019). Fish species recognition using VGG16 deep convolutional neural network. *Journal of Computing Science and Engineering*, 13(3), 124–130.
- Jasanoff, S. (2001). Image and imagination: The formation of global environmental consciousness. In C. A. Miller & P. N. Edwards (Eds.), *Changing the atmosphere. Expert knowledge and environmental Governance* (pp. 309–338). MIT Press. <https://doi.org/10.7551/mitpress/1789.003.0013>.
- Levy, D. L., & Spicer, A. (2013). Contested imaginaries and the cultural political economy of climate change. *Organization*, 20(5), 659–678.
- Machin, D. (2004). Building the world's visual language: The increasing global importance of image banks in corporate media. *Visual Communication*, 3(3), 316–336.
- Manovich, L. (2009). How to follow global digital cultures, or cultural analytics for beginners, <http://manovich.net/index.php/projects/how-to-follow-global-digital-cultures> (accessed 23 July 2023).
- Manovich, L. (2020). *Cultural analytics*. MIT Press.
- Mitchell, W. J. T. (2007). World pictures: Globalization and visual culture. *Neohelicon*, 34(2), 49–59.
- O'Neill, S. (2022). Defining a visual metonym: A hauntological study of polar bear imagery in climate communication. *Transactions of the Institute of British Geographers*, 47(4), 1104–1119.
- O'Neill, S. J. (2013). Image matters: Climate change imagery in U.S., U.K., and Australian newspapers. *Geoforum*, 49, 10–19.
- O'Neill, S. J., & Smith, N. (2014). Climate change and visual imagery. *Wiley Interdisciplinary Reviews: Climate Change*, 5(1), 73–87.
- Pearce, W., Niederer, S., Özkula, S. M., & Sánchez Querubín, N. (2018). The social media life of climate change: Platforms, publics, and future imaginaries. *Wiley Interdisciplinary Reviews: Climate Change*, 10(2), ISSN 1757-7780, <https://doi.org/10.1002/wcc.569>.
- Purgar, K. (2014). Visual studies and the pictorial turn: Twenty years later. *Images*, 2(2).
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.
- Ratner, Y., Dvir Gvirsman, S., & Ben-David, A. (2023). “Saving journalism from Facebook’s death grip”? The implications of content-recommendation platforms on publishers and their audiences. *Digital Journalism*, 11(8), 1410–1431. <https://doi.org/10.1080/21670811.2023.2180403>.
- Rebich-Hespanha, S., & Rice, R. (2016). Climate and sustainability| Dominant visual frames in climate change news stories: Implications for formative evaluation in climate change campaigns. *International Journal of Communication*, 10, 33. <https://ijoc.org/index.php/ijoc/article/view/4701>.
- Rebich-Hespanha, S., Rice, R. E., Montello, D. R., Retzlöff, S., Tien, S., & Hespanha, J. P. (2015). Image themes and frames in U.S. print news stories about climate change. *Environmental Communication*, 9(4), 491–519.
- Rigney, A. (2005). Plenitude, scarcity and the circulation of cultural memory. *Journal of European Studies*, 35(1), 11–28.
- Rogers, B. (2019). Rebranding climate change: From a realist perspective. *International Journal of Climate Change: Impacts & Responses*, 11(4), 33–44.
- Rogers, R. (2021). Visual media analysis for Instagram and other online platforms. *Big Data & Society*, January–June, 1–23.
- Schäfer, M. S. (2020). News media images of climate change: Reviewing the research. In D. C. Holmes & L. M. Richardson (Eds.), *Research handbook on communicating climate change* (pp. 143–152). Elgaronline.
- Schneider, B., & Nocke, T. (Eds.) (2014). *Image politics of climate change: Visualizations, imaginations, documentations*. Transcript Verlag.

Iconography in flux

- Schwarz, Garbielle (2022, Nov 5). 'It was like an apocalyptic movie': 20 climate photographs that changed the world. *The Guardian*. Archived 5 November 2022. <https://web.archive.org/web/20221105110328/www.theguardian.com/environment/2022/nov/05/20-climate-photographs-that-changed-the-world>.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Smith, T. (2007). The architecture of aftermath: Iconomy and contemporaneity. *Bulletin of the Institute of Humanity, Human and Social Sciences*, 88, 33–47.
- Smith, T. (2021). Iconomy, iconoclasm ≠ iconomics. *The Nordic Journal of Aesthetics*, 30(61–62), 172–194.
- Smith, T. (2022). *Iconomy: Towards a political economy of images*. Anthem Press.
- Thelwall, M., & Vaughan, L. (2004). A fair history of the web? Examining country balance in the Internet Archive. *Library & Information Science Research*, 26(2), 162–176.
- Wevers, M., & Smits, T. (2020). The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities*, 35(1), 194–207.
- Wozniak, A., Lück, J., & Wessler, H. (2015). Frames, stories, and images: The advantages of a multimodal approach in comparative media content research on climate change. *Environmental Communication*, 9(4), 469–490.

4

COMPARING THE HOLDINGS OF CLOSED NATIONAL WEB ARCHIVES THROUGH SUMMARIES

Janne Nielsen, Yves Maurer, and Eld Zierau

Introduction

A recurring theme in the web archive studies community in recent years has been the relationship between the national and transnational. Although the internet is a global infrastructure, the use and preservation of the web are often situated in a national context. While there are international web archives, a lot of web archiving takes place in national archives with the aim of collecting and preserving the national web as part of the nation's cultural heritage. Most transnational studies using archived web will, therefore, likely need to or benefit from including materials from different archives. This is complicated by the scarcity of knowledge about the holdings of individual archives, the closed nature of many national web archives and their different aims, scopes, and legal frameworks.

In this chapter, we ask: how can we better understand what is in the holdings of different archives, how can we compare holdings across archives, and what types of analyses can be done based on the comparisons? The chapter addresses these questions from both a researcher's perspective and a library perspective, where both parties are interested in data being accessible and used for research. We present a method we call *cdx-summarize*, which addresses the main obstacles in comparing web archive holdings by creating summaries that can be used to understand the holdings of each archive and to compare across archives. To show the potential of *cdx-summarize* as a method and a possible standard for describing web archive holdings, we present different types of transnational analyses based on the summaries. The focus is on explaining the method and the analytical potential, so the analyses in this chapter should be understood not as exhaustive analyses but as illustrative examples of the types of questions that can be addressed using this approach, and which would be difficult to answer otherwise. The transnational aspect is central in two ways: we create a way of analyzing web archive holdings to allow for transnational comparisons, and we show how national web archives are also in a way transnational, as they include parts of other countries' national web domains. Our approach does not solve all problems facing transnational web archive studies,

Comparing the holdings of closed national web archives

but it provides a first step in understanding the possibilities and challenges of transnational web archive studies by shedding light on what is available to study and compare.

The chapter explains why such a method is useful and what similar initiatives exist, and then the method itself, including the data generated and used, the legal framework, and requirements and limitations. Finally, the chapter presents analytical examples based on the summaries of selected web archives. The sample archives are the national web archives placed at the national libraries of Denmark, France, Hungary, and Luxembourg as well as what has been preserved of these countries' national country code top-level domains (ccTLDs) in the collections of the Internet Archive and Common Crawl, which are both international, US-based non-profits that offer open access to web crawl data from many countries. The archives are abbreviated as follows: Netarkivet at the Royal Danish Library (NAK), Bibliothèque nationale de France — Archives de l'Internet (BNF), Luxembourg Web Archive at Bibliothèque nationale du Luxembourg (LWA), OSZK webarchívum at National Széchényi Library (SZL), Internet Archive (IA), and Common Crawl (CC).

Context

The main challenges of comparing across archives

Several studies have grappled with questions related to national aspects of the web as seen in the edited volume *The Historical Web and Digital Humanities: The Case of National Web Domains* (Brügger & Laursen, 2019) and other publications (for instance Ben-David, 2016; Ben-David, Amram, & Bekkerman, 2018; Brügger, Nielsen, & Laursen, 2020; Hale, Yasseri, Cows, Meyer, Schroeder, & Margetts, 2014; Rogers, Weltevrede, Borra, & Niederer, 2013). However, many web researchers and archivists are also interested in transnational aspects, as evidenced by the Web ARChive studies network researching web domains and events (WARCnet), and in creating connections between the national and the transnational — and international. In some cases, it hardly makes sense to address only the national, while dealing with phenomena that cross national boundaries. In other cases, relating the national level to the transnational level can be beneficial to understand how the web is at the same time national, transnational, and international.

Transnational web archive studies are made difficult by at least two major obstacles: The first obstacle is that the mentioned aims, scopes, and legal frameworks are different across archives. This means that what is available in different archives is not necessarily comparable — or at least the comparisons must be made with certain reservations and a keen eye for how these different conditions influence the archive holdings. The second obstacle is that web archives, especially national ones, are often only accessible under certain strict regulations. The archiving and preservation by many national web archives are regulated by legal deposit laws, which means that the acquisition of web content by the web archive is legal. However, as archived websites often contain personal data and copyright-protected data, access to data is restricted to abide by national and international regulations.

Therefore, in many cases, there is no easy, openly available access to the data. Some national web archives only allow access to their collection onsite from physical reading rooms; this is the case for the French and Luxembourg web archives. On the contrary, the Danish web archive offers access to their full collection online via VPN but only for researchers affiliated with Danish research institutions and following an application procedure (Det Kgl. Bibliotek, n.d.). Some archives, like the Hungarian and British national web archives, combine access

options, making a minor part of their collection accessible to the public online (IIPC, n.d.2; British Library, n.d.), while most of the collection can only be accessed onsite. Finally, some archives offer free online access to their holdings, among these the Internet Archive and Common Crawl, but also the national web archives of Iceland (vefsafn.is), Croatia (haw.nsk.hr), and Portugal (arquivo.pt).¹ Still, the closed nature and legal frameworks of most of the national web archives make it difficult for researchers from other countries to get access to the materials (Brügger, 2024).

Similar initiatives to compare web archive holdings

Different methods have been investigated to describe large-scale web archive collections or provide cross-archive information about resources. One very successful initiative by Van de Sompel et al. (2009) resulted in the Memento Protocol, which was adopted as an IETF RFC (Van de Sompel et al., 2013). It allows for remote querying of multiple web archives for particular URLs in time. Like this, even closed archives can advertise their holdings publicly, while restricting usage of the resource itself. It does not produce a downloadable file that can be statistically analyzed though. On top of this, other technologies have been built like MementoMap by Alam (2020), which uses profiling of the web archive holdings to make the Memento framework more efficient.

In Klein, Balakireva, Holub, Celjak, and Rudomino (2022), bloom filters are used to efficiently check whether a web archive holds a particular resource, while it also allows for the sharing of the holdings information in a format that is unencumbered by personal data and copyright considerations. A bloom filter is a probabilistic data structure that can tell you whether some value (in our case a URI) is probably in a collection or if it is definitely not in there. It does this through hash functions, which means that the original URI is irreversibly transformed into a fixed-length string and cannot be retrieved later. While this approach fulfills the requirement that by sharing the bloom filter, no personal or copyrighted data is made available, it also means that researchers cannot infer which domains and media types have been archived through this data structure. Klein et al. (2022) note that “CDX indexes would be the much better choice as they only contain metadata about the crawled resources such as their URI, date time of crawl, etc. However, we have observed a reluctance by institutions to share information about their holdings in plain text.” That reluctance, in addition to technical hurdles, also explains why so few web archives make CDX API endpoints publicly available. All these points are addressed with our cdx-summarize program.

Other tools exist to characterize collections of WARC files, like cdx-summary² presented in Alam and Graham (2022), which gives aggregate information about media types (also known as MIME types, see glossary) and status codes, but no information broken down by domains so is not suited for exploring domains and overlaps between holdings.

A combination of methods was used by Brügger, Laursen, and Nielsen (2019, 2020) to compute statistics that show the developments of a national web, but the researchers needed access to protected material to run the analyses.

Our idea

The work presented in this chapter springs from the work in WARCnet Working Group 1: Comparing entire web domains. The overall objective was to “support research that aims to compare entire national and transnational European web domains” (WARCnet,

Comparing the holdings of closed national web archives

n.d.) by having web researchers and web archivists join forces to develop new approaches while drawing on their joint expertise and experiences from working with web archives — in research and preservation, from both a technical and a legal point of view.

Our idea for a solution to the twin challenges of protected data and lack of access to web archives is to provide a summary of an archive's holdings in the form of descriptive statistics, which offers researchers an overview of the holdings in terms of number and size of files and different media types preserved in the different years covered by the archive, and which does not contain protected data, and is, therefore, not subject to the same legal restrictions (Maurer, 2022). Such an aggregated data set is a middle course between, on the one side, full access, and on the other side, very general overall descriptions of, for instance, the total number or size of files in the archive or a general description of the archive's scope. The aggregated data is computed in a predefined way so that data sets from different web archives with different policies and different infrastructures remain comparable. The summaries should preferably be small enough so that an individual researcher can process the data on a personal computer to allow for large-scale comparisons between archives without requiring massive resources for storing and high-performance computing.

As a summary does not contain copyrighted information or personal data, an institution with a closed web archive can make the summary file of its collection openly available to the public. That would enable researchers to explore the statistics of the archive's holdings to better understand the collection and — in the case of more archives making summaries available — facilitate cross-archive and transnational comparisons. A format including yearly distributions allows for the study and comparison of historical developments. These insights could be relevant for researchers from a multitude of fields, and for other interested parties.

Method

How cdx-summarize works

As mentioned, the restrictions on access might still allow for some ways of mapping the holdings, and the approach taken here is to summarize web archives by using available index files to get simple, comparable statistics. During the WARCnet project, Yves Maurer at the Luxembourg Web Archive developed a program named *cdx-summarize*, which can easily be run by a web archive to generate a summary of the archive's holdings, detailing the number of domains and their sizes in terms of media types per year. The resulting summary files are simple and comparatively small but still contain rich data.

Most web archives today use open-source tools that work with WARC files and function in a similar way. To replay the web archive, they rely on indexes into the WARC files and *cdx-summarize* uses those index files to calculate its statistics. This is intentional as WARC files are around 50 times bigger than the index files, which affects the effort required on the part of the operators of the web archives in running *cdx-summarize*. There is support for different flavors of CDX, including the one used by Internet Archive's CDX server.³ It also supports CDXJ files, such as the ones used by Common Crawl or pywb.⁴ There is a special version of the program (*cdx-summarize-outbackcdx.py*) for web archives which have an OutbackCDX server.⁵ Finally, for web archives using WARC-indexer from the British Library⁶ and with a backend with a Solr index, there is a version that pulls the data from Solr efficiently and directly.⁷

The summary file format

The summary files contain statistics about domains and the number and sizes of different media types across the collection years. The file consists of JSON lines prefixed by the domain name. This simplifies using Unix tools such as “sort,” or “join,” on the summary files, or even just working with domain names by using the first column. In the JSON part, there are keys for each year, and inside the year, keys for each (simplified) media type for the number of mementos (n_) and their sizes (s_):

```
host.tld {"year": {"n_html": A, ... "s_html": B}}
```

```
A sample entry could be: bnl.lu {"2003": {"n_audio": 0,"n_css": 8,"n_font": 0,"n_html": 639,"n_http": 728,"n_https": 0,"n_image": 44,"n_js": 0,"n_json": 0,"n_other": 7,"n_pdf": 30,"n_total": 728,"n_video": 0,"s_audio": 0,"s_css": 5268,"s_font": 0,"s_html": 1295481,"s_http": 4680354,"s_https": 0,"s_image": 295235,"s_js": 0,"s_json": 0,"s_other": 13156,"s_pdf": 3071214,"s_total": 4680354,"s_video": 0}}
```

For details on how the media types are simplified, see GitHub.⁸

The summary files can be calculated cumulatively if you base them off disjoint sets. For example, if you have a summary file A for a broad crawl done in 2021 and a different summary file B for a broad crawl done in 2022, you can just combine those two summary files by adding up the fields for each domain that appears in both crawls. This is implemented in the program combine-summary.py. It is helpful for generating an initial summary file from a large collection, as you can usually separate it into smaller sub-collections on which you run cdx-summarize. It is also useful for adding information about new crawls to an existing summary file. If the summary file for the whole collection up to 2022 exists and you would like to just add 2023, you will only need to process the CDX files/index for 2023 and then combine the resulting summary file with the one you already had. This simplifies operations for web archive operators.

Advantages, limitations, and requirements: researchers and archives

The extract is meant to be a standard that researchers can know to ask for, even if they are new to web archives and do not know much about what would be useful to ask for. To get, use, and maintain the extracts, there are various advantages, limitations, and requirements both for researchers using the extracts and web archives producing the extracts, which relate to:

- Availability of data
- Extraction practices
- Legal framework
- Requirements for the researcher.

Availability of data

Even though all web archives have indexes that can be used for cdx-summarize, not all can necessarily provide all the data that the extracts are supposed to include. That is because not all replay systems need, for example, the file size or media type and so it is sometimes not

Comparing the holdings of closed national web archives

in the index. One way to ensure that all the requested information that is present could be for an archive to expand its index to cover all the information included in cdx-summarize. Another way could be to run special extracts on the full archive. Such operations will in most cases be both resource and time-consuming for the web archive. The alternative is that the researcher should be aware of the missing information, which is the case in the following analysis. As more data are archived over time, the extracts need to be updated, which can easily be done by the archive, as explained above.

Extraction process

Different web archives have different platforms and infrastructures for their web material, which means different setups for running the extract, and a strength of the extract program is that it is designed for a wide number of platforms.

Most of the web archives in this study have provided the summaries for free. The Internet Archive and Common Crawl summaries have been created by us via APIs. If API access is given, it is theoretically possible for a researcher to compute the summary but in practice, it can still pose significant hurdles. Downloading the 90317 parts of the .fr CDX via Internet Archive's API took several weeks and many retries for broken connections or other errors. For the web archive operator, the amount of data and its index type determines the effort necessary to extract the data. For LWA, it took about 12 hours to run on a collection of half a million CDXJ files stored on an NFS drive. For SZL, the cdx-summarize-outbackcdx.py was used and it took around 24 hours to run. For the BNF, it was only possible to provide an extract from one year with available resources, which took around 4 hours to run. For all web archives above, the programs were provided entirely by the author and no development was needed on the part of the web archive operators. Only running the scripts and copying the data had to be done. NAK on the other hand had to ask for coverage of costs for tuning the extraction scripts for the NAK platform and for engineering time for running the scripts for the extract as they were the only ones running SolrWayback on a large scale.

All index types currently known to the authors are covered by the cdx-summarize scripts so applying them in a different web archive will only demand time and computing resources to run them. For these, different web archive operators will probably have different policies and researchers might have to reserve funds for extraction.

Legal framework

Different countries have different legal setups, and in this study, there have been various considerations on whether the summaries from the extracts could pose any legal challenges, both regarding the second-level domains and the statistics in the summaries. The section that follows is not meant as a comprehensive legal analysis but only as an indication of the challenges and how they can be addressed. It involves EU and Luxembourg law as one of the authors is working at the National Library of Luxembourg and is familiar with Luxembourg law. For the domain names themselves, there could be implications for intellectual property rights, moral rights, trademarks, and finally GDPR.

The consensus around the signatory countries of the Berne Convention seems to be that domain names are not subject to copyright (Copyright.gov, n.d.). The applicable Luxembourg law states that only original artistic or literary expressions are copyrightable but

not isolated words (Loi du 18 avril 2001 sur les droits d’auteur, les droits voisins et les bases de données). A domain name consists of a single or isolated word. The 2009 judgment of the European Court of Justice in the case Infopaq International A/S v Danske Dagblades Forening (EUR-Lex, Case C-5/08, point 45) also underlines that law. Insofar as they are not copyrighted, secondary intellectual property rights such as moral rights, etc. do not apply.

Trademarks are an issue when someone is trying to register a domain name. So, the holder of a domain name may perform trademark infringement against a trademark holder. As an example, trying to register cocacola.lu would be trademark infringement. However, there is no problem when referring to a website known by its trademark name, as long as it is not intentionally misspelled.

Concerning GDPR, there was a worry, as the Danish web archive has one known case where personal information was exposed in the archived URLs, and thus URLs, in general, cannot be disclosed (thelocal.dk, 2014). However, as the second-level domain is registered, there is no personal data in the domain name itself except unless the domain holder has included that information explicitly (like <http://yvesmaurer.ch/>). By using that domain name, the domain holder has given consent to all DNS servers, browsers, etc. to process the domain name so that users can access the site. It is meant as the opposite of private since it is the public address that the holder has chosen.

So there seem to be no strong arguments against publishing lists of second-level domain names. This is also seen in practice as some European registrars like Sweden and France make the whole zone file (list of registered domains in .se and .fr respectively) available to anyone as open data.

The statistics in the summaries are about the holdings of the web archive itself. For each domain, the statistics hold information about how many URLs were saved with which media types and which sizes in that particular web archive. There are no third-party rights to that. This is entirely information about the web archive collection itself. There seem to be two relevant and applicable laws; copyright and the open data and the re-use of public sector information Directive (EU) 2019/1024 Article 3.

The EU has the concept of database rights. Even if each piece of information in the database is a fact (not copyrightable) or public domain or otherwise not subject to intellectual property rights, the database itself can be considered to be under copyright. This is a right that is held by the web archive and can be waived by it to enable publishing a database under an open license. If the web archive is a public institution, the EU directive 2019/1024 says that in principle it should waive its property rights (including database copyright) to enable re-use.

Requirements for researchers

Researchers need some technical background and skills to work with summaries. One of the most important requirements for the researcher is a basic understanding of the web and the archived web. One example is the changeability of web material, where a website can be an important website for literary scholars in one year and a site for toaster reviews the next year, as was the case for litlive.dk (2007, 2018). This means that if web archives include captures of a website from different timeframes, the sites may be completely different even though it is the same domain. Our method does not solve this issue, but the yearly distributions are likely a more accurate way of comparing overlaps than if the entire accumulated holdings were compared.

Comparing the holdings of closed national web archives

Besides general web knowledge, using this method also requires an understanding of what is and is not in a summary. For example, only the second-level domain is included, so, e.g., all Facebook pages and profiles are grouped under facebook.com. So, it will not be possible to check whether different web archives have copies of the same Facebook profiles.

Excel can be used for simple statistics on small summaries and on the output from `host_year_total.py` and `overlap.py` for larger summaries. If researchers would like to run more advanced analytics, they need some experience with a programming language such as Python or R, as done in this study.

Data

Archives included

The following illustrative analyses use material from the national web archives of Denmark, France, Hungary, and Luxembourg as well as what has been preserved of these ccTLDs in the Internet Archive and Common Crawl. Including these open, trans-national archives gives us a common comparison point to the closed web archives. The national webs are of different sizes so directly comparing the number of domains in each one does not offer any insights into differences in collection strategies. Assuming that the collection policies, frequencies, and curatorial decisions for IA and CC are similar for all national web spheres, we can use them as a common comparison point to show how the holdings of the different national archives are influenced by their collection strategies. For example, if the BNF's web archive for the .fr ccTLD is larger than IA's and the one for .hu of the NSZL is smaller than IA's, it means that it is likely that the BNF's domain crawls are very vast, but the NSZL does only selective or partial domain crawls. Another interesting case could be an almost complete overlap between a national web archive and IA if the web archive outsources some of or all its technical operations to IA's crawling service (IA, n.d.).

CC is increasingly relevant as their data is used in a lot of the datasets that form the basis of large language models, for example, The Pile, introduced by Gao et al. (2020). So, its intersection with the national web archives can give researchers a first indication of how representative it is.

As IA has one of the oldest web archives, its inclusion in this study allows for exploring a longer period compared to the national web archives, which started later. Finally, the inclusion gives researchers an indication of whether they could work with freely available collections for their purposes instead of relying on the closed national web archives.

Limitations in the data

Although the program produces the same type of output for each web archive, there are idiosyncrasies that need to be taken into consideration. First, the summary files do not try to eliminate duplicates so if an archive crawls frequently and produces duplicates, the holdings reflect this. Typically, IA has the highest crawl frequency and probably has more duplicates as it collates web data from lots of different sources like its own teams, Archive-It partners, The GDELT Project, etc.⁹

Two archives in our study do not have complete file size information. The CDX format used by the BNF has no size information, so their summary also does not provide that information. CC cuts off large files at 1 MB (Nagel, 2018), so it erroneously looks like they only harvest small files.

As the BNF provided the summary for one single domain crawl from 2021, we do not have data about event or thematic crawls or other years from them. It is also noteworthy that CC has no data from 2011.

Examples of analyses using the sample archives

Analytical focus

The summary files allow for different types of analysis on different levels with a focus on one archive or cross-archive comparisons. The following examples have two focal points, the national and the transnational, and illustrate how our method can be used to analyze and compare national domains transnationally, thus adding insights to the complicated issue of how to do transnational studies across archives with different scopes and holdings. After a visualization of the holdings of each archive, showing the number of domains over time, we focus on three aspects that illustrate the analytical potential: First, we study how much of a ccTLD is available in the different archives. The second part examines and compares what top-level domains (TLDs) are the most common in the holdings of the different national archives. Finally, we study overlaps between the archives. The presentation here is very selective and not comprehensive; the aim is to show illustrative examples of the types of questions that can be asked and the types of insights that can be gained.

The overview in Figure 4.1 is based on the summaries and shows that the number of domains and the periods covered are very different, which can be explained by different scopes, policies, and strategies for archiving. All the national archives have many more domains than their ccTLD, corresponding to the definition of a national web as something more than just the ccTLD.

How much of a ccTLD is available in different archives?

One transnational aspect — and a driving force behind this study — is to find out how much of a country's web has been archived in other web archives. Here, we define the national web domain differently than most national archives, where nationally relevant domains on TLDs other than the ccTLD are included. We use only the ccTLD because it is impossible for us to determine what other domains could or should be considered part of the nation's web. To ascertain this, we would need the full curated list of domains included in the conceptualization of the national web from each of the national archives for each of the years included — and to exclude years before a national archive exists, even if others were harvesting domains from the country's national web. Also, some domains, like social media, may be considered relevant to more than one country's national web.

Figure 4.2 shows how many second-level domains from the national ccTLD are available in the sample archives. The historical data allows us to examine the development over time, to see when the archive first captured domains from the ccTLD, and whether there are interesting patterns in the development. As mentioned, our analysis includes four national web archives and the two international archives IA and CC. When visualizing the

Comparing the holdings of closed national web archives

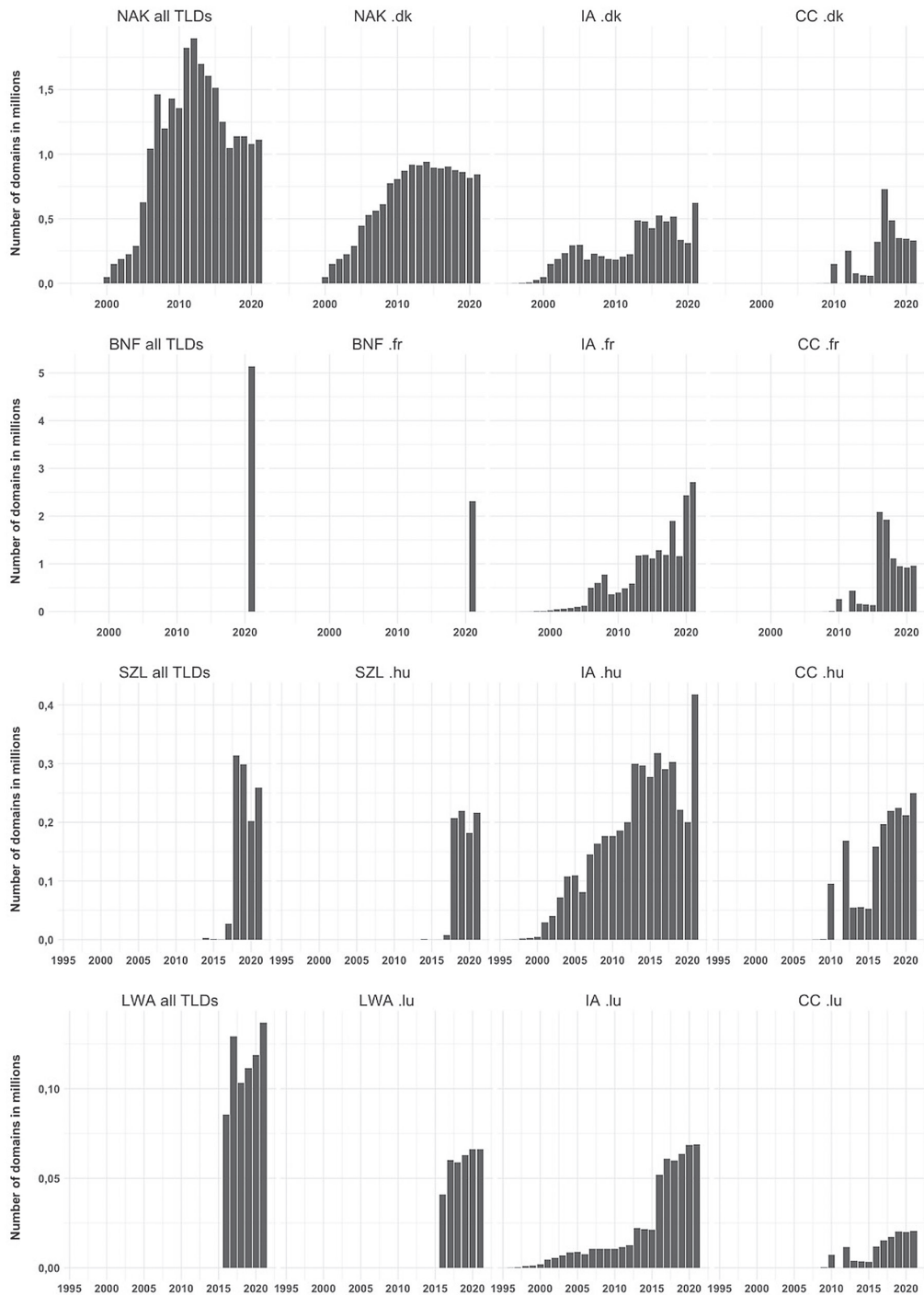


Figure 4.1 Overview of the web archives' holdings.

The Routledge Companion to Transnational Web Archive Studies

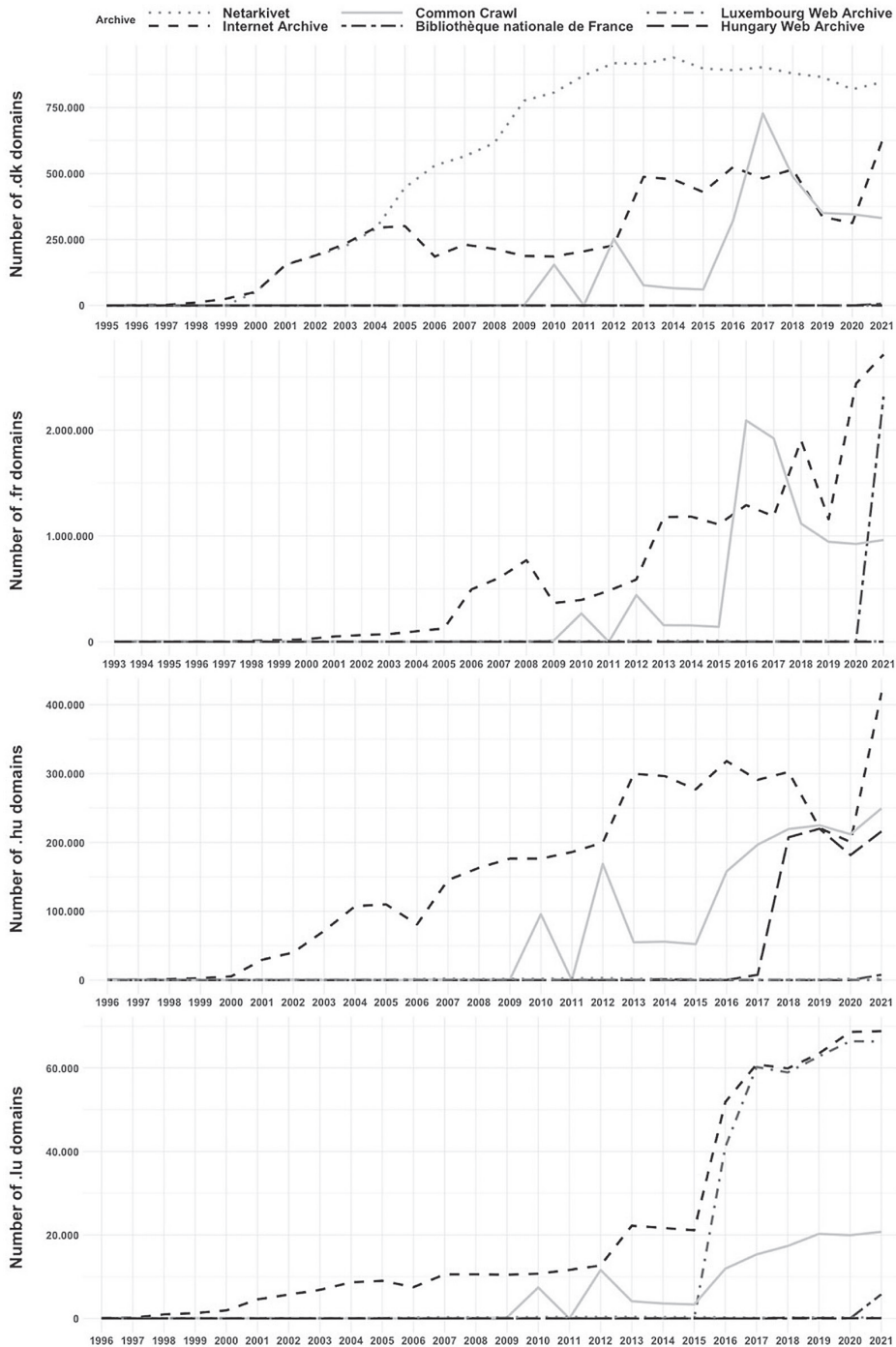


Figure 4.2 Each ccTLD across web archives.

Comparing the holdings of closed national web archives

development, two things stand out: (1) In all cases, IA has the first archived versions of domains with the ccTLDs, which is not surprising as IA predates the other archives by several years, and (2) only IA and CC have holdings of each of the national ccTLDs that are large enough to be properly visualized alongside the national web archive's holdings. The numbers for the other national archives are so small that they are mostly not distinguishable in the figures. For instance, LWA and SZL have very low numbers with less than 400 domains on average per year, except .fr where they have on average 2,300 and 600 domains per year, respectively. NAK also has substantially more .fr domains than .lu and .hu domains, corresponding to the French web being by far the largest of the four in number of domains on the ccTLD. For the one year we have, BNF has the highest numbers of domains from the other ccTLDs (5,800–7,800), which might indicate a relationship between the size of the national web, the archive, and the inclusion of other national webs, intentionally or not. This would be interesting to examine.

Using the summaries to study a ccTLD across different archives allows for questions like, whether the archives have a similar development in the number of domains archived, how collaborations between archives influence the holdings, and whether the international archives have similar or differing strategies for different ccTLDs.

Comparing the holdings of *.dk domains* across the archives, for example, the Danish NAK holds the largest amount of distinct .dk domains with a curve steadily increasing until 2012, where it stabilizes around approximately 900,000 domains a year with a tendency to a small decline in the later years. Before 2005, the curves for NAK and IA were similar with NAK a bit lower. NAK began their broad crawls in December 2005, so their collection from before then is a combination of static web documents (included according to the Danish Legal Deposit Act, 1997) and early .dk content from 2000 to 2004 bought from IA (Nielsen, 2009), which explains the relationship between the curves. Even with .dk domains obtained from IA, there are still domains that are in IA but not in NAK. NAK uses a list of all registered active (and not suspended) .dk second-level domains as a seed list but the unique IA domains can be explained by a possible higher crawl frequency resulting in the capture of domains that were not on NAK's list.

The number of yearly .dk domains in IA declines from 2005 to 2006, coinciding with the start of NAK. In the .hu data, we see a similar decline in IA captures coinciding with the start of the national web archive. It would be interesting to study if such co-occurrences are coincidental or due to IA's strategies.

BNF and LWA have also collaborated with IA. From 2004 to 2009, BNF and IA collaborated on annual broad crawls of the French domain (IIPC, n.d.1), and the 2008 peak in the IA data is due to the largest crawl (IA, 2012). The IA numbers drop momentarily after the collaboration before generally increasing but fluctuating, so again the relationships are interesting. LWA started its first broad crawls in collaboration with IA in the summer of 2016 (Noguera, Nielsen, Maurer, & Els, 2024), resulting in a steep increase (245%) in the IA numbers in 2016 and a very similar development in IA and LWA numbers from 2017 forward. IA has 1,400 .lu domains more on average than LWA, though, because of a higher crawl frequency (ibid.).

Generally, there are both large jumps, outliers, and periods with relatively stable numbers in the data, which indicate changes not just in the number of domains on the ccTLDs but also in harvesting policies and settings. It requires a further in-depth analysis of the fluctuations to fully understand the holdings, but the summaries offer a starting point for mapping the developments and asking questions about them.

Looking at the curves of the international archives across the different ccTLDs, there is a definite pattern in the CC holdings, especially from the first years until 2015. From 2015 we see an increase in the numbers to a degree matching the sizes of the national webs. The larger the web, the steeper the increase. The development is not entirely the same but for .fr and .dk we see a prominent peak in 2016/2017, indicating a lot of CC crawls. Similarly, there seem to be comparable highs and lows in the IA data for .dk, .fr, and .hu, especially from 2012 onwards. The pattern becomes clearer as more data is available (the higher the numbers), meaning the highs and lows are the most noticeable in the .fr data, and the least in the .hu data. This points to the difficulties encountered in assessing whether relatively minor changes in small data may still signify changes in harvesting strategies or settings.

What TLDs are represented in the holdings of the different national archives?

While the national web archives have a similar intent of harvesting and preserving the national web, the archives still differ in the scope, settings, and strategies for how to do this. The sample archives, however, all interpret the national web as something that includes not only the assigned national web domain (ccTLD) but also domains on other TLDs if they are considered culturally relevant for the country. One way to help us understand the holdings of different web archives is to look at the distribution of domains on different TLDs.

Figure 4.3 shows the percentage of domains that are located on the top 20 TLDs in the four national archives, including an “Other” category showing the percentage of domains below the top 20. The total number of TLDs differs depending on the scope and size of the archive. The French archive is by far the largest and even with just one year of data there are 870 different TLDs in the BNF summary — more TLDs than in the entire holdings of each of the other archives. Several TLDs are only represented by one or a few domains, so the importance of the number of individual TLDs should not be overstated.

In all four national archives the most common TLD is — not surprisingly — the ccTLD, and the second most common is .com. In NAK, LWA and BNF the ccTLD accounts for 38–45% of the holdings, and domains using .com comprise approximately a third, whereas in SZL the ccTLD takes up 63% and .com only 19%. Interpreting the meaning of the prevalence of .com domains poses a challenge because we do not know whether all the .com domains are considered part of the national web domains. Depending on an archive’s scope and strategies, .com domains may be harvested due to being on a seed list, because domains on the seed list link to it and the crawler follows links one or more steps, or because resources like images or videos used on webpages in national domains are hosted on platforms using .com. The characteristic linked structure of the web makes it hard to delimit a national web (Brügger et al., 2020), and some .coms may be considered part of the national web, while others might be considered peripheral or outside the national domain.

The linked structure also explains the high probability of national web archives including parts of other countries’ national web domains. The ccTLD featured the most in the other countries’ web archives are Germany with .de taking third place in both the Danish and the Luxembourgish archives and fifth place in the French and Hungarian archives. The size of Germany and the close geographical proximity may explain why .de sites are so prominent in the archives. Generally, there seems to be a geographical aspect to the holdings. High in the list for NAK are Sweden’s .se and Norway’s .no, which place lower in the other lists. The Belgian .be and the Swiss .ch rank higher in the Luxembourgish and French archives, which SZL has the

Comparing the holdings of closed national web archives

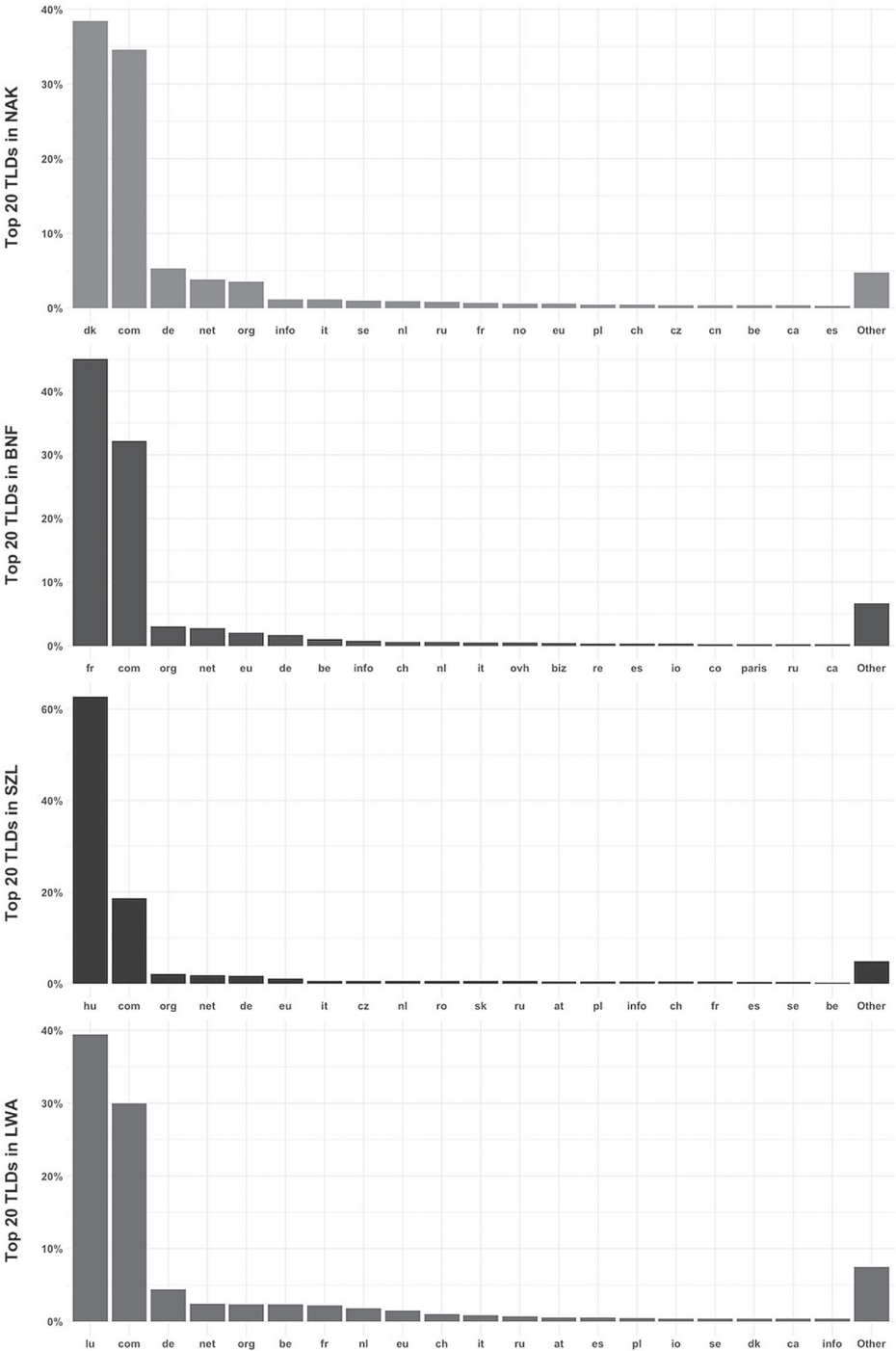


Figure 4.3 Top 20 TLDs in the national web archives.

Czech .cz, Romanian .ro, and Slovakian .sk. However, in all but one case, the others mentioned place in the top 40, so they are still represented, just in lower numbers. Again, we should consider the different sizes of the archives when comparing because 1% equals a very different number of distinct domains in the different archives due to size (between approximately 2.800 and 62.100 domains), so for the larger archives, the number of domains separating the rankings are much higher. Note also that the numbers here do not indicate factors like how many times the different domains have been harvested, how many URLs from these domains are included, or how large the sites are. This could be explored, though, using the summary files as they include the number and size of URLs across the years.

Like .de, .org places either third or fifth, and in all the archives .net places fourth, so the top placements are alike. Also recurring is the immense difference between second and third place with the TLDs in third place having only around 2–5% of the total number of distinct second-level domains. Below the top 10, all numbers are less than 1%, creating a long tail with very low numbers. The “Other” category with the combined percentage for all TLDs below the top 20 comprises around 5–7.5% on more than 500 TLDs. So, while the national domains are conceptualized as something more than the ccTLD, it seems clear that most of these other domains are .com domains. Other generic TLDs and ccTLDs play a relevant role but individually they take up small portions of the holdings.

Overlap analyses

This part illustrates how summaries can be used to study not just where we can find parts of a national web but also how much the archive holdings overlap. As the number of archived domains from the other countries’ ccTLDs are relatively small, as shown above, we focus on the overlaps between a national archive and the international archives, exemplified by analyses of .dk and .hu for all years included (for an analysis of .lu, see Noguera et al., 2024) and of all four ccTLDs for the year of 2021. What is compared in the overlap analyses are the distinct second-level domains from each archive within a year, but the URLs might be different across archives, and even if two archives hold the exact same URLs, the content is not necessarily identical. On the contrary, as they were probably not archived at the exact same time, they might differ due to the dynamics of updating and the temporality of archiving (Brügger, 2018). The national archives have a grey color, IA has vertical lines, CC has horizontal lines, and overlaps combine the relevant signatures (e.g. dark vertical lines signify the overlap between the national archive and IA).

The overlap analysis for .dk in Figure 4.4 clearly shows that the largest part are domains unique to NAK. This speaks volumes about the efforts of the national archive in capturing .dk domains and their ability to do so because they use lists from the Danish domain name registrar as seed lists for their broad crawls (in combination with a curated list of relevant domains on other TLDs). Another obvious finding is that in all the years where all three archives are represented (no data from CC for 2011), there are some domains that are shared between all three, some that are shared by just two (in all three combinations), and some that are unique to each archive.

There are overlaps between IA and NAK almost from the beginning, due to the mentioned collaboration, but we also see that both NAK and IA hold domains that are not available in the other archive. From 2010 forward, all three archives have shown some overlap, the size of which seems to correlate with the development in the CC harvests (confer Figure 4.2). The most prominent year for CC is 2017 with large overlaps both for all three archives and

Comparing the holdings of closed national web archives

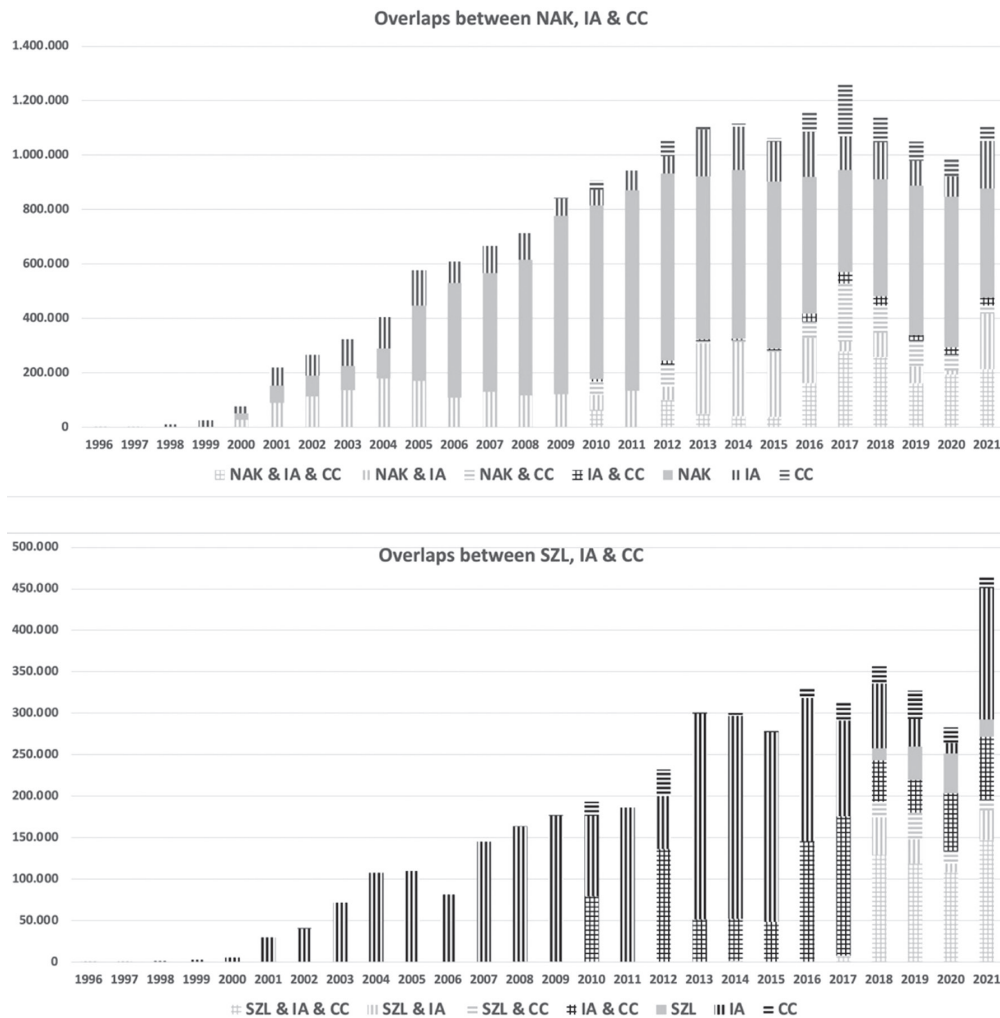


Figure 4.4 Overlap analyses.

for NAK and CC, and around 191k unique domains. This year CC did one crawl per month for the first time (CC, n.d.). Comparatively, NAK crawls the entire Danish domain four times a year, which might explain the amount of unique .dk domains in CC's holdings. The overlap between IA and CC, which is not shared with NAK, is small across the years. There are more domains unique to IA than CC in the early years, whereas in the later years, the numbers are more similar indicating an increased effort from CC to be comprehensive.

The .hu overlap analysis also shows overlaps between all the archives but the overlap here is a larger proportion of the numbers (but fewer domains). IA is the sole archive for the first years, and when CC starts archiving .hu, almost all the domains in most years overlap with IA. Compared to the Danish case, few domains are unique to the national archive, which can be explained by the amount of archiving of .hu domains done by IA and CC and by SZL's strategy. Whereas the Danish web is archived as fully as possible in line with legal deposit

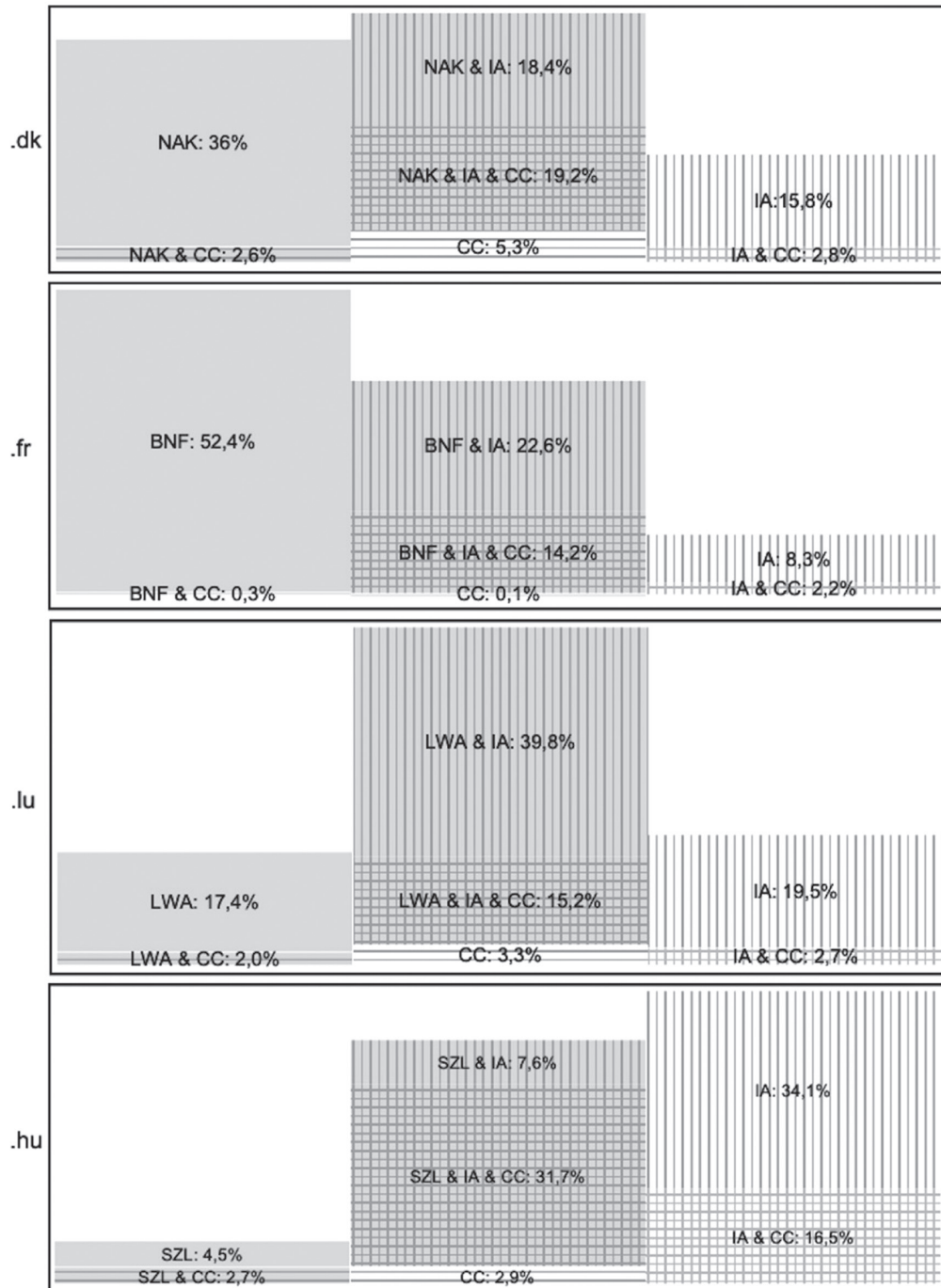


Figure 4.5 Overlaps 2021.

Comparing the holdings of closed national web archives

law, the scope of SZL is related to their three-part strategy: “At first making a selection from the most important Hungarian websites, secondly making event-based harvests from main news resources and thirdly making general, representative harvests from the Hungarian web space” (OSZK webarchívum, n.d.). The intention is not to capture all ccTLDs, and the 48,000 domains archived by SZL are likely important sites for Hungarian culture. Numbers alone do not necessarily indicate the value of a cultural heritage collection. After the establishment of SZL, relatively few domains are unique to CC, while the IA numbers fluctuate from less than 12,000 to more than 158,000 in 2021.

Another way to use overlap analyses to study national webs from a transnational perspective is to compare the relationship between the national archives and the international archives across different countries. To illustrate, Figure 4.5 shows the proportions (of 100%) of the overlap of second-level domains in the year 2021.

Visualizing the overlaps in this way makes the comparisons across countries easier and the differences more obvious. For example, for all the national web archives except SZL, the proportions that are not covered by the archive itself (i.e. IA, CC, and IA & CC) are less than 26%. For SZL it is more than 50% due to the aforementioned strategy. The proportion covered only by the national archive varies significantly (from 4.5% to 52.5%), which indicates, again, different strategies of the national archives but possibly also that IA and CC might have different strategies, including possible collaborations, for different ccTLDs.

Comparing one year only makes it very important to consider yearly differences, though, as they might greatly impact the interpretations. For instance, NAK only did two broad crawls in 2021 due to COVID-19, which likely influenced the numbers, as the average proportion of NAK and IA unique domains from 2012 to 2021 is 48.3% and 11.9%, respectively, compared to 36% and 15.8%. Visualizing the overlaps in different years or visualizing an average across years might prove more appropriate but was impossible here due to only one year of French data. Either way, overlap analyses might highlight interesting patterns or relationships to examine further.

Conclusion

We have shown that by providing a simple method for web archive operators to run a program that generates comparable statistics, we can get rich summary statistics that enable a high-level analysis of the holdings of the web archive. By showing for the first time that the holdings of four closed web archives can be analyzed systematically and by providing strong arguments for the proposition that summary files can be released as open data by web archive operators, we hope to inspire web archives to run these programs themselves and make the resulting summary files available for the research community. Compared to WARC and CDX files, summaries are smaller, simpler files and may lower the barriers for researchers starting in web archiving research. To enable researchers to use comparable statistics from different archives, the community should settle on one standard, and we propose one here. For researchers considering traveling to a physical reading room, it might also help discover whether there is a likelihood that the archive will have the sought-after data.

We illustrate the power of the summary files with some comparative studies and how they can show the overlap between archives, thus illustrating the transnational nature of web archives, even while focusing on a national web as defined by the archives. Including IA and CC collections in the comparative analysis shows how interesting they are as sources for transnational studies and frames them as a common reference point to the closed

national collections. The seeming patterns but also differences underline the importance of thinking about web archive holdings as not only representations of a web sphere but also representations of harvesting strategies and settings.

Other types of studies can be done using summary files. For example, it is possible to document the gradual switch of the web from http to secured https by using their relative frequencies. The prevalence of audio files could give indications of the rise of podcasts. One can analyze the linguistic properties of domain names in different archives and at different times or do quantitative analyses about the size of archived websites with respect to their longevity.

More work is needed to convince web archive operators to provide summary files, develop software and analysis tools capable of interpreting summaries, popularize them with researchers, and find new and innovative uses for them.

Notes

- 1 Arquivo.pt, as an exception, is not run by a national library but has a legal mandate which allows access to the data (see Brügger, 2024).
- 2 Internet Archive (2021). CDX Summary. Github. <https://github.com/internetarchive/cdx-summary>.
- 3 Internet Archive (2013). Wayback/wayback-cdx-server. Github. <https://github.com/internetarchive/wayback/tree/master/wayback-cdx-server>.
- 4 Webrecorder (2013). Webrecorder pywb 2.8. Github. <https://github.com/webrecorder/pywb>.
- 5 National Library of Australia (2015). OutbackCDX. Github. <https://github.com/nla/outbackcdx>.
- 6 UK Web Archive (2012). WARC Indexer. Github. <https://github.com/ukwa/webarchive-discovery/tree/master/warc-indexer>.
- 7 Yves Maurer (2022). CDX-summarize-warc-indexer. Github. <https://github.com/ymaurer/cdx-summarize-warc-indexer>.
- 8 Yves Maurer (2021). Summarize CDX. Github. <https://github.com/ymaurer/cdx-summarize>.
- 9 Archive-It is a web archiving subscription service by Internet Archive aimed at institutions that collect and preserve web content. The GDELT Project monitors and categorizes broadcast, print, and web news from around the world.

References

Note: Archived web references are provided as PWID (Persistent Web Identifier) URNs (Uniform Resource Names), a way to create persistent references to archived web sources. PWID is registered in IANA's Registry of URN Namespaces (<https://web.archive.org/web/20240306083352/www.iana.org/assignments/urn-namespaces/urn-namespaces.xhtml>), which include a description of how to use and resolve a PWID: <https://web.archive.org/web/20231012093849/www.iana.org/assignments/urn-formal/pwid>. The PWID identifies the archive, timestamp, whether it is a part or page, and the URI archived. Thus, the PWID for the link above is urn:pwid:archive.org:2023-10-12T09:38:49Z:page: www.iana.org/assignments/urn-formal/pwid

- Alam, S. (2020). MementoMap: A web archive profiling framework for efficient memento routing. PhD Dissertation, Computer Science, Old Dominion University. <https://doi.org/10.25777/5vnk-s536>.
- Alam, S., & Graham, M. (2022). CDX summary: Web archival collection insights. In G. Silvello et al. (Eds.), *Linking Theory and Practice of Digital Libraries. TPD L 2022. Lecture Notes in Computer Science*, vol 13541. Springer. https://doi.org/10.1007/978-3-031-16802-4_25.
- Ben-David, A. (2016). What does the web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. *New Media & Society*, 18(7), 1103–19. <https://doi.org/10.1177/1461444816643790>.

Comparing the holdings of closed national web archives

- Ben-David, A., Amram, A., & Bekkerman, R. (2018). The colors of the national Web: Visual data analysis of the historical Yugoslav Web domain. *International Journal on Digital Libraries*, 19, 95–106.
- British Library (n.d.). UK Web Archive. urn:pwid:archive.org: 2023-05-28T09:18:21Z;page: www.bl.uk/collection-guides/uk-web-archive
- Brügger, N. (2018). *The archived web: Doing history in the digital age*. MIT Press.
- Brügger, N. (2024). Comparing national webs across national web archives: Methodological challenges of doing transnational studies of entire web domains. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge Companion to Transnational Web Archive Studies*. Routledge.
- Brügger, N., & Laursen, D. (Eds.) (2019). *The historical web and Digital Humanities: The case of national web domains*. Routledge.
- Brügger, N., Laursen, D., & Nielsen, J. (2019). Establishing a corpus of the archived web: The case of the Danish web from 2005 to 2015. In N. Brügger & D. Laursen (Eds.), *The historical web and Digital Humanities: The case of national web domains* (pp. 124–142). Routledge.
- Brügger, N., Nielsen, J., & Laursen, D. (2020). Big data experiments with the archived Web: Methodological reflections on studying the development of a nation's Web. *First Monday*, 25(3), n.p. <https://doi.org/10.5210/fm.v25i3.10384>
- Common Crawl (CC) (n.d.). So you're ready to get started. urn:pwid:archive.org:2023-06-26T09:54:38Z;page:<https://commoncrawl.org/the-data/get-started/>
- Copyright.gov (n.d.). What does copyright protect? urn:pwid:archive.org:2023-07-22T18:12:09Z;page:www.copyright.gov/help/faq/faq-protect.html
- Det Kgl. Bibliotek/The Royal Danish Library (n.d.). Research access to netarkivet. urn:pwid:archive.org:2023-09-18T15:56:52;page:www.kb.dk/en/find-materials/collections/netarkivet/research-access
- EUR-Lex, Case C-5/08. Infopaq International A/S v Danske Dagblades Forening. urn:pwid:archive.org:2022-08-11T22:23:04Z;page:<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:62008CJ0005&from=en>
- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A. Nabeshima, N., Presser, S., & Leahy, C. (2020). The pile: An 800 gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027.
- Hale, S., Yasseri, T., Cows, J., Meyer, E. T., Schroeder, R., & Margetts, H. (2014). Mapping the UK webspace: Fifteen years of British Universities on the web. *WebSci 14 Proceedings of the 2014 ACM Conference on Web Science*, 62–70, ACM. <https://doi.org/10.48550/arXiv.1405.2856>.
- International Internet Preservation Consortium (IIPC) (n.d.1). Bibliothèque nationale de France. urn:pwid:archive.org:2023-06-02T19:36:42Z;page:<https://netpreserve.org/about-us/members/bibliothèque-nationale-de-france-national-library-france/>
- International Internet Preservation Consortium (IIPC) (n.d.2). Országos Széchényi Könyvtár. urn:pwid:archive.org:2023-06-11T22:52:11Z;page:<https://netpreserve.org/about-us/members/orszagos-szechenyi-konyvtar/>
- Internet Archive (IA) (2012-07-21). Crawl data from Internet Archive from 2008-04-04T22:43:24PDT to 2008-04-05T01:50:32PDT. Retrieved 05.07.2023 from <https://archive.org/details/BNF-CRAWL-003-20080404224324-00133-crawling03-c>
- Internet Archive (IA) (n.d.). Domain-scale Web Harvests. urn:pwid:archive.org:2023-03-09T03:52:17Z;page:<https://webservices.archive.org/pages/crawling>
- Klein, M., Balakireva, L., Holub, K., Celjak, D., & Rudomino, I. (2022). Investigating bloom filters for web archives' holdings. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (JCDL '22)*, Article 14, 1–10.
- Legal Deposit Act (1997). Lov om pligt aflevering af udgivne værker. LOV nr. 423 af 10/06/1997. urn:pwid:archive.org/2024-02-23T13:53:32Z;page: www.retsinformation.dk/eli/acn/A19970042330
- Litlive.dk (2007). litlive Jun 07. urn:pwid:archive.org:2007-06-10T17:33:59Z;page:www.litlive.dk/
- Litlive.dk (2018). Levende litteratur, bøger og anmeldelser. urn:pwid:archive.org:2018-08-31T11:12:58Z;page: www.litlive.dk/
- Loi du 18 avril 2001 sur les droits d'auteur, les droits voisins et les bases de données (2001). Le gouvènement luxembourgeois. urn:pwid:archive.org:2022-04-01T10:33:54Z;page:<https://legi.lux.public.lu/eli/etat/leg/loi/2001/04/18/n2/jo>.

The Routledge Companion to Transnational Web Archive Studies

- Maurer, Y. (2022-08-10). Investigate holdings of web archives through summaries: cdx-summarize. IIPC blog. urn:pwd:archive.org:2023-08-11T11:19:22Z:page:https://netpreserveblog.wordpress.com/tag/cdx-summarize/
- Nagel, S. (2018-08-26). August crawl archive introduces language annotations. Common Crawl. urn:pwd:archive.org:2018-09-04T06:37:25Z:page:https://commoncrawl.org/2018/08/august-2018-crawl-archive-now-available/
- Nielsen, E. K. (2009). Forord. In L. Ømann (Ed.), *Årsberetning 2009*. Det Kongelige Bibliotek. urn:pwd:archive.org:2016-03-17T17:04:13Z:page:http://kb.dk/export/sites/kb_dk/da/kb/aarsberetning/aarsberetninger/2009.pdf
- Noguera, C., Nielsen, J., Maurer, Y. & Els, B. (2024). Exploring the evolution of the .lu domain names through a transnational comparative: similarities and differences between .lu and .dk. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge Companion to Transnational Web Archive Studies*. Routledge.
- OSZK webarchívum (n.d.). About this website. urn:pwd:archive.org:2023-04-23T03:12:53:page:https://webarchivum.oszk.hu/en/for-users/short-description/
- Rogers, R., Weltevrede, E., Borra, E., & Niederer, S. (2013). National web studies: The case of Iran Online. In J. Hartley, J. Burgess, & A. Bruns (Eds.), *A Companion to New Media Dynamics* (pp. 142–66). Wiley-Blackwell.
- Thelocal.dk (2014-06-20). ‘Hactivist’ arrested for leak of MPs’ ID numbers. urn:pwd:archive.org:2022-07-02T06:07:27Z:page:www.thelocal.dk/20140620/hactivist-arrested-for-leak-of-politicians-id-numbers/
- Van de Sompel, H., Nelson, M., & Sanderson, R. (2013). HTTP framework for time-based access to resource states — memento. RFC 7089. urn:pwd:archive.org: 2023-03-21T14:52:16Z:page:www.ietf.org/rfc/rfc7089.txt
- Van de Sompel, H., Nelson, M., & Sanderson, R., Balakireva, L.L., Ainsworth, S., & Shankar, H. (2009). Memento: Time travel for the web. https://doi.org/10.48550/arXiv.0911.1112
- WARCnet (n.d.). *Working Groups*. urn:pwd:archive.org:2023-02-01T00:20:06:page:https://cc.au.dk/en/warcnet/working-groups

5

EXPLORING THE EVOLUTION OF .LU DOMAIN NAMES THROUGH A TRANSNATIONAL COMPARISON

Similarities and differences between .lu and .dk

*Carmen Noguera, Janne Nielsen, Yves Maurer,
and Ben Els*

Let us now have a closer look at the challenges of remaking this study in another European nation state, with a view to making transnational web studies of national web domains
(Brügger, 2019b, p. 39)

Introduction

This chapter aims to study the development of the .lu domain by analyzing the domain names from 1992 to 2022 and putting the results in perspective using a transnational approach. Part of the objective is to evaluate whether a method used in the national web study from Denmark by Brügger et al. (2017) can be reproduced in another national setting and to what extent it is possible to adapt it to the specific reality of another nation's web. Brügger et al. (2017) investigate how the domain names list provided by the national registry — a source that has been “often overlooked” (Brügger et al., 2017, p. 64) — can be used to study the national web. In this chapter, we follow the basic ideas in the Danish study but adapted to the Luxembourg case and data. We use the list from the Luxembourgish registrar, Restena, to analyze .lu and retrace the history of the Luxembourg web in a similar way, and then we compare our analysis of .lu with the analysis of the Danish .dk in Brügger et al. (2017). The comparison examines the possible common denominators and trends across the two country code top-level domains (ccTLDs) to offer insights relevant for transnational studies.

The chapter is divided into three main parts. The first part will study the evolution of the Luxembourg web through its domain names, because of the domain names registration list provided by Restena (Réseau Téléinformatique de l'Éducation Nationale et de la Recherche), the national registry since 1992. According to Brügger et al. (2017): “The institutionalized national delimitation mirrored in the ccTLD constitutes an appropriate first

step in identifying national web” (Brügger et al., 2017, p. 65). Taking this delimitation as a premise, we asked: What can domain name registrations and deletions tell us about the beginning of the Internet in Luxembourg? To reply to this, we will approach more descriptive questions such as: How has the number of domain names grown over time? How many have been registered and disappeared over time?

As a next step, we added a more specific set of questions related to the aliveness of the domain names. Following Brügger et al. (2017, p. 63) we measure aliveness in regard to “new/disappeared domain names and frequency of updating”. How many domain names from the early days are still alive — remain active? To what type of websites do they relate? What is the average lifetime of the Luxembourg web domains? If a domain name is active, does it mean that it has a website? If there is a live website, has the topic changed over the years? Which categories have been more affected by the disappearance of domain names?

In the second part, we will compare the Luxembourg Web Archive (LWA) from the National Library of Luxembourg (BnL) from August 2016 onward with the collection of .lu domains in (1) the international Internet Archive (IA) from 1996 onward and (2) in Common Crawl (CC) from 2009 onward to analyze how these sources combined, together with the ccTLD list, can help us create a picture of the evolution of the .lu domain. More specifically, we asked: How representative of the Luxembourg web are these web archives in comparison with the Restena registry list? Additionally, we wanted to analyze the synergies between the three archiving institutions (national and international) for the specific case of Luxembourg, their collaboration, the (in) completeness of their collections, and how they complement each other in the preservation of the Luxembourg web.

The last part will reflect on the reproducibility of methodology to conduct transnational studies of national web domains by comparing the Luxembourg analysis with the Danish one and establishing some trends and common denominators to apply in transnational studies despite the differences in the collections, with their specific criteria, temporalities, archiving strategies, and access conditions. To what extent is it possible to make transnational comparisons of the historical evolution of national web domains? What specificities and challenges do we need to consider when doing transnational studies? How does the difference between temporalities and sources affect the comparisons between countries, and can we still find some trends?

A growing research area

National web domains have been studied since the early 2000s, and they can be categorized into “three waves with a different focus”, according to Brügger and Laursen (2019a, p. 415). While the first wave was mainly focused on analyzing national web domains from a technical point of view, studying elements such as hyperlinks, language, page size, among others, as we can see with studies from Baeza-Yates et al. (2006), Baeza-Yates et al. (2007), Miranda and Gomes (2009), Gomes and Silva (2005), Lasfargues et al. (2008), Zabicka and Matejka (2007), the second wave studied the national webs as a source “in their own right” (Brügger & Laursen, 2019a, p. 415) to be used as part of research projects within the social sciences and the humanities. They mainly focused on the current live web as seen in the studies by Rogers et al. (2013) and Ben-David (2014). In contrast, the third wave, which emerged almost in parallel to the second, is characterized by historical approaches, often using archived web as the main source. An early project in this wave was “Probing a Nation’s Web Sphere — the

Historical Development of the Danish Web”, which began in 2014 with the aim “to analyze the development of the Danish national web from 2005 to 2015 as it was archived in the Danish national web archive, Netarkivet” (Brügger et al., 2017, p. 63). Another example of a historical project, albeit with a very different methodology to the Danish one, is “Web90—Patrimoine, Mémoires et Histoire du Web dans les années 1990” (<http://web90.hypothesos.org>), whose objective was analyzing in a more qualitative way the historical development of an entire national web, the French one, based on multiple sources (Schafer, 2018; Schafer & Thierry, 2012, 2016). Other examples of studies include Ben-David’s (2016) reconstruction of the history of the web domain of the former Yugoslavia (.yu) with the IA as a source and her later study of the web of Palestine and Kosovo, countries located at the fringe of the web (Ben-David, 2019). Hale et al.’s (2014) studied the evolution of the presence of British universities with data obtained from the IA; Musso and Merletti’s (2016) reconstructed the UK business web space (1996–2001) using web directories, which draws from a project “Big UK Domain Data for the Arts and Humanities (BUDDAH)”, which analyzed the UK web domain with crawls conducted by the IA¹.

National web studies have been based on the challenges of defining, analyzing, and delimiting the national web, especially in countries where the national ccTLD has been barely used or the web is fragmented for several reasons, such as different national subdomains (like .co.uk and .ac.uk) and/or several archiving institutions with differing archiving methods and access criteria. More specifically, different scopes and crawling technologies result in different types of collections (for instance, crawls of an entire national domain versus selective crawls including only a curated selection of websites), which are often not stored in the same place or do not have the same modes of access. This is specifically relevant for transnational studies when comparing national webs, where it would be necessary to combine all the collections from one country, which might not correspond to how collections are built (and can be combined) in another country. Milligan and Smyth (2019) highlighted the challenges of studying the .ca domain in a country where generic top-level domains (gTLD) such as the .com predominate, and there are different archiving institutions with different solutions and legislative contexts. Teszelszky (2019) underlines the challenges of studying the national web when there is no legal deposit, such as the Dutch National web, and where other top-level domain (the Frisian region .frl) and other local TLDs, such as Amsterdam (.amsterdam) coexists. Winters (2019) presented the challenges of studying the .uk ccTLD with different subnational collections. Chambers et al. (2019) presented the challenges in Belgium, where no national web archive exists; instead, there are different web archiving initiatives.

Niels Brügger published the first edited volumes, giving space to scholars to explore the specificities of the different national webs (Brügger & Schroeder, 2017; Brügger & Milligan, 2019; Brügger & Laursen, 2019b). A national web can also be analyzed from a variety of angles taking into consideration different characteristics. For instance, Brügger (2017) analyzed how a national web sphere might be reconstructed based on the archived web and its inherent challenges. Brügger et al. (2020) explored how large-scale web history studies based on the archived web can be performed within a national web domain. Brügger and Laursen (2019b) focused on the importance of studying the history of national web domains, its challenges and historical evolution of the research focus. In addition, regarding the importance of national web studies, we can highlight Brügger (2019a), where the author proposed creating a national and even a transnational “Web Trend Index” as an essential source to understand national webs and their historical development.

For the purpose of this chapter, which is to assess the feasibility of reproducibility of the methodology used on a national level to transnational studies, we took Brügger et al. (2017) as the starting point to first assess the viability of using the national domain name list as a source to study the national web adapting to the specificities of each country, and to compare web archive collections and national registries for transnational studies.

An overview of the domain names of the Luxembourg web

As many authors have argued, delimiting a national web is challenging. Only using the ccTLD to determine a national web has challenges and limitations, especially in countries where they are barely used (Hale et al. 2014; Milligan & Smyth, 2019; Brügger et al., 2020). As Brügger stated: “There is not a one-to-one relation between the physical borders of the nation and the nation’s borders on the Web” (Brügger, 2017, p. 63).

In Luxembourg, where 47.1% of the population are foreign nationals, and 46.3% of the working labor consists of cross-border workers (Klein & Peltier, 2022), how do we define the national web, and how representative is the .lu? The BnL defines the Luxembourg web as:

- All websites with the ccTLD “.lu”
- Websites published by a resident or legal entity registered in Luxembourg
- Websites hosted on servers in Luxembourg
- Websites published by nonresident Luxembourg citizens or websites in relation to Luxembourg, determined by manual research, as contributions by website owners or third-party subject experts.

According to the BnL December 2022 domain crawl seed list, which acts as the base to preserve an essential part of the national web, and to which we had access because of two of the authors of this book chapter managing the LWAs, most of the domain names, 94.8%, correspond to the .lu top-level domain name, followed by .com (3.6%) and .eu (0.6%). The .fr, .be, and .de domains account for 0.06%, 0.09% and 0.14%, respectively. Although there are slight differences in the numbers, this high representation of the .lu domain is constant throughout all the domain crawl seed lists of different years. Nevertheless, we cannot confirm that this is entirely representative of the Luxembourg web as, although the BnL has a complete list for the .lu, the domain names from other top levels are based on their best effort, thus not exhaustive.

In this study, we focus on .lu domains, firstly because we aim to assess the viability of using the domain name list from the national registry, Restena, as a source to study the national web, and secondly for the feasibility of making transnational comparisons, in this case with .dk. However, it would be interesting for future studies to attempt to get closer to the full definition. Still, it would be challenging to compare national webs across international archives as the inclusion of gTLDs make it hard to assess nationality from the URLs.

The development of the domain names of the Luxembourg web

Restena, launched in 1989 by the Ministry of Education to meet the communication needs of education and scientific research institutions in Luxembourg (Restena, 2020), is today the foundation that manages the network for research and education institutions, together with the health, culture, and administration sectors. Restena has been managing the ccTLD .lu since 1992, when the institution connected Luxembourg to the Internet (Chrillesen &

Exploring the evolution of .lu domain names

Swain, 2020). Since then, it has overseen all aspects of domain name registration. It first allocated domain names free of charge to anyone requesting them and later with a registration and maintenance fee (Restena, 1997).

On 16 March 2001, the government set up a working group to create a new domain name charter for the .lu (Dns.lu, 2001). In 2006, accredited registrars were created and took over registration requests (Restena, 2022). As of February 2023, there were 55 accredited registrars, although registrations can be completed online through *my.lu*.

Evolution of domain names registration and deletions

To analyze the evolution of domain names from 1992 to 2022, we took the data from the public statistics published on *dns.lu*. In February 2023, there were 114,944 domain names registered, 76.51% representing companies and organizations and 23.49% representing individuals. The nationalities of the owners of .lu domains are mainly: 54.96% from Luxembourg, 7.72% from Germany, 8.61% from Belgium, and 9.86% from France according to the official national domain name registry (Dns.lu, 2023).

To analyze the yearly evolution of the domain name registrations, we extracted the annual and monthly data published by Restena on the *dns* website, which allowed us to establish patterns and differentiators between the years. Restena data (Figure 5.1²) shows that the web started slowly in Luxembourg, with fewer than 20 domain names registered yearly in

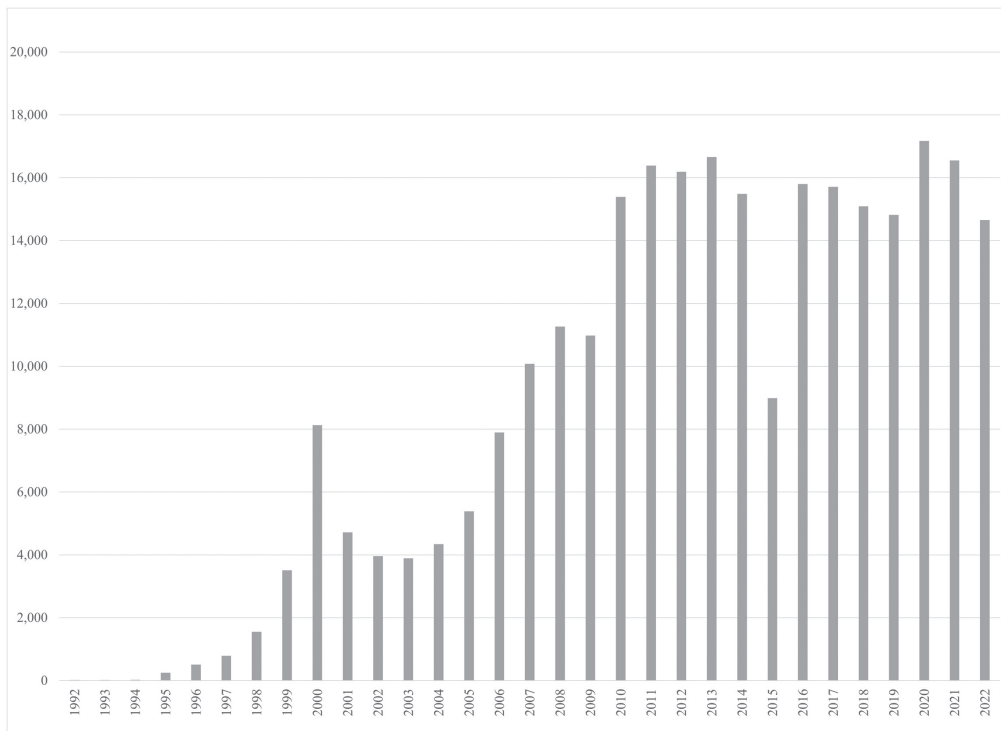


Figure 5.1 Number of .lu domains registered per year (1992–2022).

1992 and 1993, remaining low until registrations took off in 1995, with more than 250 new registrations. Since then, registrations have increased, reaching the highest peak of the first decade in 2000, with more than 8,000 new domain name registrations, representing approximately 55% of all domain names registered since 1992, with an average of registrations close to 1,650 per year.

It should be noted that the number of registrations fell in 2001, with an average in the first half of the 2000s (2001–2005) of approximately 4,450 domain names registered per year, which increased by 150% in the second half of the 2000s (more than 11,100 registration per year) from 2006 to 2010. 2010 ended with more than 92,000 registrations³.

From 2011 to 2022, the yearly numbers were relatively stable, except for 2015, when registrations seemed to decrease. However, the data for 2015 is incomplete⁴ as there are no registrations or deletions listed from September to December. These numbers should be interpreted with additional caution. The annual average during this period was around 15,292 domain registrations, with a steady monthly rate of 1,000–2,000 new registrations over a decade (Restena, 2022). Luxembourg finished 2022 with 114,693 domain names registered.

We also analyzed the deletion rate⁵ based on the premise that the disappearance of domain names could give us as much information as the registrations. For that, we used the same public statistics and calculated the disappeared domain names based on the following formula:

$$\begin{aligned} & ((1) \text{ Total number of domain names at the beginning of the year} \\ & + (2) \text{ Number of domain names registered in the relevant year}) \\ & - (3) \text{ Total number of domain names at the end of the year} \end{aligned}$$

Nevertheless, this methodology has encountered its limitations. Based on Restena's public data, as of 2006, the public *dns.lu* website started to include the numbers of registered, deleted, and transferred domain names. While these numbers generally display the same trends as our inferred calculations, we could see that the exact numbers of deleted domain names showcased on the website do not match them individually. Comparing the same available time series, the individual differences between the two datasets stand at approximately 10%, translating into a maximum absolute error of 1% in the percentage of inferred deleted domain names. At the time of writing, we have not found a plausible reason to explain these differences. Although we explored possibilities of potential data duplications, the timings on the year that the information was registered, etc., none of them seems entirely convincing.

We use the inferred numbers instead of those supplied by Restena because their numbers only cover part of the period studied, part of the data for 2015 is missing, and the number of deletions paired with registrations do not seem to correspond to the total number of registered domains. Either way, we interpret the data cautiously. We used the inferred deletions not as absolute numbers but as providers of some trends in the registration and deletion rates over time. A closer look at the data gave us interesting information.

The average ratio of deleted domains during the entire time series between 1992 and 2022 was 11.1% (Figure 5.2). The average number of domains deleted per year from 1992 until 2000 was slightly over 40, totaling approximately 330. 2001 was the year with the highest number (30%) of domain names disappearing⁶, followed by 2002 with 23.1%, and 2003 with 15.3%. The average deletion rate in 2001–2005 was 18%, which decreased to 11.7% in 2006–2010. This ratio stabilized from 2011 to 2022, with an average of 12.1%.

Exploring the evolution of .lu domain names

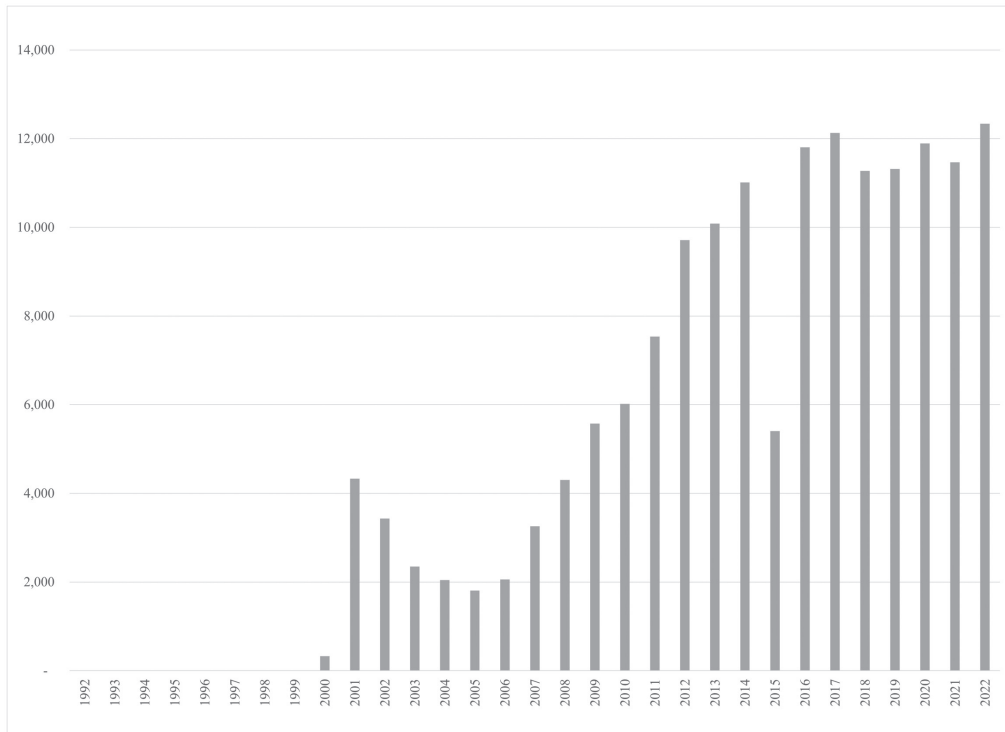


Figure 5.2 Number of disappeared .lu domains per year (1992–2022).

What can domain name registrations and deletions tell us about the beginning of the Internet in Luxembourg?

Calculating the percentage of registrations and disappearances has allowed us to better understand the evolution of the Luxembourg web by identifying key moments and contextualizing the meaning of this data to explain the evolution of the Internet in the country in the early years.

To go further, we contextualized it within the national context using other sources, such as statistics, reports, and press articles, based on the premise that Internet appropriation and adoption was not a linear process, and several factors had an impact: Different stakeholders, paths — taken or not — or imaginaries, as well as the national and international context (Schafer, 2015; Schafer et al., 2016).

For example, from the World Bank data, we can see that only 2% of Luxembourg’s population were Internet users in 1995. In 2000, 23% of the population were Internet users. In 2000–2001, 36% of the population, and in 2001–2002, 40% of the population. The numbers grew in 2003 and 2004 to 55% and 66% of the population (World Bank, 2023).

Press articles (Vandermeir, 2001; “8% (seulement) des entreprises”, 2001) and specific studies stated companies’ slow web usage in Luxembourg, especially for e-commerce activities (Meretens et al., 2000; Otjacques, 2001; Poussing, 2002). According to Poussing (2002), 48.9% of the companies studied in 2001 were on the web, whether with their site or as a reference on another site, where the information of the company was accessible via links or banners.

Going back to the numbers, we already mentioned that 2000 was the year with the highest increase of registrations while the years 2001, 2002, and 2003 displayed the highest deletion rate of the 30 years of our analysis (1992–2022). This could be interpreted as the effect of the potential impact of the .com burst. Nonetheless, we are cautious in our conclusions on the implications for a nascent sector based on just looking at one dimension (deletion rate) and more analysis should be done in this respect. It will be explored further by one of the authors, Noguera in her forthcoming PhD thesis conducted on the history of the Internet and the web in Luxembourg, to be disseminated in 2025.

Lifetime domain names

For the study of the lifespan of domains, Restena provided data for the domains registered between 1 January 2000 and 31 December 2016. The most prevalent lifespan — time-line that a domain has been active in the national registry list — of the domain names is 12 months (22,529, representing 12.8% of the total sample), followed by 24 months (5.6%), 13 months (4.4%), 36 months (3.4%), and 25 months (3.4%). This would seem to reconcile with the offers advertised on the Restena platform *my.lu*, where subscription and renewal are available for a period of 1 or 2 years. Figure 5.3 shows that, out of all domains registered after 2000, only 43 domains have lived — have been active — more than 23 years (0.02%), and only 3.4% of domains stayed alive for more than 19 years.

Taking as a premise that the leading governmental and cultural institutions and the largest enterprises in the country may have been registered before 2000 and that these institutions might have been stable across the years, we decided to do a close reading of the domain names registered and deleted from 1992 to 1995 to understand which of the registered ones are still active today, which categories have survived the longest and which categories have

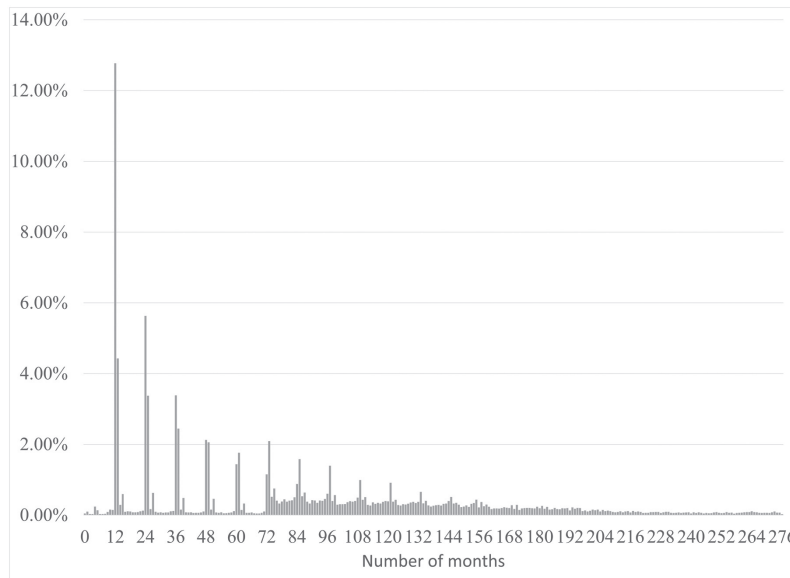


Figure 5.3 .lu domain name lifetimes (2000–2016).

Exploring the evolution of .lu domain names

recorded a more significant number of disappearances. To do that we filtered the domain names created on 31 December 1995 or before and checked whether they still existed in March 2023.

We could see that 57.9% of the domain names registered in 1992 are still active today (as of March 2023). Of those registered in 1993, 72.7% are still active, 61.9% from 1994, and 47.8% from 1995. That means that for the entire period from 1992 to 1995, 50.3% of the domain names registered still exist (153 out of 304). Most of the domain names that are still alive today correspond to education sites (especially the high schools), representing 31.4%, followed by commercial sites (29.4%), miscellaneous (organizations, sports clubs, etc., 9.2%), research (8.5%), and art and culture (5.2%) sites. Notably, these percentages are impacted by the lower number of domains in the first years, with a relevant increase in 1995, adding more volatility to the results and explaining why the trend is downward due to the higher number of registrations and deletions. Doing a closer analysis, we can see that from the existing domain names, 56.9% still have a website, of which 25.3% have been redirected to a new URL. In addition, by checking the URLs manually, we saw that 72.4% of the domain names with a website today cover the same topic as in the previous years, against the 21.8% that have changed or the 5.8% that we could not assess due to error pages — which could have been for several reasons, among them as an outcome of link rot⁷ — or that the IA did not archive them.

The categories most affected by the disappearance of domain names from 1992 to 1995, in aggregate, were education (39.7%), art and culture (20.5%), and commercial (15.9%). Nevertheless, we need to take this data cautiously. In 1995, we saw that some educational, art, and cultural institutions registered several domain names playing with different variants, in many cases showcasing the country's multilingualism, so these variants are more likely not to be renewed. Finally, we could see that most of the sites were archived by the IA (86.9%), although with some limitations, as mentioned.

Luxembourg web through web archives

Luxembourg started to preserve the national web as part of its digital heritage through the BnL in the mid-2010s⁸, first with a collection of state sites in 2015 with the collaboration of the *Centre des technologies de l'information de l'État* (CTIE), then with the launch of its first extensive collections in 2016 (Els et al., 2021).

It was not until 2004, with the law for the reorganization of the cultural institutions of the state, that Luxembourg officially introduced the preservation of websites as part of its legal deposit, specifically under the mission of the National Library of Luxembourg, confirmed by the Grand-Ducal Regulation on legal deposit in 2009, later amended on 21 December 2017. In December 2022, a new law was put in place, modifying the one of 2004. Since 2016, the BnL has been preserving the national web, acting as the web archive's operator and responsible for the legal deposit of the Luxembourg web. The IA has been their service provider for harvests since 2016. These services were expanded in a public tender in 2021 and 2023, which was awarded to the IA on both occasions.

To collect as much of the Luxembourg web as possible, the LWA has followed three main methods using mainly Heritrix to harvest⁹.

- 1 The domain crawls method, which aims to collect many websites a year, has been operated four times a year since 2020, previously twice a year. The crawls take around 5 weeks to

complete, and due to the fast-paced nature of the web, some sites might have changed or disappeared between (or even within) domain crawls.

- 2 Event collections to harvest exhaustive information about an event in a limited timeframe; the seed lists are restricted, but the capture frequency is higher. An example of this was the coronavirus disease 2019 (COVID-19) collection.
- 3 Thematic collections cover a specific topic or field of interest, where the seed lists are evolving, and websites can also be harvested more frequently.

Comparing IA, LWA, and CC

To compare the collections of the IA, LWA from the BnL, and CC for the .lu TLD (Figure 5.4¹⁰), we used a summary of the holdings of each of the archives¹¹, which contains statistics about what domains are in the archive in the different years. The summary was created using a method described in Nielsen, Maurer, and Zierau (forthcoming). The summary includes only successfully archived domains, so domains with no active website or just error pages are not counted.

One of the main findings from this analysis is the positive impact that national and international partnerships have on the completeness of the national collections of the web archiving institutions. Since the beginning of the BnL activities on web preservation in 2016, they have worked closely with key institutions, such as the IA and Restena, positively impacting the development of Luxembourg web archiving. At the national level, the BnL has been working closely with Restena, which regularly provides the complete list of registered .lu domains. This list facilitates BnL's task as it contributes to the crawl's seed list, which is complemented with other seeds found by the IA through GeoIP or language analysis identifying relevant websites outside the .lu TLD. As BnL's service provider for harvests, the IA has vastly improved and enlarged its collection of Luxembourg websites, so the collaboration not only benefits the BnL but also the IA as an international organization as it keeps a copy of the archived data from the BnL.

Before the collaboration and the large-scale national efforts on web archiving, the IA was the main institution preserving the Luxembourg web. It started to archive .lu with 97 second-level domain names. This number grew in the following years, with a greater increase in 2001 with 4,565 domain names. The growth persisted in the following years, except for a decrease in 2006. The biggest increase happened in 2016, reaching 51,871, following which, it stayed within a range between 60,000 and 68,000 domain names until 2021. The increase can be explained by the impact of the partnership between the BnL/IA and BnL/Restena, creating an interesting dynamic in which the institutions both benefit from their collaboration.

On the other hand, the size of the CC collection of the Luxembourg web is much smaller compared to the IA. However, there are relevant overlaps with the IA, especially in 2012 with 6,501, followed by 2010, with 4,259 domain names overlapping. CC started to have a more extensive collection of the Luxembourg web from 2016 onward, with the highest number in 2021 with 20,751.

In turn, the LWA from BnL became a key player as soon as it started preserving the national web with 41,103 domain names in 2016, and then increasing to around 60–66,000 in the following years. As explained, the collaboration with Restena and the IA has been key to guaranteeing the success of its activities. The three had relevant overlaps from 2016 to 2021, with numbers from 7,872 and steadily increasing to 13,576 domain names in the

Exploring the evolution of .lu domain names

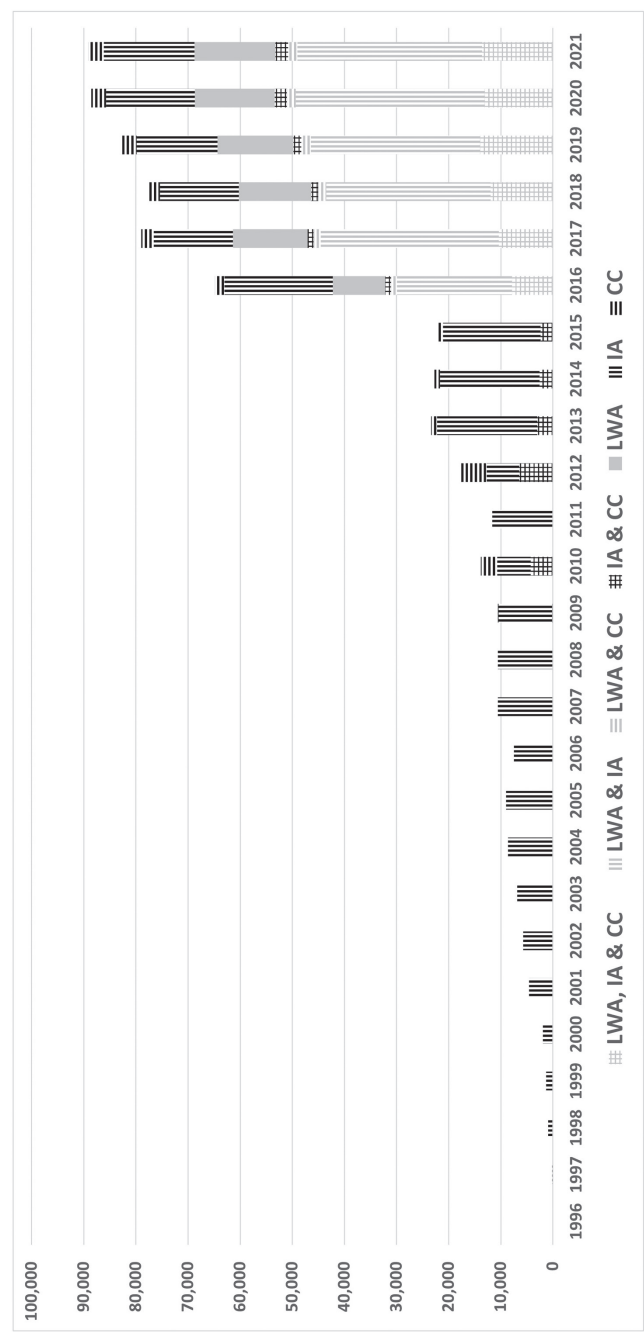


Figure 5.4 LWA, IA, CC .lu domain names overlapping.

three archives simultaneously. Considering only the LWA and IA, we see several overlaps, starting in 2016, which can be explained by the beginning of the activities of the BnL and the transnational collaboration.

To evaluate how different archives represent the Luxembourg web, we compare the data from the three archives with the numbers registered by Restena for each year, focusing only on .lu domains. As the national registry for .lu, Restena has information about all second-level domain names registered in the .lu TLD. Therefore, the number of domain names registered at Restena is an upper bound for any web archive. Some second-level domains are not archived anywhere due to different reasons. The most common one is that no web server was ever registered for that domain, but it could also be that the web server was not answering the crawlers. It is possible that the web server only returned error messages and never a real page with HTTP status 200. While the latter case could be retraced in web archives, those sites are not counted in the CDX Summary and do thus not appear in Figures 5.4 and 5.5. As Restena's number is the upper bound, we can calculate the percentage of domains archived in each web archive. The objective of this comparison is to estimate the representativeness of the web archives.

The IA has historically been the largest preserver of the Luxembourg web (Figure 5.5), with a high percentage of domain names included in its archives in proportion to the numbers of the registry, especially as of 2016. This increase in the percentage from 2016 to 2020 — as we can see in Figure 5.5 — was caused by the partnership between the BnL and IA and the

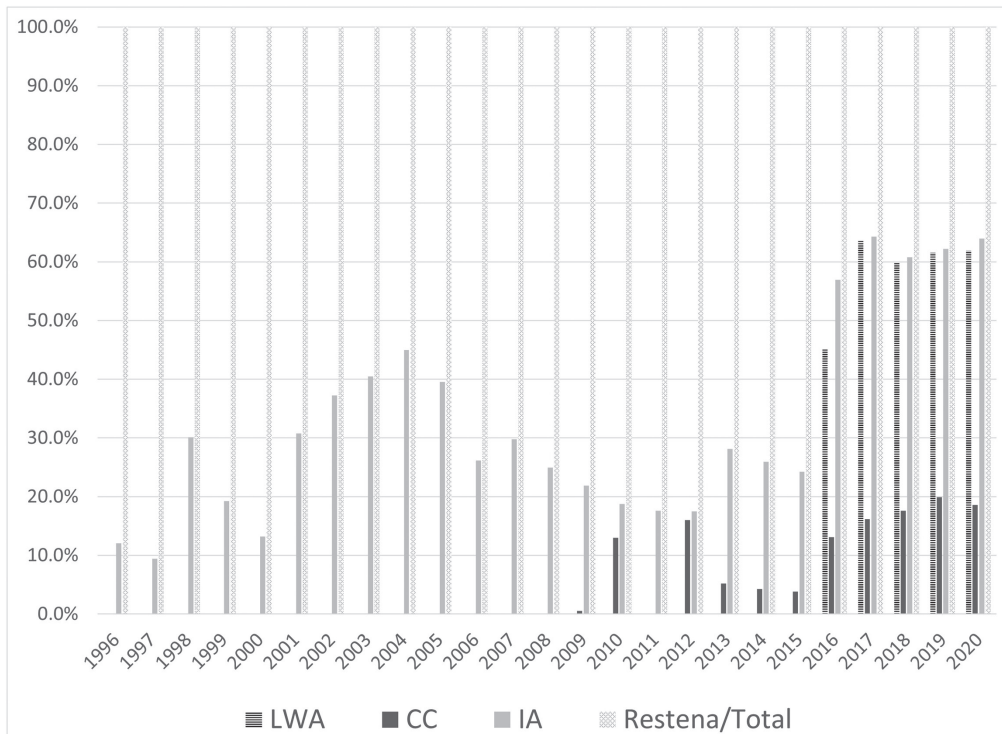


Figure 5.5 Completeness of web archives for the .lu domain in Luxembourg (1996–2020).

Exploring the evolution of .lu domain names

fact that the BnL receives the annual registry list from Restena. CC has been complementary to both, while its share is still less prominent in comparison with the total number of domain names registered at Restena.

In conclusion, we can say that LWA and IA archives represent a similar proportion of the Luxembourg web, partly because they crawl more often than CC, so they can get short-lived domains. In addition to the four domain crawls contracted by the BnL, the IA runs additional crawls for .lu websites, which might vary from one site to another. This can explain why some domain names are only included in the IA collection.

Comparing the Luxembourg web and the Danish web

An obvious place to start our comparisons of the two national webs through the lens of the development of the ccTLDs, is to study the growth of each web as represented by the total number of domains over time. Even though we are inspired by Brügger et al. (2017), we do not use the numbers from the chapter in our comparison of the evolution of the national domains as they only include the years 2005, 2009, 2012, and 2015. To make it more comparable with Luxemburg, we use data from the website of the Danish registrar, DK Hostmaster, which includes statistics of the development from 1997 to 2022 (.dk statistics, n.d.).

First, it is essential to note that the number of domains is very different, hence the use of two y-axes (.dk on the left and .lu on the right, Figure 5.6). In 2022, the number of .lu

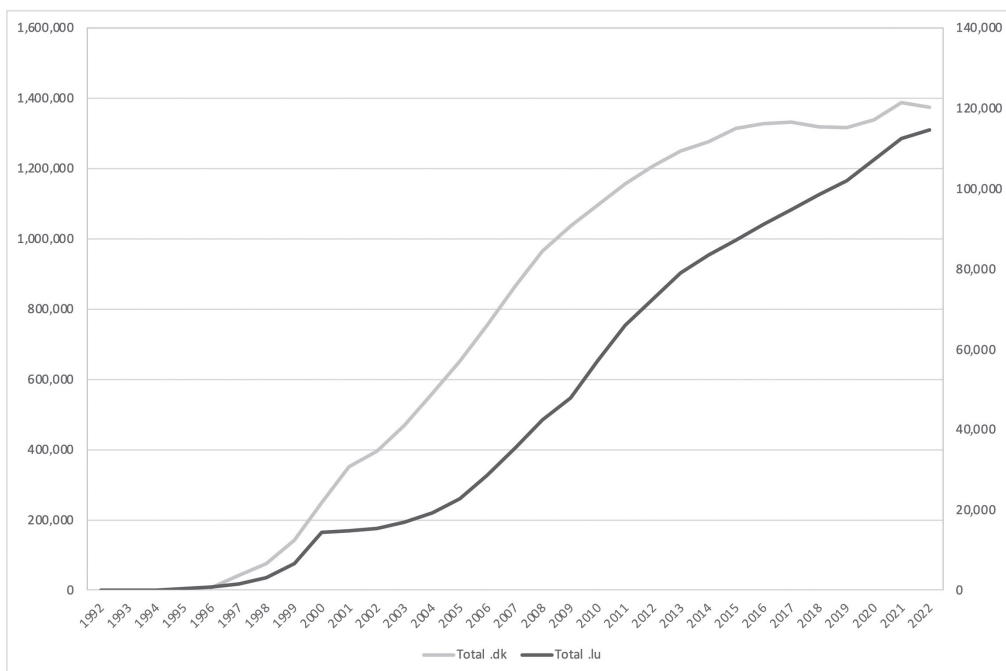


Figure 5.6 Comparison of total registered domain names .dk and .lu (1992–2022, left y-axis .dk, right y-axis .lu).

domains was 114,693, while the number of .dk domains was 1,373,174, so the size of the Danish web as defined by the number of domains on the ccTLD was almost 12 times larger at that point in time. The population of Denmark is approximately nine times larger than the population of Luxembourg (World Population Review, n.d.), so this is part of why there are so many more domains.

We can still compare the developments, though, and see that they follow a similar curve with a relatively steady increase in the total number of domains for most of the period with a tendency to slower growth over time. The .dk numbers increase before the .lu numbers suggesting a later start for many .lu domains, which corresponds to the previous point about many Luxembourg companies being reluctant to go online in the early period. Brügger et al. (2017) point out that even if the number of domains has constantly been increasing over the years measure, “the increase is decelerating. This could indicate that the number of domain names is stabilizing” (Brügger et al., 2017, p. 68). For the .dk the yearly increase from 2009 to 2011 is 6–7% before the increase drops to 2–4% in the following 4 years. From 2016 to 2019, the values are around 1% to –1%, and then there is a slight increase again in 2020 and 2021 (2% and 4%). Looking at the curve, it may seem like the leveling out is a deviation from the norm, but may be the last 2 years are deviating. The rise in the number of domains, following what seems to be a relatively stable number, could possibly be explained by the COVID-19 pandemic, because many people might have tried to move their business online or create websites for hobbies or similar activities. However, more research should be done to support that. In the .lu data, we see a similar pattern of slowing increases, and — once again — this seems to happen a few years later, with numbers increasing around 4–6% from 2014 forward. We do not see the curve level out for the .lu data, but maybe it would have if not for the pandemic. As we do not have data from after the pandemic yet, this is hard to confirm or dismiss.

Another aspect we can compare is the registered and disappearing domains. In the former study, “Figure 3.3 Registered and disappearing .dk domain names over time” (Brügger et al., 2017, p. 69) shows the registered and disappearing domains in three intervals. The number of registered domains seems to decrease steadily over time, while the number of disappearing domains increases but much less so in the later years (a 78% increase from the first to the second period, in contrast to a 13% increase from the second to the third). See Figure 5.7.

In the .lu data, we have the numbers from all the years between 2005 and 2022, except 2015, so instead of intervals we map all the years, which results in a more detailed representation of the development (see Figure 5.8). As we saw in Figure 5.1 and 5.2, there is a relatively steady increase in the number of both registered and disappearing domains leading up to a plateau, first for the registered domains from around 2010 to 2011, then for the number of disappearing domains from around 2012 to 2013. The number of registered domains per year seems to stabilize, while the number of disappearing domains continues to increase, albeit very slowly. The difference between the two lines seems to be relatively stable, especially in the later years, which matches the stable increase in the number of total domains over time.

The most obvious difference between the .lu and .dk data is that the number of registered .dk domains decreases throughout the period, whereas the registered .lu domains increase in most of the early years. Had we only mapped the development until 2015, it would have looked like the numbers were decreasing significantly in 2015, but with the added data, it seems more like a levelling out. As the DK Hostmaster statistics online do not state the number of registered and deleted domains, we cannot determine whether the Danish development would have looked more similar if we had more than three data points and the years

Exploring the evolution of .lu domain names

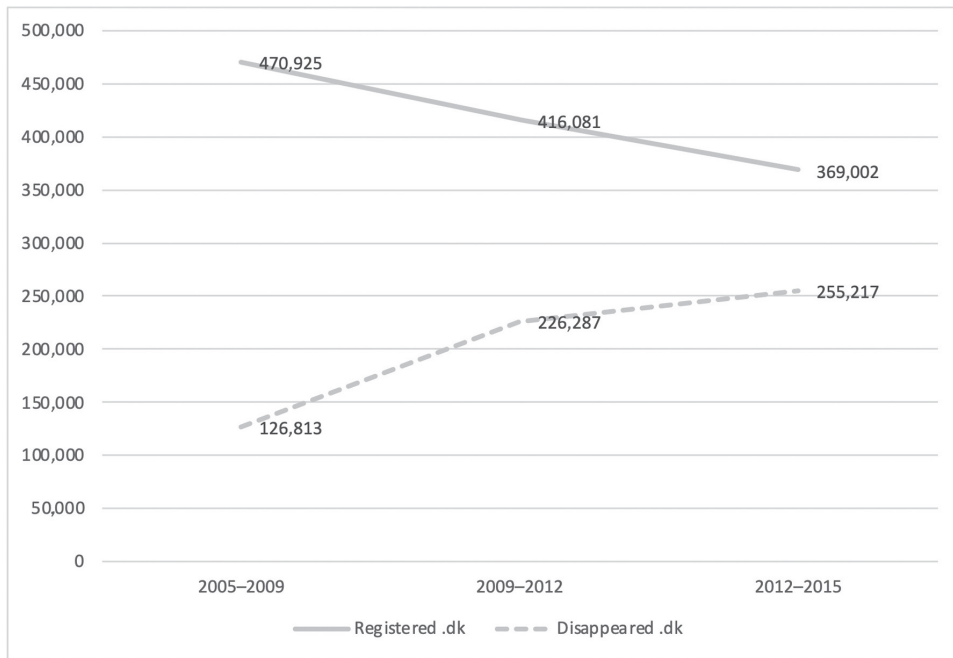


Figure 5.7 Registered vs. disappeared domain names .dk.

Source: Figure from Brügger et al. (2017, p. 69), printed by permission from the author.

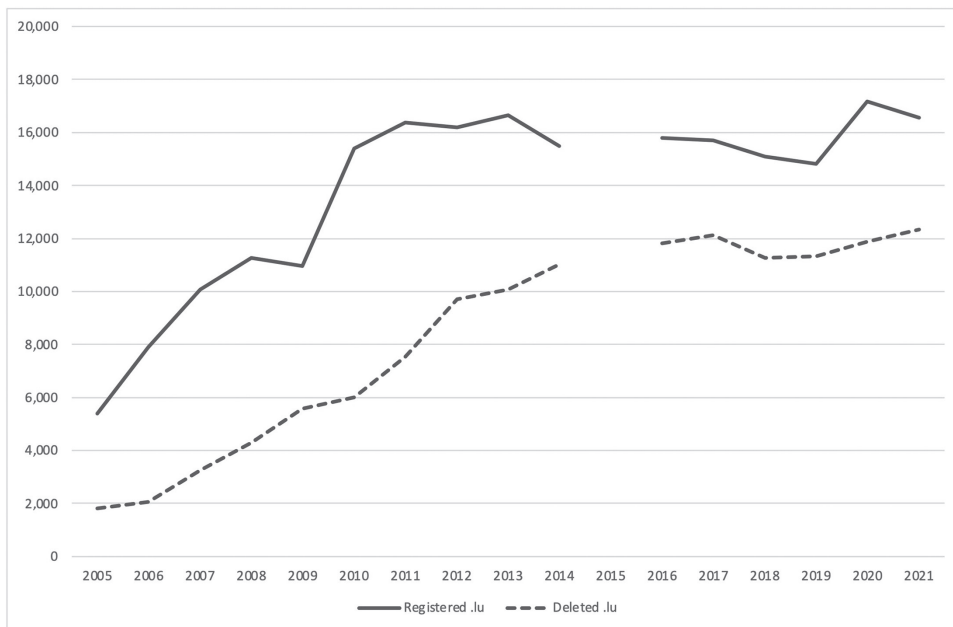


Figure 5.8 Registered vs. disappeared domain names .lu.

after 2015. However, seeing as the growth in the total numbers of domains show similar tendencies across the two ccTLDs, this might be the case, and would be interesting to study further. Either way, the claim made by Brügger et al. that “the relationship over time between the increase in disappeared domain names and the decrease in registered domain names is stable” (Brügger et al., 2017, pp. 69–70) seems to apply to the .lu data as well.

Comparing web archives collections and national registries for transnational studies

Our comparison between registrar lists and archives is inspired by a similar comparison in the Danish study but there are differences that mean that the results are not directly comparable. The .lu analysis is based on Restena’s annual registry list as of the end of the year, whereas the registry lists in the Danish study were lists given by the Danish domain name registrar DK Hostmaster to Netarkivet to be used as seed lists for their harvests. This means that the lists follow the temporality of the broad crawls they were used in. Brügger et al. (2017) include a figure of the number of domains in the .dk registry list and in Netarkivet, respectively, which mentions the dates of the registry lists — 6 October 2005 (for 2006, the very first broad crawl ran from December 2005 to May 2006), 20 January 2009, 1 February 2012, and 5 January 2015. Thus, in contrast to the .lu lists, the .dk lists are from the beginning of each year (except 2006). The data from the archives is not based on the same sources, either. The data from the .lu study is based on the CDX summaries from the web archives (metadata describing the holdings of the archive), whereas the Danish data is based on crawl logs from the archives (metadata describing what has been crawled and archived — or not). The use of different data sources means that we must consider the effect of the differences in temporalities and what is included when we compare, so we are not looking to see whether the numbers are the same but whether there are similar patterns across the studies.

One of the main points from the comparison between the registry lists and the archives in the Danish study are that “more .dk domains are found in the crawled data than on the domain name registry list” (Brügger et al., 2017, p. 72). This difference increases over time, and fewer .dk domains are found in the IA data compared to the registry lists and the Netarkivet data. Still, there are .dk domains only in the IA data (Brügger et al., 2017, pp. 73–74). Whereas the Danish study finds more data in Netarkivet than in the domain name list because new domains appear in the period between when the list is made and until the crawl finishes, in the .lu case the archives only contain around 60% of what is in the registrar lists. The numbers for IA and Netarkivet are very different in the Danish study, whereas our data show a very broad overlap between the LWA and IA data because the collaboration means that the .lu data in the two archives, to a large extent, are the same. However, as Figure 5.5 shows, there are still some .lu domains found only in the IA, similar to the Danish study, which might be explained by the higher frequency of the IA crawls. The numbers suggest that Netarkivet offers the most comprehensive and thorough archiving of the Danish web but that the IA still has a significant role as an archive of the .dk web. Both for .dk and .lu, the IA has a lot of data from before the start of the national archives, and even after, they still have unique domains and versions that are not in the national archives due to the differences in temporality and frequency. Last but not least, you cannot stress too much the importance of the IA’s data being openly available, which facilitates comparative studies.

Conclusion

Throughout this chapter, we have demonstrated how the domain names list from the national registry can be a relevant source to study the evolution of the national web and as a starting point for transnational studies. Through the data (total registrations, inferred registrations, and deletions), we can see the main peaks and declines over the years that can be complemented with other sources such as reports, web archives, press articles, oral interviews, among others, to have a more comprehensive view of the historical evolution of the web. In addition, by doing a close reading of these numbers, we can study who was at the forefront of the web in the early years, the most represented categories, and the most affected by the disappearances.

Because of this data, we have been able to compare the Danish and the Luxembourgish web. We have seen that despite the differences in the numbers, because of the higher population in Denmark, and the methodological questions, such as the different temporalities of the lists impacting the comparisons, we have seen how the registrations and deletions in both countries follow similar patterns. However, with different speeds, probably because of the different rhythm in the appropriation of the web by both countries in the early years (see section “Comparing the Luxembourg web and the Danish web”): The number of domains is stabilizing, and the relationship over time between the increase in the disappearance and the decrease in registered domains is stable.

In comparing the national registry holdings with the national and international web archives institutions to assess the (in) completeness of the web archives, we show how national and international partnerships have been key to increasing the national web’s representativeness by the web archiving institutions, as we have seen with the Luxembourg case. Comparing these results with Denmark, we have shown that, even if we cannot directly compare the numbers in the results, it is still relevant to compare the development of ccTLDs and the holdings of the archives because it helps highlight, once again, the peculiarities between national webs and web archives. Different sources can be used to address similar questions, and the comparisons show the importance of grappling with issues like the temporality and structure of the sources (what is included, what is not, and from when). The ways of representing the developments of a national web frame our interpretations, so it is important to remember that the representations are not absolute but different lenses we apply, which keep some things in focus while leaving others out. Still, we have seen that these lists may have deficits, which underlines a recurring point in web archives studies about the need for documentation. This is a need that is only more pertinent in transnational studies because it can help us understand the influencing factors we must consider in our comparisons.

Acknowledgements

We want to thank Restena for the information provided for the research.

Notes

- 1 See Cows (2017), for an overview of the project.
- 2 Data from new registrations per year based on the monthly data available at dns.lu.
- 3 Total registered domain names for 1992–2010, irrespectively of whether those domain names were deleted or not.

- 4 The detailed statistics published by the website dns.lu, do not give figures for registrations and deletions from September to December 2015 and do not specify the reason for this gap.
- 5 According to Restena, deleted domains are those not renewed at the end of the registration period.
- 6 1993, with an inferred deletion rate of 35% (six domains being deleted), was disregarded due to the low granularity of the data.
- 7 See more about link rot in (Taylor & Hudson, 2000; Wallace, 2004; Dimitrova & Bugeja, 2006; Liebler & Liebert, 2013).
- 8 The CNA (Centre national de l'audiovisuel) also holds the legal deposit for audio-visual web content but does not operate a web archive and has signed an agreement with the BnL to keep the responsibilities of archiving the Luxembourg web with the BnL. In return, the CNA supports the BnL by providing seed lists.
- 9 The BnL uses also a combination of Heritrix and Brozzler web crawlers in some cases, and Webrecorder/Conifer, but mostly Heritrix.
- 10 The IA data for 2020 is not complete (only until November), explaining the greater prominence of LWA without any overlapping on that year.
- 11 Data extracted between September 2022 and April 2023.

References

- 8% (seulement) des entreprises luxembourgeoises sont actives dans l'e-commerce. (1 April 2001). *Le Mensuel d'Agéfi Luxembourg*. <https://agefi.lu/Mensuel-Article.aspx?date=Apr-2001&mens=69&rubr=539&art=5492&query=commerce>
- Baeza-Yates, R., Castillo, C., & Efthimiadis, E. N. (2007). Characterization of national Web domains. *ACM Transactions on Internet Technology*, 7(2). <http://doi.acm.org/10.1145/1239971.1239973>
- Baeza-Yates, R., Castillo, C., & Lopez, V. (2006). Características de la Web de España. *El profesional de la información*, 15(1), 6–17.
- Ben-David, A. (2014). Mapping minority webspaces: The case of the Arabic webspace in Israel. In D. Caspi & N. Elias (Eds.), *Ethnic Minorities and Media in the Holy Land* (pp. 137–157). Vallentine Mitchell Academic.
- Ben-David, A. (2016). What does the web remember of its deleted past? An archival reconstruction of the former Yugoslav top-level domain. *New Media & Society*, 18(7), 1103–19. <https://doi.org/10.1177/1461444816643790>
- Ben-David, A. (2019). National web histories at the fringe of the web. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities: The Case of National Web Domains* (pp. 89–109). Routledge.
- Brügger, N. (2017). Probing a nation's web domain: A new approach to web history and a new kind of historical source. In G. Goggin & M. McLelland (Eds.), *The Routledge Companion to Global Internet Histories* (pp. 178–187). Routledge.
- Brügger, N. (2019a). A national Web Trend Index. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities: The Case of National Web Domains* (pp. 178–187). Routledge. <https://doi.org/10.4324/9781315231662>
- Brügger, N. (2019b). National web archives: The land of promise for researchers. *Bladen voor documentatie/Cahiers de la documentation*, 2020(2), 35–42. https://web.archive.org/web/20231030115818/www.abd-bvd.be/wp-content/uploads/2020-2_Brugger.pdf
- Brügger, N., & Laursen, D. (2019a). Historical studies of national Web domains. In N. Brügger & I. Milligan (Eds.), *Sage Handbook of Web History* (pp. 413–427). Sage.
- Brügger, N., & Laursen, D. (Eds.). (2019b). *The Historical Web and Digital Humanities: The Case of National Web Domains*. Routledge. <https://doi-org.proxy.bnl.lu/10.4324/9781315231662>
- Brügger, N., Laursen, D., & Nielsen, J. (2017). Exploring the domain names of the Danish web. In N. Brügger & R. Schroeder (Eds.), *The Web as History: Using Web Archives to Understand the Past and the Present* (pp. 61–80). UCL Press.
- Brügger, N., & Milligan, I. (2019). *The SAGE Handbook of Web History*. SAGE Publications Ltd. <https://doi.org/10.4135/9781526470546>

Exploring the evolution of .lu domain names

- Brügger, N., Nielsen, J., & Laursen, D. (2020). Big data experiments with the archived Web: Methodological reflections on studying the development of a nation's Web. *First Monday*, 25(3). <https://doi.org/10.5210/fm.v25i3.10384>.
- Brügger, N., & Schroeder, R. (Eds.). (2017). *The Web as History: The First Two Decades*. UCL Press.
- Chambers, S., Mechant, P., & Geeraert, F. (2019). Towards a national web archive in a federated country: A Belgian case study. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities: The Case of National Web Domains* (pp. 29–44). Routledge.
- Chrillesen, N., & Swain, D. (2020). *Restena: L'épopée D'une Infrastructure = Restena: The Story of an Infrastructure*. Fondation Restena.
- Cowls, J. (2017). Cultures of the UK web. In N. Brügger & R. Schroeder (Eds.), *The Web as History: Using Web Archives to Understand the Past and the Present* (pp. 220–237). UCL Press. <https://doi.org/10.14324/111.9781911307563>
- Dimitrova, D. V., & Bugeja, M. (2006). Consider the source: Predictors of online citation permanence in communication journals. *Portal-Libraries and the Academy* 6(3), 269–283. <https://doi.org/10.1353/pla.2006.0032>
- .dk statistics. (n.d). *Registered Domain Names*. https://stats.punktum.dk/en/domains/total_domains/ (accessed 24 May 2023).
- Dns.lu. (2001). *La nouvelle Charte de Nommage en application depuis le 1er novembre 2001*. <https://web.archive.org/web/20011116221851/www.dns.lu:80/fr/> Archived 16 November 2001. Accessed 4 April 2022.
- Dns.lu. (20 March 2023). *A Propos/Chiffres Clefs*. <https://web.archive.org/web/20230320232231/https://dns.lu/fr/a-propos/chiffres-clefs> (accessed 4 April 2023).
- Els, B., Maurer, Y., & Schafer, V. (2021). Les archives du Web luxembourgeois. Histoire, enjeux et perspectives. *Hémecht: Revue d'histoire luxembourgeoise*, 73(2), 183–204.
- Gomes, D., & Silva, M. (2005). Characterizing a national community web. *ACM Transactions on Internet Technology*, 5(3), 508–531. <https://doi-org.proxy.bnl.lu/10.1145/1084772.1084775>
- Hale, S., Yasseri, T., Cowls, J., Meyer, E. T., Schroeder, R., & Margetts, H. (2014). Mapping the UK webspace: Fifteen years of British universities on the web. In *Proceedings of the 2014 ACM Conference on Web Science (WebSci 14)* (pp. 62–70). Association for Computing Machinery. <https://doi.org/10.48550/arXiv.1405.2856>
- Klein, C., & Peltier, F. (2022). *La démographie luxembourgeoise en chiffres 2022*. Statec. <https://web.archive.org/web/20231030085813/https://statistiques.public.lu/fr/publications/series/en-chiffres/2022/20220511.html>. Archived 30 October 2023.
- Lasfargues, F., Oury, C., & Wendland, B. (September 2008). Legal deposit of the French Web: Harvesting strategies for a national domain. In *8th International Web Archiving Workshop (IWA08)*. Aarhus, Denmark. <https://web.archive.org/web/20240303111017/https://bnf.hal.science/hal-01098538/document>
- Liebler, R., & Liebert, J. (2013). Something Rotten in the State of Legal Citation: The Life Span of a United States Supreme Court Citation Containing an Internet Link (1996–2010). *Yale Journal of Law and Technology*, 15(2), 273.
- Meretens, J., McNeill, G., & Tonhofer, F. (2000). *Luxembourg Business in the New Digital Economy: Report*. Centre d'études prospectives.
- Milligan, I., & Smyth, T. J. (2019). Studying the Web in the shadow of Uncle Sam: The case of the .ca domain. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities: The Case of National Web Domains* (pp. 45–63). Routledge. <https://doi.org/10.4324/9781315231662>
- Miranda, J., & Gomes, D. (2009). Trends in Web characteristics. In *Proceedings of the 2009 Latin American Web Congress (la-web 2009)* (pp. 146–153). IEEE Computer Society.
- Musso, M., & Merletti, F. (2016). This is the future: A reconstruction of the UK business web space (1996–2001). *New Media & Society*, 18(7), 1120–1142. <https://doi.org/10.1177/1461444816643791>
- Nielsen, J., Maurer, Y. & Zierau, E. (forthcoming). Comparing the holdings of closed national web archives through summaries. IN THIS BOOK.
- Otjacques, B. (2001). *Commerce électronique et Internet dans les entreprises du Grand-Duché de Luxembourg: État de la situation à la fin de l'année 2000: Rapport de l'enquête*. Chambre de Commerce.
- Poussing, N. (2002). *Les entreprises et les Technologies de l'Information et de la Communication*. Statec.

The Routledge Companion to Transnational Web Archive Studies

- Restena. (1997). *Fees for Domain Name Registration*. https://web.archive.org/web/19980204052010/www.dns.lu/fees_1997_01.html (accessed 4 April 2022).
- Restena Foundation Press Release. (21 December 2020). *30 Years of History for Restena's 20 Anniversary*. <https://web.archive.org/web/20231025133657/www.restena.lu/en/news/250-30-years-history-restenas-20th-anniversary>. Archived 25 October 2023.
- Restena Foundation Press Release. (29 April 2022). *30 Years for .lu*. https://web.archive.org/web/20231025134316/www.restena.lu/files/inline-images/COM_30ansLu_EN_20220429.pdf. Archived 25 October 2023.
- Rogers, R., Weltevrede, E., Borra, E., & Niederer, S. (2013). National web studies: The case of Iran Online. In J. Hartley, J. Burgess, & A. Bruns (Eds.), *A Companion to New Media Dynamics* (pp. 142–66). Wiley-Blackwell.
- Schafer, V. (2015). Part of a Whole: RENATER, a Twenty-Year-Old Network within the Internet. *Information & Culture*, 50(2), 217–35. <https://doi.org/10.7560/IC50205>
- Schafer, V. (2018). Exploring the “French Web” of the 1990s. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities: The Case of National Web Domains* (pp. 145–160). Routledge.
- Schafer, V., Musiani, F., & Borelli, M. (2016). Negotiating the web of the past. *French Journal for Media Research*. La toile négociée/Negotiating the web, 6. <https://web.archive.org/web/20231025134525/http://frenchjournalformediaresearch.com/lodel-1.0/main/index.php?id=952>. Archived 25 October 2023.
- Schafer, V., & Thierry, B. (2012). *Le Minitel: L'enfance numérique de la France*. Nuvis.
- Schafer, V., & Thierry, B. (2016). The “Web of pros” in the 1990s: The professional acclimation of the World Wide Web in France. *New Media & Society*, 18(7), 1143–1158. <https://doi.org/10.1177/1461444816643792>
- Taylor, M. K., & Hudson, D. (2000). “Linkrot” and the usefulness of Web site bibliographies. *Reference & User Services Quarterly*, 39(3), 273–277.
- Teszelszky, K. (2019). The historic context of web archiving and the web archive: Reconstructing and saving the Dutch national web using historical methods. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities. The Case of National Web Domains* (pp. 13–28). Routledge.
- Vandermeir, M. (7 April 2001). Les entreprises ignorent internet. *Le Soir*.
- Wallace, K. (2004). A longitudinal study of Web pages continued: A consideration of document persistence. *Information Research*, 9(2). <https://web.archive.org/web/20040122034748/http://informationr.net/ir/9-2/paper174.html>. Archived 22 January 2004.
- Winters, J. (2019). Negotiating the archives of UK web space. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities. The Case of National Web Domains* (pp. 75–88). Routledge.
- World Bank. (2023). *Utilisateurs d'Internet (% de la population) — Luxembourg*. <https://web.archive.org/web/20231030124058/https://donnees.banquemondiale.org/indicateur/IT.NET.USER.ZS?locations=LU> Archived (accessed 30 October 2023).
- World Population Review. (n.d.). *Total Population by Country 2023*. <https://worldpopulationreview.com/countries> (accessed 24 May 2023).
- Zabicka, P., & Matejka, L. (2007). Czech Web archive analysis. *New Review of Hypermedia and Multimedia*, 13(1), 27–37.

6

COMPARING NATIONAL WEB DOMAINS ACROSS NATIONAL WEB ARCHIVES

Methodological and practical challenges of doing transnational studies

Niels Brügger

Introduction

This chapter aims to discuss the use of national web archives in transnational studies. The topic is explored by investigating to what extent it is possible to make a comparative study of several countries' national web domains, based on the holdings in different national European web archives.

Within the last decade studies of national web domains have started to emerge (Brügger & Laursen, 2019b; for overviews, see Brügger & Laursen, 2019a, pp. 415–416; Brügger et al. 2020, pp. 1–2), whereas publications on transnational web archive-based studies are not widespread. A few journal articles exist that establish an inventory of national web archives' collections, strategies, and access forms (e.g. Vlassenroot et al., 2019), but with less focus on the researcher's perspective.

One possible reason for this scarcity is that researchers find it difficult to get insight into the feasibility of such studies. This gap in the academic literature was the main reason for launching the international researcher network WARCnet, 'Web ARChive studies network researching web domains and events'. One of the Working Groups, 'Comparing entire web domains', aimed at comparing national web domains, including:

reflections on how to define a national web domain, as well as standards for describing web domains, for mapping the available archived web content in different web archives, for establishing a corpus that can be said to represent a national web domain, and for the possible tools to be used.

(WARCnet, Working Groups, 2023a)

The WARCnet team provided valuable insights into how web archives reflect on some of the topics mentioned in the quote above, in particular in relation to a transnational pandemic like

COVID-19 (see WARCnet, Papers and Special Reports, 2023b). However, the ambition of providing more detailed knowledge about the relevant web archives with a view to enabling transnational comparisons proved to be more difficult to live up to than expected. This is where this book chapter starts.

With a view to providing a standard of comparison for this investigation an already existing historical study of a national web domain — the Danish web — is used as a point of departure. The study ‘Probing a nation’s web domain: The development of the Danish web 2005–2015’ was based on the holdings of the national Danish web archive Netarkivet at the Royal Library, it was launched in 2015, and it has resulted in a number of publications (Brügger, 2018, 2021a; Brügger et al., 2017, 2019, 2020; Nielsen, 2019, 2021a, 2021b). In the following, it is discussed if this study can be replicated in four other European countries — France, Luxembourg, the Netherlands, and Great Britain — based on their national web archiving activities. The Danish study and its potential expansion to a transnational project in the present context constitutes an extreme case (Flyvbjerg, 2006, pp. 229–230) in the sense that its scale helps further our general understanding of the landscape that comparative web archive studies are facing, no matter the size of the study.

The chapter is guided by the following research questions: What does the web archive landscape look like for a researcher who wants to make a transnational study of different nations’ web domains across different web archives? — and how do the differences and similarities of the web collections affect how the researcher can navigate this landscape?

The historical development of a national web domain: The Danish case

Before debating the research questions, a brief introduction to the Danish study is needed. The project aimed at making the first digs into how the Danish web domain had developed over a 10-year period, including studies of file types, website size, hyperlinks, trackers, and the written text on all web pages. In addition, the project aimed at becoming familiar with the archived web in Netarkivet.

The project is described in greater detail in Brügger (2018), Brügger et al. (2020), and Nielsen (2021b), but for the present chapter a few issues must be foregrounded because in some cases Netarkivet constitutes an outlier among national web archives (about Netarkivet, see Laursen & Møldrup-Dalum, 2017). First, a high-performance computer was available which was an essential prerequisite for performing the study at all. In addition, Netarkivet makes four broad crawls of the Danish web domain every year, which is not a very widespread practice, but it implied that it had to be determined which broad crawl should be used for each year as the project only wanted one instance of each individual web entity per year. Also, Netarkivet uses harvest IDs to identify each crawl, and job IDs to identify the individual crawl jobs that form a harvest, which enabled the combination of broad crawls with elements from selective crawls, based on harvest and job IDs and an analysis of temporal overlaps between the broad crawl and the candidates to be included from the selective crawls. Finally, it became clear that the same web entity (e.g. an HTML document on a specific URL) was very often archived more than one time in the same broad crawl because links in what was already archived pointed to it. These instances that were not necessarily identical were called ‘versions’, in contrast to duplicates that are exactly identical files (about versions/duplicates see Brügger et al., 2020, pp. 5–6; Brügger et al., 2023). Analyses showed that approximately 50% of a broad crawl was versions, and therefore an algorithm was developed to filter the versions to have only one version in the dataset. This algorithm was run on the seed list of

Comparing national web domains across national web archives

any given crawl job's ID; hence it was imperative to have access to seed lists, harvest ID and job ID that Netarkivet provided.

Analytical focus, methods, and sources

Instead of providing a detailed account of each web archive's activities as such, the chapter focuses on identifying several continua or *scales* on which these activities can be placed. These different scales represent the structural extremities of the landscape that a researcher must navigate. Thus, the main objective is not to know that web archive X is based on legal deposit and considers 'the web nation' in one way, whereas web archive Y is not based on legal deposit and conceives 'the web nation' in another way, but rather to identify 'legal framework for web archiving' and 'conceptualizing the borders of the web nation' as the scales on which the two different legal approaches and understandings of 'the web nation' are placed. This approach enables the potential use of the insights beyond this chapter in relation to other web archive collections and other research projects, and it has a greater chance of remaining relevant in the future because it is not dependent on how individual web archives act here and now.

Interviewing

One of the key takeaway points from the study of the Danish web was that it takes curators, technical, and legal staff to know a collection, and for researchers, it is pivotal to have access to knowledge about how and why the web archive was made the way it was, and which researcher access forms are possible. Thus, the best way of providing in-depth knowledge about the collections and their potential researcher use is to conduct research interviews with relevant staff which is why this chapter is based on qualitative research interviews (van Selm & Helberger, 2019, p. 162), or to be more precise expert interviews (van Audenhove & Donders, 2019; Bogner et al., 2009; van Selm & Helberger, 2019, p. 162). Expert interviews prove to be important when the aim is to provide technical knowledge, process knowledge, or explanatory knowledge (van Audenhove & Donders, 2019, p. 183; see also, Bogner & Menz, 2009, pp. 52–53).

All interviews were semi-structured using an interview guide that was developed based on the knowledge from the Danish study referred to above, and to ensure systematic comparisons an identical version was used in all cases. In contrast to an iterative and cyclic interviewing process (van Selm & Helberger, 2019, p. 166), the analysis only started once all interviews were made. However, as interviews were conducted over a period of 13 weeks themes that came up in earlier interviews affected the conversation in the following interviews — new themes resonated in the head of the interviewer, and de facto the initial interview guide was continuously and indirectly expanded during the entire interview period. As a consequence, the latest interviews became broader than the first ones, but all in all the conversations remained consistent in terms of overall themes.

One to 2 months before the interviews a contact person at each institution was emailed to set up agreements (this person was known to the interviewer beforehand). The interview guide was emailed to the contact person to enable preparation of the interview session, including inviting the relevant staff to participate. Deciding the exact number of participants and the sampling was made by the contact person, but not in a random way. Each group of questions in the interview guide stipulated relevant staff types to involve

(curator, legal, technical, management), and the contact person then set the team to be interviewed, based on resources, availability, and expertise. In practice, no sampling was needed as, in most cases, only very few, or only one person, would be relevant. Therefore, a satisfactory level of saturation was achieved (van Selm & Helberger, 2019, p. 167), and the interview participants mirror the available expertise at each web archive. Because of this selection procedure, interviews became rather heterogeneous regarding the number of participants and staff type, ranging from 2 to 8 persons, and from only technical and management or technical and curator to technical, curator, legal, management, and research support.

All interviews were conducted by the author of this chapter; they took place at the premises of the interviewees, and they were recorded. After each interview, the interviewer made brief field notes to keep track of immediate impressions and reflections. Interviews were coded by using audio-coding in NVivo (Brinkmann & Kvale, 2015, p. 206; Stonehouse, 2019). Before publication interview participants have had the opportunity to correct errors or add relevant facts.

Document analysis

Document analysis has also been used in cases where the web archives provided written documents. In most cases documents were referred to during the interview session, either they were mentioned in passing or they were shown on a screen in the room. In four cases documents were handed over to the author, either right after the interview session or in the following weeks.

Karppinen and Moe (2012) make an analytical distinction between documents as a source and as a topic of study (pp. 184–189). As the present study does not focus on the discourses related to the web archiving process, but rather aims at better understanding the web archiving processes documents are primarily used as sources. Karppinen and Moe (2012, p. 186) also reflect on how documents can be used to understand the context in which they are embedded, including the relation between document and interview. The present study not only focuses on the content of documents, but it also understands the mere existence of specific types of documents (or of documents at all) as something that bears witness to how the web archive functions as an organization which in itself constitutes valuable contextual information for the interviews. In contrast to Karppinen, Moe interviews were the primary data-gathering method, and documents appeared to constitute unforeseen, but valuable spin-off material that provided additional background knowledge. Thus, documents play a role at two stages in the study, first during the interview session, in cases where they were presented, and they pushed the interview in whatever directions that emerged from the documents, second when preparing the analysis, because they were read in extenso and could be integrated in the analysis. In that sense interviews and documents have cross-fertilized each other.

In combination interviews and documents are used to identify and analyze themes and patterns that are relevant to the research questions, be that themes that were already given via the interview guide, or themes that emerged during the interviews and/or in the provided documents. In practice, the analytical approach is inspired by ‘Thematic Analysis’ as a method, including the six phases involved in carrying out a thematic analysis (Herzog et al., 2019).

Sources

Five interview sessions were performed with the following web archiving institutions: Institut National de l'Audiovisuel (INA), Bibliothèque nationale de France (BnF), Bibliothèque nationale du Luxembourg (BnL), the Royal Library of the Netherlands (KB), and the British Library (BL). These web archives were chosen because they constitute a very diverse group regarding legal frameworks, size, archiving strategy, and time of existence, which can help map the extremities of the landscape that a transnational study is facing. However, despite their differences, they do not cover all types of web archives. The interview sessions generated five interview datasets:

- 1 INA, France, 14 December 2022, 2 hours, two participants (technical, management),
- 2 BnF, France, 15 December 2022, 6 hours, six participants (technical, curator, management, access),
- 3 BnL, Luxembourg, 18 January 2023, 2 hours, four participants (technical, curator, management),
- 4 KB, the Netherlands, 9 March 2023, 4 hours, eight participants (technical, curator, legal, management, research support),
- 5 BL, UK, 13 March, 2 hours, two participants (technical, curator).

In all cases, at least one person participated during the entire interview session. Interviews are referred to in the following by using the acronyms in parenthesis above.

Four web archives provided documents that were either shown or just mentioned during the interview session and then handed over or later shipped to the interviewer. Documents were mainly about web archiving activities and collection policy and strategy.

Analysis and results

The analysis focuses on three main themes: (1) Are there any data to support a study of a national web, and if affirmative, what characterizes these data and their potential usefulness for the researcher?; (2) How can the researcher get access to the data?; (3) Is it possible to enrich the data by combining it with data from other web collections? All themes come with a number of sub-themes as is shown in the following. As previously mentioned, the analysis does not give a detailed account of each web archive's operations as such, rather it identifies various scales on which these operations can be placed within each of the three main themes and their sub-themes.

Finally, a note on terminology. A web crawling that aims at collecting a national web domain is termed 'broad crawl', although some web archives use the term 'domain crawl'. This is done to avoid the confusion that may arise as 'domain' is also part of 'domain name', for instance in a seed list of domain names.

Reflections on organization

How web archives operate within their organization is not in itself related to the main focus of this chapter. However, it constitutes important background information that may explain the characteristics of the web collection and researchers' possibilities of interacting with it, and it constitutes valuable knowledge for researchers who set out to study the web across web archives. Two themes were highlighted in the interviews and when reading documents.

First, although all web archives are part of a hierarchical organization in which they have to find their place, they do this in various ways. Some constitute an organizational unit that fits into the organizational system as would any other unit, whereas others are organized in a more flat and network-like way where staff is only part-time working with web archives because they also have other obligations. Also, the size of the web archive units vary, some are very small and consist of very few full-time staff, whereas others are quite big.

Second, the differences as to the means of communication that underpin the web archive activities are striking: on the one end of a scale one web archive is very paper-based, and produces a wide array of documents, on the other end one web archive is almost exclusively orally based, and the rest is somewhere in-between.

It is important that researchers who want to do transnational studies acknowledge these cultural and structural differences, because they may explain why in one web archive things may easily be done because the web archive unit acts as an almost independent unit, whereas in other cases a heavy and bureaucratic hierarchy may impede things happening that the web archive unit itself would like to do. Also, as to the differences regarding means of communication, it is important to keep in mind that if a web archive is mainly paper-based this may be a function of having to adhere to the general rules of the hierarchical organization, but it may also be a function of having existed for many years, with routines in place, and with changing staff, so a written memory is needed, whereas the more recently established web archive units tend to be less paper-based because their short existence has not yet led to written routines. For the researcher, it can be a challenge to have the needed documentation provided in organizations that are mainly orally based.

Are there any data to study?

The first overall theme is whether web archives can provide any data to support a study of a national web, and what characterizes these data and their potential usefulness for the researcher. This theme comes with four sub-themes: (a) time, (b) space, (c) the web archiving process, and (d) the available documentation.

(a) *Time*: Two topics are relevant to the debate regarding time: when did the web archive start? And are broad crawls performed, and in case they are, how many per year?

One would expect that it is relatively easy to identify when a web archive started, but although a precise start date is often communicated publicly this may cover for a more messy prehistory. In some cases, the web archive as such has started more than once, for instance, one or more test crawls or similar were made before the actual launch of the archive, or the web archive started by doing selective crawls and then later scaled up to do broad crawls. In other cases, material from the Internet Archive that predates the Web Archive's own collections was acquired, either based on a partnership where the Internet Archive performed the archiving for a national library, or based on a retrospective ingestion of relevant material. And in yet another case a broad crawl was performed several years before the library started doing web archiving on a regular basis, but this material was not kept, only the list of archived domains was. Also, web archiving may not have started with a national actor such as a national library, but with research institutions, or national and city archives. Thus, for various reasons identifying a clear-cut start date for all web archives, including the initiating of broad crawls, tends not to be straightforward. The consequence for researchers is

Comparing national web domains across national web archives

that when determining when to start the study they have to evaluate the different start dates on a scale from a precise start date to a variety of start periods. This is important because the material is often made in different ways in the start period, and these historical differences within a web archive have to be taken into consideration.

As the researcher wants to study the development of an entire national web domain year by year it is obviously important to know how often the national web domain was collected. The interviewed web archives cover a scale from no broad crawls are done, to only one per year, to more than one per year. This leaves the researcher with three challenges. First, in the case no national broad crawl exists one must decide how to compensate for this, which could either be done by searching for relevant material in other web collections (see more below), or by selecting one of the many versions that are available in the collection (e.g. the biggest, the first, or patched of more collected items of the same web domain (see also below about versions)). Second, if one or more broad crawls exist they may not have been made at the same time during the year in all web archives which may challenge the temporal consistency across all collections, and differences may also apply as to when the annual domain crawl is made in the same web archive (in one case it follows the same schedule every year, in another, it does not). Third, if more broad crawls are made each year one has to choose between them.

(b) *Space*: The main topic related to space is how each web archive determines the perimeter of their nation on the web. There is a high degree of consistency between web archives as to which criteria to consider when delimiting the nation's web: being related to the nation, written or registered by someone from the nation, or hosted on servers in the country. Usually, more of these are used, but the differences emerge when it comes to which ones are used, and in particular how they are operationalized, including the number of resources (money, time, manpower) that are committed to identifying the web domains that are in line with the criteria.

All broad crawls use the list of country-code Top-Level Domains (ccTLDs) as a basis (acquired from the national domain name registrar), but it is supplemented in different ways and based on different source types. All libraries supplement with relevant generic top-level domains (gTLDs) such as .com, .org, .biz, etc., and where relevant web archives also add ccTLDs from parts of their nation with their own ccTLD, be that overseas territories (e.g. the French DOMTOM), regions within the nation (e.g. .bzh for Brittany), or cities (.london, .paris), and also TLDs related to e.g. products of relevance for a nation, such as .lancome.

These domain names are found through various source types: Yellow Pages, geo-IP, address and phone number on websites, a list from the Internet Archive of websites written in a unique language (e.g. Luxembourgish, but not hosted on .lu), a list with all the domain names registered by someone with an address within the country provided by domain name hosting companies, a list with all hostnames registered by state institutions provided by the National Computing Center, information from the Internet Corporation for Assigned Names and Numbers (ICANN) and from associations handling regional domain names such as .alsace, .corsica, etc. In some cases, only a limited number of these sources are used to identify the extra-ccTLD domain names, and not based on a detailed procedure as to how to identify them, whereas in other cases the final seed list is the fruit of a very extensive and time-consuming work that goes on for months, and that includes a large number of sources.

Finally, in relation to broad crawls space and time interfere with each other as, during the time, it takes to compile the seed list domain names may be deleted, and new ones may be registered. This dynamic of the domain name registration (Brügger et al., 2017, pp. 72–76) is an inherent dilemma of compiling seed lists: the further away from the web archiving's starting point the list is finalized, the more inaccurate the web archiving result will be compared to the changing domain names in that period. In cases where this challenge is acknowledged web archives strive to minimize it in different ways.

In addition, a number of differences as to how web archives configure and scope their broad crawl also apply: some limit their broad crawl by the number of URLs on a website, others scope by download time, some accept crawling to include web domains outside the nation's ccTLD, others do not, some respect robots.txt others do not, just to give a few examples.

As the nation's perimeter on the web is only partly determined in the same way across web archives, depending on the criteria, the resources, and the systematics used to identify domain names, these differences must be acknowledged and taken into consideration where relevant because they have an impact on the nature of the 'nation on the web' in the different collections. This also includes the different ways that the dynamic of registration as well as configurations and scoping is handled.

- (c) *The web archiving process*: In most cases, the web crawling is performed by the web archive itself, but in some cases, it has been outsourced to other actors such as the Internet Archive or Amazon Cloud. These differences may have an impact on the material that researchers can study.

As briefly touched upon above there are differences as to how much work and resources go into the planning of a broad crawl. One web archive prepares the seed list in a few days, others do this in a week or two, whereas one web archive spends several months on this task. Obviously, the degree of detail and the formalization of the preparatory work vary accordingly, from being less detailed to being very detailed and formalized.

In addition, the level of testing and preprocessing vary a lot, from no testing to some testing and checking to running several pre-crawls and similar to heighten the quality of the result and to stay on budget. This also includes different ways of describing the broad crawl, from simply setting up the crawling software to drafting a detailed order form for the management, with a description of the seed list, configurations, etc. The level of preparation and detailing is also partly a function of how the web archiving unit is organized and placed within the organization and how much funding is available, in some cases the preparation is done by one person only, in other cases, several staff members are involved. Also, preparation routines change over time in the same web archive, e.g. the formalized order form was initially only emails with descriptions of settings and more.

Whether or not the level of preparation and continuous monitoring has an impact on the quality of what is archived cannot be evaluated based on the interviews alone. However, it is obvious that it affects the level of documentation that is available for the researcher to study in the form of different types of metadata.

- (d) *Available documentation*: When it comes to documentation a number of differences prevail. The following five scales as to what to take into consideration when evaluating the level of documentation were identified in the interviews:

Comparing national web domains across national web archives

- the level of *preparation* of the broad crawl affects the level of documentation: the longer the preparation takes, the longer the paper trail that documents the process, and the more documentation is potentially available for the researcher,
- the level of *attention paid to systematically documenting* also plays a role, from documentation being merely an almost unnoticed spin-off of the crawl activity, and with no systematic overview of the available documentation to documentation being an integral and acknowledged part of the crawl activity, including an overview of the available documentation is available,
- the *format and the nature* of the documentation vary between raw technical data that is automatically generated as part of the crawl process (crawl log, and the like), via technical information that is handmade (crawl configurations, and similar), to curatorial information about reflections and decisions preserved in an Excel sheet, a wiki, a database, or other; as for the raw technical data this may either just be left as is, or it is analyzed and compiled in extensive reports,
- the level of *publicness* of the documentation is also varying, from the documentation that is meant for internal use only, via internal documentation that may be made available to researchers, to publicly available information (e.g. on a website),
- for any study that wants to look at the past, it is pivotal that *historical documentation* is available; in some cases, documentation is mainly thought of as a here-and-now activity made to make the crawl process run, and documentation is then discarded or forgotten when no more relevant, or documentation may exist back in time, in some cases not preserved in any systematic way and with no registry that can help find it, in other cases preserved as an integral part of the broad crawl.

One of the most important things for the researcher who wants study material in any web archive is to have access to the needed documentation, but the researcher who wants to do a transnational study across the interviewed web archives will have to navigate a very heterogeneous landscape where some web archives provide extensive and very diverse types of documentation, easy to find, ready to use, whereas others only start finding, treating and preparing documentation when researchers are standing on their doorstep. These different levels of transparency and readiness make it difficult for the researcher to make informed choices as to what can be studied across borders. The main reasons for these differences probably relate to the available resources to how the web archive is organized and how its ‘mother organization’ is operating. Creating and providing high-quality and researcher-relevant documentation is time and resource-demanding.

How can the researcher get access to the data?

The first thing to be taken under consideration regarding access is whether a foreigner can get access to the web archive at all. Most web archives give researchers from abroad access to the web collections, but in all cases, the default access is limited to reading room access. Studying an entire national web is not feasible using a reading room access only, and therefore the follow-up theme is whether other access forms are available. This theme is investigated by focusing on three sub-themes: (a) what the researcher can get access to, (b) the available in-house research infrastructure and support, and (c) extraction of data.

- (a) *What the researcher can get access to:* As the default reading room access to the archived web data is not useful for large-scale studies other solutions have to be investigated (see following two points below).

Regarding metadata, that is data about the archived data, all web archives create index files about the collection's content, in most cases CDX files are made because they are needed to replay the content in a Wayback Machine, but one collection (INA) creates three indexes, one for websites, one for videos, and one for Twitter content. In some cases, indexes are openly available for everyone through API access, in others a researcher may get access on an on-demand basis.

In addition to the indexes, the collections are surrounded by a variety of other metadata, some of which are not ready to use for the researcher (e.g. crawl logs, and similar) whereas others can immediately be handed over to researchers (e.g. reports on the content), while in other cases customized lists can be generated (WAT files with links, CSV-files with different types of statistics, etc.). A recurring theme for more web archives is the lack of time and resources to have the raw metadata transformed into something useful for researchers such as customized data files, summary reports, or similar.

The interviews also touched upon two important topics related to access: deduplication and versions (for a definition of 'version' see above). Deduplication is the process where a web entity is removed from a web crawl if a 100% identical copy is already sitting in the collection. Researchers asking web archives if deduplication is performed may get an affirmative answer in most cases, but when digging more closely into this there are differences as to how deduplication is handled. In one web archive (KB), deduplication is not an issue because websites are archived individually and in their totality each time they are archived. Where deduplication is performed it is done in different ways: BnL makes four annual broad crawls of which the first one is considered the base crawl (where nothing is deduplicated) and against which the following three broad crawls are deduplicated, but for web archives where only one annual broad crawl is made deduplication is either not performed, or it is made against the previous years' broad crawls. What is deduplicated also varies; BL deduplicates everything, whereas the Danish Netarkivet which is the starting point of this chapter does not deduplicate HTML files.

As previously explained, the question of selecting among versions arises when the same web entity has been archived more than once in the same archiving. In the Danish study, versions were handled through an algorithm that selected only one of the versions, namely the one that was mentioned in the seed list of any given crawl job. Thus, the question is whether this algorithm could be run on all collections. A prerequisite for doing this is that the seed list is available. In the case of INA the seed list may not be constructed the same way as in the other web archives because crawling is not made using the same software, and in the case of the KB the question of versions is difficult to investigate because websites are archived individually, each in their WARC file. In the BL case, the algorithm will be challenged by the fact that the entire broad crawl runs in just one job, and therefore there is no distinction between harvest ID and job ID (as in the Danish case). The BnL and the BnF both can give access to useful seed lists and seed reports, and it is thus probably possible to run the identical algorithm on these collections.

The researcher who sets out to do a transnational study of a national web is challenged regarding what metadata one can get access to. The good news is that CDX indexes appear to be the type of metadata that almost all collections provide access to, one way or another,

Comparing national web domains across national web archives

although CDX files are also different and therefore can be complex to compare in practice (Nielsen et al., 2024). Beyond CDX files it may not be possible to have generated identical types of metadata in all web archives, for technical reasons or because of resource constraints. In addition, transnational studies across collections are challenged by the fact that some web archives are not deduplicating whereas others are, but in different ways. Finally, the question of selecting only one version is challenging to handle because in some cases this problem does not exist whereas it does in others, but the initial way of handling this through an algorithm may not be applicable in all cases.

- (b) *The available in-house research infrastructure and support:* As reading room access is not a relevant solution for a study of a national web domain the theme of having an in-house research infrastructure and researcher support emerges.

None of the web archives provides a research infrastructure to enable the study of an entire national web domain. However, some research infrastructure and researcher support are available which can be used to support some steps of a large-scale study. Only one web archive has a research infrastructure in place that offers extensive dedicated staff to help the researcher, along with a medium-sized computer environment. The other web archives do not provide a ready-to-use customizable setup, but offer staff to help the researcher and technical setups can be established on a case-by-case basis, in some cases delimited by the available resources, in other cases more extensive help is offered.

Among web archives there is general agreement that they should build close collaborations with researchers, and web archives should not only provide data but rather (also) act as ‘explainers of the data’, as formulated by one web archive, including discussing how the scope of the research aligns with the holdings of the web archive. These interactions come in a variety of forms, in many cases also within the same web archive: researcher in residence, associated researcher, PhD scholars being co-supervised, partnerships with large research projects, and ad hoc help to researchers, which is very widespread. In all cases, researchers are not charged for the help, in almost all cases a contract is signed that describes the collaboration, and in most cases, the more formalized types of collaboration are based on the evaluation of an application. However, the general picture is that web archives are not flooded with researchers who want to study the archived web, which is probably one of the reasons why web archives mainly handle researcher engagement on a case-by-case basis. They all seem to struggle with squaring the circle of, on the one hand, very few researchers wanting to study the collections because a research infrastructure and staff to help are not available, and, on the other hand, web archives do not make these support functions available, because the researcher demand is too little.

- (c) *Extraction of data:* As a research infrastructure for doing large-scale analyses is lacking in all web archives the question of extracting data becomes relevant.

Extraction has not played a significant role for web archives yet, mainly because only a few researchers want to study the archived web, as mentioned above. Thus, when asked about the possibility of extracting web material web archives balance between a clear ‘yes’ or a clear ‘no’, with a ‘yes and no, we’ll have to look into that’ (and similar) as a middle ground. Also, in one case, it is possible to extract and hand over metadata (word frequencies, file types, or

similar), but because this is resource-demanding it cannot be done at scale. As for the web archives where extraction is possible data can be extracted to a researcher at an institution outside of the country of the web archive, based on a contract being signed.

Obviously, the biggest challenge is the web collections where extraction is not possible, but in some cases, valuable metadata about word frequencies, file types, or similar may be extracted and handed over to a researcher. However, extraction comes at a prize for the researcher and for the web archive: web archives would like to be ‘explainers of the data’, and if data are just handed over to researchers this contextual knowledge is lost, even if shared standards for documentation and associated metadata should be established (see ‘Concluding remarks’). Thus, the extremities of this dilemma are best understood as a continuum where some sort of collaboration is established between web archive and researcher, and once researchers are familiar enough with the data it is handed over to them for analysis, in cases where extraction is possible.

Enriching the data by combining collections

The third main theme revolves around enriching data by combining it with data from other web collections. The theme comes with three sub-themes: (a) combining with the web archive’s own holdings, (b) supplementing with other national collections, (c) supplementing with transnational and other nations’ collections.

- (a) *Combining with the web archive’s own holdings*: In cases where web archives not only do broad crawls, but also different types of selective crawls the question arises whether one could supplement the broad crawl with parts of a more complete selective crawl.

Researchers who want to experiment with combining two types of collections within the same web archive are facing various challenges to which the answer is not the same in all web archives. In some cases, combining a broad crawl with a selective crawl is not an issue because there exists no broad crawl (INA, KB), instead one could compile a version of each website that is as complete as possible within the time frame of the time slice one investigates. In cases where combining is possible some web archives suggest going by CDX index, others by using the crawllog, and in the Danish case the combination of more collections was based on harvest and job IDs and an analysis of temporal overlaps between the broad crawl and the candidates to be included. In addition, it is very likely that what is combined with a broad crawl is unevenly distributed, as selective crawls vary in terms of number, depth, frequencies, etc.

- (b) *Supplementing with other national collections*: Supplementing a broad crawl by combining it with other collections from the same country was only discussed in the interviews in relation to the two French national web archives (INA, BnF), and the biggest challenge was evaluated to be the organizational differences between the two institutions. The Dutch web constitutes a special case as no one institution performs a broad national crawl, in contrast, there exists an array of small web archive collections. In this case, the major challenge is to get access to these collections and have material extracted if possible, whereby all the types of differences identified above are multiplied by the number of individual web collections.

Comparing national web domains across national web archives

- (c) *Supplementing with transnational and other nations' collections:* A collection could also be supplemented by combining with relevant web material from outside the country in question, be that transnational collections such as the Internet Archive or Common Crawl, or holdings of other national web archives that deliberately or by accident has archived web content on another country's web domain. Some web archives already have experience with ingesting web material from the Internet Archive that predates the establishment of the web archive, whereas using the Common Crawl is uncharted territory.

Although such combinations may sound tempting, they constitute a challenge for researchers, in the main because they have to identify the relevant material themselves, as these collections are placed outside the web archive in question. This can be done by using these collections' CDX files if available (they are publicly available for both the Internet Archive and the Common Crawl collections), but it demands some technical skills to make this information useful, and only limited information is available which is definitely better than nothing, but it is not the data one may (also) want to study. The data itself then has to be retrieved which is relatively easy for Common Crawl, but less so for the Internet Archive. As for material found in web archives in other countries all of the above topics may make such a study challenging, that is: getting access to the data itself, the level of support, the possibilities of extraction, etc., including just getting access to information about whether web domains from the country in question have been archived at all.

Mapping the landscape

Before discussing the insights of the analysis let us briefly summarize what the landscape looks like that a researcher has to navigate who sets out to make a transnational study across web archives.

This researcher is facing a very heterogeneous landscape in which the many differences between web archives can best be identified, understood, and discussed as a number of scales regarding the organization, what can be studied, what can be accessed, and if/how a web collection can be combined with other web collections. These scales are shared by all of the interviewed web archives, and although similarities exist they all place themselves differently on all the scales.

That web archives are different is not a big surprise, but it is striking to note that they are different in almost all areas. The consequence for researchers is that they must start by familiarizing themselves with how each web archive is placed on each of the scales, followed by an evaluation of how feasible the planned transnational study is, given the many differences between web archives.

Discussion

The interviews revealed that challenges are piling up for the researcher who wants to make a transnational study, and it is obvious that these challenges affect if and how such a study can be made at all, from having only little effect on the study to making it impossible.

However, as the differences are very closely connected to the core of how each web archive operates, how it was established, and how it has developed they are not very likely to disappear (for an infrastructure approach to web archives, see also Maemura, 2021, 2022).

Therefore, it is relevant to discuss how researchers can do transnational studies, given these circumstances, and how they may engage in activities to potentially change the circumstances (the latter question is debated in the ‘Concluding remarks’).

First, researchers have to decide if or to what extent any given difference between web archives is important for the investigation of their research question. For instance, if the study is about entire national web domains it is an important limitation if a country does not perform broad crawls, but if the ‘nation on the web’ is not delimited in exactly the same manner across web archives this limitation may be considered minor.

Second, researchers should consider possible strategies for mitigating the consequences of the differences. An obvious approach is simply to discard any given country from the study if its web archive does not allow for investigating any parts of the research question (the material is not useful, access/extraction is not possible, etc.). This will make the study less complete, but probably more consistent.

Another strategy is to supplement web archive collections that are not as complete as the researcher would like them to be vis-à-vis the research question, which could be done by combining their holdings with other collections. The analysis has indicated that this may be possible, but to a large extent, it is the researcher who has to find and combine the archived web material (about the use of metadata from different web collections, see Nielsen et al., 2024).

A probably more feasible strategy is to identify a minimal part of the research question that can be investigated in all web archives. In cases where access to the data itself is not possible, a minimal study could, for instance, be done by using metadata datasets, and as the analysis has shown CDX files are widespread and they can help do at least some comparative studies (see Nielsen et al., 2024).

This minimal approach can be extended by stratifying what will be studied. If we imagine a study including six web archives a minimum as outlined above can be investigated in all six web archives, another part of the research question may only be possible to study in four to five web archives, and yet other elements of the research question can be studied in only two to three web archives. Such a stratified approach does not allow for a full-scale comparative study but comparing a few web archive holdings is better than comparing none at all.

Third, two prerequisites shall be highlighted that can help researchers, no matter which of the above approaches is chosen. On the one hand, an ongoing dialogue with the involved web archives is key, because the archived web is a malleable source that can be handled in so many ways that knowledge from web archive staff is needed to tailor the object of study to the concrete research project (Brügger, 2021b, pp. 219–221). On the other hand, documentation is key for the researcher — and for the web archive as well — to enable making informed choices.

Concluding remarks

The chapter has shown that a researcher who wants to make a transnational study of different nations’ web domains across web archives has to deal with a very varied landscape of web archives, each characterized by their own way of creating their web collections and making them available for research.

The analytical focus on establishing a number of scales along which similarities and differences between web archives can be identified enables the potential use of the insights of this chapter in other contexts, in relation to other web archive collections and research

Comparing national web domains across national web archives

projects. Obviously, some of the scales could to some extent have been identified without conducting expert interviews, whereas others could not, and in any case the interviews help make the scales concrete and flesh out their nature. It is also obvious that other scales could probably have been identified, had other web archives been included. Also, one theme could have been investigated more in-depth, namely, the organization in which web archives are embedded; this topic was not asked in the interview guide, rather information about it came by accident and in passing, but it could have been investigated in a more targeted way.

To conclude this chapter, three themes shall be highlighted that can help improve the circumstances under which transnational web archive studies are taking place. First, the analysis identified an overall recurring theme, namely that much more could be done in terms of supporting transnational studies had there been more resources available at the different web archives, be that funding and/or staff. The need for more resources is, of course, a difficult agenda to push forward, but the first step is to raise awareness of the need.

Second, as indirectly indicated by the analysis one possible solution to mitigate the many differences between web archives is the development of shared standards for documentation, be that for how a broad crawl is described, or for how metadata is structured and provided, just to mention a couple of examples. ISO standards for web archives exist (Information and Documentation, 2013), but what is needed is shared documentation standards aimed at researchers (see Brügger, 2019, pp. 179, 185–186). Standardized documentation across web archives would constitute an invaluable help for researchers to familiarize themselves with the holdings of the different web collections and to make informed choices as to which comparisons are possible, and it would probably also be beneficial to web archives themselves. Such standardization could be developed within the framework of the International Internet Preservation Consortium (IIPC), with the involvement of relevant researcher communities, and it could be based on best practices in some of the very extended and systematic forms of documentation that have been identified in this analysis.

Finally, web archives and researchers should join hands and engage in transnational studies that can be used as examples, on the one hand, to showcase what can be done with the existing available web collections, and, on the other hand, to highlight the existing limitations and the possible paths to be taken to overcome these shortcomings.

References

- van Audenhove, L., & Donders, K. (2019). Talking to people III: Expert interviews and elite interviews. In H. Van den Bulck, M. Puppis, K. Donders, & L. Van Audenhove (Eds.), *The Palgrave handbook of methods for media policy research* (pp. 179–197). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-16065-4_10.
- Bogner, A., Littig, B., & Menz, W. (Eds.). (2009). *Interviewing experts*. Palgrave Macmillan.
- Bogner, A., & Menz, W. (2009). The theory-generating expert interview: Epistemological interest, forms of knowledge, interaction. In A. Bogner, B. Littig, & W. Menz (Eds.), *Interviewing experts* (pp. 42–80). Palgrave Macmillan.
- Brinkmann, S., & Kvale, S. (2015). *InterViews: Learning the craft of qualitative research interviewing*. Los Angeles.
- Brügger, N. (2018). Probing a nation's web domain: A new approach to web history and a new kind of historical source. In G. Goggin & M. McLelland (Eds.), *The Routledge companion to global internet histories* (pp. 61–73). Routledge.
- Brügger, N. (2019). A national web trend index. In N. Brügger & D. Laursen (Eds.), *The historical web and digital humanities: The case of national web domains* (pp. 178–187). Routledge.
- Brügger, N. (2021a). Digital humanities and web archives: Possible new paths for combining datasets. *International Journal of Digital Humanities*, 2(1), 145–168.

- Brügger, N. (2021b). The need for research infrastructures for the study of web archives. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The past web: Exploring web archives* (pp. 217–224). Springer.
- Brügger, N., Clavert, F., & De Wild, K. (2023). How to handle web archive versions? In *Paper to be presented at the 5th RESAW Conference, Exploring the Archived Web During a Highly Transformative Age*, Marseille, 6–7 June.
- Brügger, N., & Laursen, D. (2019a). Historical studies of national web domains. In N. Brügger & I. Milligan (Eds.), *The SAGE handbook of web history* (pp. 413–427). SAGE.
- Brügger, N., & Laursen, D. (Eds.). (2019b). *The historical web and digital humanities: The case of national web domains*. Routledge.
- Brügger, N., Laursen, D., Nielsen, J. (2017). Exploring the domain names of the Danish web. In N. Brügger & R. Schroeder (Eds.), *The web as history: Using web archives to understand the past and the present* (pp. 62–80). UCL Press.
- Brügger, N., Laursen, D., & Nielsen, J. (2019). Establishing a corpus of the archived web: The case of the Danish web from 2005 to 2015. In N. Brügger & D. Laursen (Eds.), *The historical web and digital humanities: The case of national web domains* (pp. 124–142). Routledge.
- Brügger, N., Nielsen, J., & Laursen, D. (2020). Big data experiments with the archived Web: Methodological reflections on studying the development of a nation's Web. *First Monday*, 25(3). <https://journals.uic.edu/ojs/index.php/fm/article/view/10384>.
- Flyvbjerg, B. (2006). Five misunderstandings about case-study research. *Qualitative Inquiry*, 12(2), 219–245. <http://dx.doi.org/10.1177/1077800405284363>.
- Herzog, C., Handke, C., & Hitters, E. (2019). Analyzing talk and text II: Thematic analysis. In H. Van den Bulck, M. Puppis, K. Donders, & L. Van Audenhove (Eds.), *The Palgrave handbook of methods for media policy research* (pp. 385–401). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-16065-4_22.
- ISO/TR14873:2013(E). Information and documentation — Statistics and quality issues for web archiving. www.iso.org/obp/ui/#iso:std:iso:tr:14873:ed-1:v1:en.
- Karppinen, K., & Moe, H. (2012). What we talk about when we talk about document analysis. In N. Just & M. Puppis (Eds.), *Trends in communication policy research: New theories, methods and subjects* (pp. 177–194). Intellect. <https://doi.org/10.2307/j.ctv36xvj36.12>.
- Laursen, D., & Møldrup-Dalum, P. (2017). Looking back, looking forward: 10 years of development to collect, preserve, and access the Danish web. In N. Brügger (Ed.), *Web 25: Histories from the first 25 years of the world wide web* (pp 207–227). Peter Lang.
- Maemura, E. (2021). *Data here and there: Studying web archives research infrastructures in Danish and Canadian settings*. Faculty of Information, University of Toronto.
- Maemura, E. (2022). *Towards an infrastructural description of archived web data*. Aarhus: WARCnet Papers. https://web.archive.org/web/20230902214729/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Maemura_Towards_an_Infrastructural_Description.pdf
- Nielsen, J. (2019). Experimenting with computational methods for large-scale studies of tracking technologies in web archives. *Internet Histories*, 3(3–4), 293–315.
- Nielsen, J. (2021a). Using mixed methods to study the historical use of web beacons in web tracking. *International Journal of Digital Humanities*. <https://doi.org/10.1007/s42803-021-00033-4>.
- Nielsen, J. (2021b). Quantitative approaches to the Danish web archive. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The past web: Exploring web archives* (pp. 165–179). Springer.
- Nielsen, J., Maurer, Y., & Zierau, E. (2024). Comparing the holdings of closed national web archives through summaries (pp. xx–xx). In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *Companion to transnational web archive studies*. Routledge.
- van Selin, M., & Helberger, N. (2019). Talking to people II: Qualitative interviews. In H. Van den Bulck, M. Puppis, K. Donders, & L. Van Audenhove (Eds.), *The Palgrave handbook of methods for media policy research* (pp. 161–177). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-16065-4_9.
- Stonehouse, P. (2019). The unnecessary prescription of transcription: The promise of audio-coding in interview research. *Research in Outdoor Education*, 17(3), 1–19. <https://doi.org/10.1353/reseout.ded.17.2019.0001>.

Comparing national web domains across national web archives

- Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*. 1(1) 85–111. <https://doi.org/10.1007/s42803-019-00007-7>.
- WARCnet (2023a). *Working Groups, Working Group 1: Comparing entire web domains*. <https://web.archive.org/web/20230905082721/https://cc.au.dk/en/warcnet/working-groups>.
- WARCnet (2023b). *Papers and Special Reports*. <https://web.archive.org/web/20230902093628/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.

7

CONVERSATION 1

Transnational

*Sophie Gebeil, Moisés Rockembach, Gabriele Balbi,
Valérie Schafer, Susan Aasman, Frédéric Clavert, Helena Byrne,
Juan-José Boté-Vericad, Sharon Healy, Ditte Laursen,
Friedel Geeraert, Lise-Anne Denis, Yves Maurer, Karin de Wild,
Janne Nielsen, Saskia Huc-Hepher*

What are the differences between transnational, cross-national, and global web histories?

Sophie Gebeil: In my research, I draw upon my background in history to apply it to the web. Following this logic, I consider cross-national history as a form of comparative international history, following in the footsteps of March Bloch's comparison (1949) of the phenomenon of Thaumaturgic Kings in France and UK during the Middle Ages. Therefore, a cross-national history of the web would be a comparative study of a common phenomenon, such as comparing the French and English web spheres dedicated to lockdowns. Global history operates on a broader scale, akin to Fernand Braudel's approach who wrote a history of structures over the long term and the connections between different geopolitical spaces (Braudel, 1958). Thus, a global history of the web would seek to understand observed phenomena by considering the process of globalisation articulated with specific dynamics according to different territories. For example, understanding the phenomenon of the spread of selfie practices on global social networks by studying the differentiated practices across regions. Finally, a transnational approach goes beyond cross-nationalism and draws from global history. It is a history attentive to circulations, particularly cultural ones, and the interconnections between different web spheres. For instance, a transnational history of migrant practices on the web emphasises how the experience of mobility impacts, through interrelation, both the communities of the migrants' countries of origin and the so-called host societies. The WARCnet project, in fact, served as a platform for exchange and allowed me to discover new interpretations and references regarding these different approaches.

Conversation 1: Transnational

Moisés Rockembach: Adding to the previous explanation, it is important to highlight the nuanced interplay between cultural and technological flows that defines transnational web histories. This transnational approach clarifies how digital culture transcends geopolitical boundaries, allowing for a richer understanding of how web technologies and content evolve in a globally interconnected space. For example, analysing the global influence of memes not only shows their widespread appeal but also how they are localised and adapted in various cultural contexts, demonstrating the web's role as a dynamic medium for cultural globalisation and transnational identity formation.

Gabriele Balbi: This question is not easy to answer and I guess it is linked to the sensitivity of each researcher and, as already mentioned, to the disciplines or fields where she/he comes from. To me, doing transnational research means focusing on flows of information at national level, e.g. analysing the history of the .eu domain requires a dose of trans-national studies with a focus on the role played by each nation in shaping the domain. A cross-national approach to web history, according to my sensitivity, focuses on facts which overcome the national relevance and are important to more than a nation and culture; a cross-national history of the impact and dominance of American web companies in Europe would be a very interesting example. With research questions that can be answered by single-nation researchers (Why did American companies become dominant in your country?), but also with a general sense and meanings. Answers can be very different from country to country, but the general sense is relevant, today, American companies are dominant in most of (all) the European countries. Finally, doing a global history of the web is again different. I start with a definition Paolo Magaudda and myself provided in *A history of digital media: An intermedia and global perspective*:

Global here does not imply that mention will be made of every country in the world but rather that the book will constantly swing between innovation centers and technological peripheries, thus analyzing specific case studies that are capable of bringing out all the geopolitical and cultural tensions and contradictions that digitization generates. As we will see, while production and appropriation of digital media are dominated by specific countries and areas of the globe (the Western world and the United States, at length, joined by Asia over the last decade), digital technologies are also making their way into other parts of the world with various degrees of complexity and some evident ambiguities.

(Balbi & Magaudda, 2018, p. 3)

Doing global histories of the web means taking into account the inspirations, complexities, and ambiguities in regulating, collecting, building infrastructures, and especially using the web. The ways in which Google is used around the globe, how people buy or not buy on Amazon, how they make use of Facebook or Instagram to create communities are just some examples. Exactly like studying how people interact in the digital platforms where these

companies do not exist or are forbidden or where local platforms are simply better and more used (like in South Korea).

How do you conceive the relation between ‘history’ and ‘the web’?

- Valérie Schafer:* The relationship between the web and history can be approached from several angles. The web itself has a history. Moreover, it is a rich source of historical content as well as a medium for disseminating historical documents. One can study how history is written on platforms like Wikipedia. Furthermore, the web serves as both a mirror and a (re) source for historians interested in exploring the digital age, as well as economic, cultural, and other histories. To capture the “spirit of the times,” media sources have long been utilised by historians. We can no longer rely solely on “traditional broadcast media” and must also consider online productions and user generated content. The web is characterised by multiple temporalities, making it a fascinating subject; its presentism or tendency towards constant updatism has been noted. Although there is a state of ephemerality, the web also embraces the past, remixing, recreating content, and constructing and co-shaping (sometimes nostalgic) digital cultures, as seen in the example of memes, to name but one phenomenon.
- Susan Aasman:* The web has multiple temporalities in terms of content, and this extends to the developments of its infrastructure and interfaces. The web today is so much different from the web of yesterday. The affordances of the early history of the web (web 1.0) have been replaced multiple times by revisions, and tracing its history will show how the internet evolved in co-evolution with society (see Sandvig, 2013). Once these underlying structures become lost, a part of the historical context is lost.
- Frédéric Clavert:* I would globally agree with Valérie Schafer and Susan Aasman’s comments. I would add something: The web is a fascinating primary source for those — in *memory studies* — who are interested in how the past is being seen, considered, reworked by groups, societies, Nations, States. It is the ideal medium to study our tendency to fight/have flame wars over the past, to deal with contested pasts.
- Helena Byrne,
Juan-José
Boté-Vericad,
Sharon Healy:* History(s) of the web, and doing history with the web might also be looked at through the lens of web historiography. There are the implications of the web’s transient nature on historical research. The web is not only a source of vast information but also a medium characterised by its ephemerality, where content can be altered or disappear with time. This presents unique challenges and opportunities for historians, prompting the development of new methodologies such as web archiving, which seeks to preserve digital content for future research. Furthermore, the web’s role in democratising access to historical information can be explored, as it allows a wider audience to engage with history, contributing to a more participatory form of historiography.

Conversation 1: Transnational

In explaining web historiography, Brügger (2012) puts forward two distinctions in terms of a web historiography aimed at writing the history of the web, or at a “web-minded” historiography, which considers “the role of the web” in modern society (p. 103). In explaining further, Brügger (2012) notes it is worth understanding that, since the development of the web, it has become “an integrated part of historiography as more and more documents are made available on the web, and the web is part of the historian’s methodological toolbox” (p. 103). Winters (2018) further points to the need to understand “the technological contexts” in which digital media has emerged, in the same way that a “palaeographer knows how parchment and ink were produced in the thirteenth century, and so too the digital historian will be required to know how the internet works, and how algorithms are constructed” (p. 285). Moreover, Winters (2018) notes how historians who are accustomed to working with manuscripts will also know how the manuscript was produced and came into being, “and what this means for any analysis of its content” (p. 286). Therefore, for Winters (2018) the “same is true for digital data and the social, cultural, and technical infrastructures which underpin its creation and transmission” (p. 286).

The web facilitates the formation of collective memories, where shared experiences and narratives are disseminated and preserved online, influencing public perceptions of history and the past. This phenomenon underscores the web’s impact not only on the professional practice of historiography but also on the broader societal understanding of history.

Gabriele Balbi: The most relevant thing which should be underlined (naive in this collection, but relevant in digital studies) is the fact that the web has history. We cannot study the web without considering the political, economic, technological, and cultural background surrounding it in different places and different times. So history is/should be a crucial lens to approach the web.

Then, in my chapter “‘History web’, ‘web history’, and ‘history of the web’” (Balbi, 2024), I claimed that the relations between the web and history have been studied by three mostly separated subfields. First, the “History Web,” where the web is mainly taken as a platform to help historians in their daily work because of digitized sources, databases, and the creation of websites dealing with history. Second, the “Web History,” which mainly focuses on websites, the material and software dimension of the past web, and the archived web. Finally, the “History of the Web,” where the political, economic, technical, and cultural history of the web itself is at the center. On the one hand, these three subfields are separated and very few scholars are active in all of them (some exceptions are present in this collection). On the other hand, we probably need more and more interactions among them and I have tried to provide some examples.

How do you see the field of transnational web archive studies, and what would be the most beneficial way(s) to advance this field of research? The answers can relate to technical, theoretical, methodological, legal and practical aspects of transnational web archive studies.

Ditte Laurssen: The diverse and often conflicting copyright laws and data privacy regulations and their implementation pose significant challenges to researchers engaged in transnational web archive studies. These disparities can hinder access to and the use of web archives across different

countries, affecting the comprehensiveness and reliability of research outcomes. Advancing the field requires a concerted effort toward harmonisation of these legal frameworks and their implementation to ensure consistency. This does not necessarily mean creating a single, universal set of laws but rather developing a set of principles or standards that can be adapted to the specific legal contexts of different countries. Such an approach would facilitate smoother cross-border research activities, enhance the accessibility of web archives, and ensure the protection of personal data in line with international standards. Maybe we could also introduce some more flexible copyright rules just for academic research, or set up international agreements that let researchers use copyrighted content without infringing on anyone's rights. It is all about finding that sweet spot that protects people's privacy and copyright but also keeps the door open for important research and discovery.

Friedel Geeraert: A copyright exception for web archiving on the European level would strengthen the position of GLAM institutions that have a web archiving programme. An exception exists on the national level in a number of European countries, but many more countries would benefit from a similar exception if it became implemented on the EU level.

Lise-Anne Denis: Web archiving activities in general should be given more importance from a legal perspective. In order to lift the obstacles that law currently imposes on these practices, there is a need to improve the copyright exceptions at the European level, both for the reproduction and communication of web material, that could be applied with certainty to web archiving practices (including the archiving itself and web archiving studies). Those exceptions should also be made mandatory across all Member States by a maximum harmonisation directive. The scope of the exception provided by the GDPR for scientific and research activities should be extended for private institutions to be able to benefit from it. Its implementation should also be harmonized throughout the European Union.

Yves Maurer: Access to web archives needs to be made more open. The biggest impediment at the moment is the legal environment. The Text and Data mining exceptions from the EU Directive 2019/790 have to be strengthened at a European level so that individual web archiving institutions in different member countries can provide researcher access remotely to the collections. In the meantime, web archiving institutions should make as much metadata and primary data available as possible. Practical solutions can be that metadata, not covered by copyright or GDPR, is made available publicly and access to restricted data is made available through a cooperation agreement between the researcher and the institution. The procedure for such access needs to be well-documented and publicised so that the barrier to entry for new researchers is not insurmountable.

Ditte Laursen: The complexity of transnational web archive studies necessitates pooling resources and sharing knowledge among a wide range of stakeholders, including academic institutions, libraries, governmental bodies, and

Conversation 1: Transnational

technology companies. Such collaborations can lead to the development of shared tools, methodologies, and best practices that could enhance the efficiency and effectiveness of research in this field. Collaborative efforts could also address the technical challenges associated with web archiving, such as the development of advanced crawling technologies, data storage solutions, and infrastructure/analytical tools tailored to the needs of transnational research. Sharing knowledge and expertise in these areas can accelerate innovation and improve the quality of research outputs.

In advancing transnational web archive studies, leveraging data-driven artificial intelligence (AI) can be pivotal. AI can analyze vast archives to identify patterns, trends, and relationships that would be unfeasible for humans to discern unaided. This approach can be used to build models that predict the cultural and societal impact of web content across borders, or to understand the global diffusion of information. It is essential to recognize that human oversight remains crucial. Researchers in the field must curate datasets, interpret results, and provide context for AI findings. Transnational web archive studies should strive to balance the capabilities of AI with the nuanced understanding and ethical considerations that human researchers bring to the field.

**What are the main challenges in building transnational web archives?
(Think, for example, on the challenge of context)**

*Helena Byrne,
Juan-José
Boté-Vericad,
Sharon Healy:*

There are a number of challenges to building transnational web archives. Interoperability and scope are just two types of challenges. Different metadata standards for transnational collections that are to be assembled from already existing collections across different web archives present interoperability challenges. Another big issue is the scope. Many institutions are limited to selecting content from their national domain. The institutions that can allow external access to their collections try to make their collections more discoverable via the Mementos Protocol. This can help users who want to explore web archiving research questions at a transnational level. The International Internet Preservation Consortium (IIPC) membership also recognized the need for a transnational approach to preserving transnational events and set up the Content Development Group (CDG). However, despite the CDG's efforts to include participation from organisations and individuals from outside the IIPC community the collections are primarily focused on countries with IIPC membership. One reason for this is because of the digital divide between the global north and south. Many countries in the global south publish information on social media platforms like Facebook. Proprietary platforms like these can be easier for people to publish on but are very challenging to archive. Thus this causes a divide in curating a truly transnational web archive collection.

Friedel Geeraert:

Having insight into the socio-political and cultural context of the geographic regions that you wish to collect content from is primordial. It is easier to scope in more crucial resources when creating seed lists when

the people involved in content curation are familiar with the context of the regions they collect content from. Transnational collaboration or collaboration with experts from different geographical regions is necessary to create transnational corpora.

Lise-Anne Denis: Being able to understand and respect the law of all the states involved is important.

Which infrastructures and tools are needed to promote the comparison of web domains, and how to have them established (in case they are not there already)?

Susan Aasman: What might be helpful is to have an infrastructure such as DARIAH (The Digital Research Infrastructure for the Arts and Humanities), or CLARIN (digital infrastructure offering data, tools, and services to support research based on language resources). WARCnet has been instrumental in developing a community of scholars, archivists, and IT experts, but a sustainable infrastructure that develops tools and hosts datasets should be the next step.

Karin de Wild: In response to Susan Aasman's comment, I agree that for transnational studies, it is helpful to sustain a European infrastructure for scholars studying web archives. It would be beneficial to support a consortium of partner institutions from various member states and coordinate efforts at both local, national, and European levels. It may be possible to sustain existing initiatives, like WARCnet and RESAW, or to integrate these initiatives as part of DARIAH or to maintain them in collaboration with the International Internet Preservation Consortium (IIPC).

Janne Nielsen: If the web domains that are to be compared are preserved in different web archives, we face several challenges (some are described in Nielsen et al., 2024), for one thing, that the archives have different strategies and scopes and that the archiving might have been done in different ways. Therefore, documentation is very important to understand what is in an archive, and how the specific strategies and harvesting settings influence the collection. I think this is already a focal point for the archives but research networks like WARCnet and RESAW can help facilitate discussions on what types of documentation are needed, how we can use the documentation to better understand the consequences of different strategies in archiving, and how this may help us compare. The tool for large-scale comparisons of the holdings of different archives (presented in Nielsen et al., 2024) is meant to help facilitate comparisons of web domains, but for more detailed analysis we also need tools for creating corpora, which allows us to define and delimit relevant corpora in the archives included, for instance in line with the work done in Brügger, Nielsen, and Laursen (2020). This type of corpus selection, again, requires a profound knowledge of the archive and the processes of archiving, and necessitates collaborations between archivists, researchers, and IT-specialists. I agree with Susan Aasman that the next step should be a sustainable infrastructure supporting such

Conversation 1: Transnational

collaborations, especially because the complexity of the work means that results are greatly enhanced by having a community where we can discuss best practices and collaborate on infrastructures and tools, similar to the work done in RESAW.

Helena Byrne: Many web archives have tight restrictions on how their collections can be accessed and used. However, the metadata produced by these institutions could potentially be made accessible to researchers. The work that Emily Maemura and I have been doing on the Datasheets for Web Archives Tool Kit will hopefully make it easier for web archives to publish their metadata. This would make it possible for researchers to do a comparison of national web domains by using metadata from multiple web archives.

Yves Maurer: To expand on Janne Nielsen and Helena Byrne's comment, I would like to include the perspective of a web archiving institution. For web archiving institutions it is important to know what type of information is of interest to researchers and how it should be organised. As a minimum, a web archive only needs to crawl websites and store the resulting files. However the research community is interested in a lot of other information as well, like a collection policy, a history of the web archiving effort itself, seed lists, crawl logs, scoping information, dissemination efforts etc. This is not necessarily information that is standardised or even kept for posterity. Researchers need to communicate these wishes to the web archiving institutions so that the objects of study exist later in the process. This is something that is starting to be addressed by Datasheets for Web Archives but needs to go further and include readily available information from institutions that perform web archiving and focus on the data points that are currently possible to include instead of abstract wishes that are unfeasible to produce.

Saskia Huc-Hepher: An EU infrastructure would definitely be a step in the right direction, but it would — yet again — risk hardening the border between the privileged “in-group” and those in states or regions on the outside. In the long term, we should seek to be more ambitious and strive for a web archiving infrastructure that reflects the borderless aspirations of the web itself and is inclusive of domains (and languages) across the world.

References

- Balbi, G. (2024). ‘History Web’, ‘Web History’, and ‘History of the Web’: Three Subfields and Why (and Why Not) Integrating Them. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *Companion to Transnational Web Archive Studies*. Routledge.
- Balbi, G., & Magaouda, P. (2018). *A History of Digital Media: An Intermedia and Global Perspective*. Routledge.
- Bloch, M. (1949). *Apologie pour l'histoire ou Métier d'historien*. Broché.
- Braudel, F. (1958) *Histoire et Sciences sociales: La longue durée*. *Annales. Économies, Sociétés, Civilisations*, 13(4), 725–753.
- Brügger, N. (2012). When the Present Web Is Later the Past: Web Historiography, Digital History, and Internet Studies. *Historical Social Research/Historische Sozialforschung*, 37(4 (142)), 102–117.

The Routledge Companion to Transnational Web Archive Studies

- Brügger, N., Nielsen, J., & Laursen, D. (2020). Big Data Experiments with the Archived Web: Methodological Reflections on Studying the Development of a Nation's Web. *First Monday*, 25(3), n.p. <https://doi.org/10.5210/fm.v25i3.10384>.
- Nielsen, J., Maurer, Y., & Zierau, E. (2024). Comparing the Holdings of Closed National Web Archives through Summaries. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *Companion to Transnational Web Archive Studies*. Routledge.
- Sandvig, C. (2013). The Internet as Infrastructure. In W. H. Dutton (Ed.), *The Oxford Handbook of Internet Studies* (pp. 86–106). Oxford Academic.
- Winters, J. (2018). Digital History. In M. Tamm & P. Burke (Eds.), *Debating New Approaches to History* (pp. 277–300). Bloomsbury Publishing.

PART II

The COVID-19 crisis as a transnational event

8

ORAL HISTORIES AND SCALABLE READING

Analysing born-digital collecting practices during the COVID-19 pandemic

*Friedel Geeraert, Nicola Bingham, Helle Strandgaard Jensen,
Jane Winters, and Frédéric Clavert*

Introduction

This chapter reflects on the unprecedented collecting activity that occurred globally in response to the COVID-19 pandemic that began in late 2019, and specifically on the efforts of cultural heritage institutions to curate web archive special collections that will form a lasting record of how the impact of the pandemic manifested in online spaces. The chapter draws on a major oral history project undertaken by members of WARCnet, which sought to understand how collecting institutions — and in one case an international network — were able to function during a period of extreme disruption; how they developed their special collections; the resources on which they were able to draw (human, financial and technical); how they addressed questions of inclusivity and representation in their activities; and the challenges of archiving a global event at the national or regional level.

The chapter touches on several important themes in the field of web archive studies, including the importance of understanding processes of collection and archiving; the need for extensive documentation of those processes; the value of special collections; the importance of oral histories for understanding histories of the web and web archives; and the need to identify and acknowledge those who co-create and curate web archives. Among those engaging with questions of documentation and positionality are Schafer et al. (2016), who highlight “the multiple socio-technical mediations, arrangements, and agencies mobilized throughout the archiving process — be they technical or human — or the choices that decide their selection, presentation, and accessibility,” noting that “Web archives can be better used as sources once they have been understood as an object of study.” Maemura et al. (2018, p. 1223) stress that “To engage scholarly uses of web archives and support researchers, we must first study web archive practices to understand and communicate the impacts on resulting collections,” and they propose “a framework for documenting key dimensions of a collection.” Similarly, Bell et al. (2022, p. 66) underscore the need to ensure that processes

“are documented openly, allowing for collaboration between web archiving institutions and a common understanding among researchers of their potential use.”

Oral histories are a key mechanism for opening up web archive “black boxes” (Schafer et al. 2016) and timeliness is essential if the nuances and particularities are to be effectively captured. Winters (2019, p. 82) notes that

The creators and users of websites, both new and old, are still available to be interviewed; as are the librarians, archivists, and technologists whose knowledge, expertise and, ultimately, decisions have shaped and will continue to shape how we will be able to access the historical web now and in decades to come.

The experiences of the former — creators and users — are increasingly being captured through oral histories and ground-breaking ethnographic research (see, for example, Nolan 2017, Locatelli 2017, and Mackinnon 2022), but the WARCnet oral history project focused the spotlight on the third group — those involved in web archiving within memory institutions.

Finally, the importance of special and/or thematic collections has been evident since web archiving began in the late 1990s. In a survey of web archiving initiatives conducted in 2003, Day identified the importance of selective web archiving and noted the existence of several special collections, for example around the Sydney Olympics of 2000 or those generated collaboratively by the Internet Archive, the Library of Congress and the Smithsonian Institution (Day, 2003). In the same year, Schneider et al. outlined a method for creating a thematic web collection, which was, “an archive of Web objects identified and captured using a set of URLs believed to be relevant to a specific theme or topic” (2003, p. 77). They were notably interested in the development of transparent collection policies for “thematic Web archives involving unfolding events,” in their case the September 11 Archive (Schneider et al. 2003, p. 87). This is early evidence of a focus on the archiving of crises or moments of national and/or global trauma as they play out online (see, for example, Schafer et al. 2019), which is nowhere more evident than in the COVID-19 special collections discussed here. This chapter seeks to demonstrate the value of special collections, both in response to crisis and to more routine events; to highlight the importance of bringing researchers and archivists together to discuss motivations and processes; and to undertake some preliminary analysis that demonstrates the richness of the oral history corpus.

The dataset analysed in this chapter comprises 12 in-depth interviews that were conducted with web archiving practitioners between July 2020 and August 2022. The purpose of the interviews was to gain insight into web archiving practices and processes. Participants were based in or representative of Denmark, France (two institutions), Hungary, Iceland, Luxembourg, Switzerland, the United Kingdom, the Netherlands, the United States, Australia, and the International Internet Preservation Consortium (IIPC). The IIPC collection is a global one, but we acknowledge that the dataset overwhelmingly presents the experience of the Global North. There are numerous factors at play here — WARCnet is a European project, web archiving is dominated by Global North institutions — but there are many other stories about the pandemic to be told. A larger oral history project might begin to address some of these imbalances, but it is beyond the scope of the current work to provide an exhaustive summary of the global landscape of COVID-19 collecting. It is hoped

Oral histories and scalable reading

that this chapter will serve as a starting point for researchers and practitioners interested in exploring pandemic web archiving.

The interviews were analysed using different, but complementary, methods. The first part of the chapter provides a high-level, comparative overview of existing COVID-19 web archive collections, comparing how they were organised and operationalised when they were initiated and the kinds of information they contain. The comparison was made through a close reading of all 12 interviews. The second part of the analysis uses a combination of close and distant reading, drawing on the Voyant suite of tools and Iramuteq for the latter, and demonstrates how the two approaches can be combined into a scalable reading method. The application of different tools and methods to the same corpus illustrates the types of analysis that are possible and also highlights the influence of technology choices. The final section of the chapter returns to close reading to highlight the challenges, opportunities, and innovative methods and approaches described in the interviews, discussing both the differences and commonalities across participants. The different methods are discussed in more detail below.

Data

The 12 interviews that are the focus of this research were semi-structured, meaning that the same set of questions was posed to each interviewee, but the interviewer was free to ask follow-up questions depending on the information provided. The questions were organized into thematic sections:¹

- The reasons for the special collection
- The organisation of the collecting process
- The scope of the COVID-19 collection
- The frame of the special collection
- Accessibility and searchability
- Partnerships and uses
- Best practices for collaborative collections

All interviews were conducted in English, except for that with the National Library of France, which was subsequently translated into English. The interview at the National Library of Luxembourg was conducted in person while the interviews with the Library of Congress, the Swiss National Library, the National Széchényi Library in Hungary, the Institut National de l'Audiovisuel (France), the National Library of France, the IIPC Collaborative collection and the UK Web Archive were conducted online and transcribed using a transcription software like Descript. The interviews with Netarkivet (Denmark), the Icelandic web archive, and the National Library of Australia were conducted in writing, the first two via a shared document (in the cloud or via screen sharing) and the last via email, with follow-up questions asked in the same manner for each of the institutions respectively. The interviewees were asked to approve the text before publication. In the second step, a spreadsheet was created in which each row was dedicated to a specific question that each interviewee was asked. The resulting grid facilitated the close reading of the transcriptions to compare the different practices between institutions, as pertinent information related to one question was sometimes mentioned in the answer to a different question.

High-level comparative overview

Given the enormous impact of the COVID-19 crisis, creating dedicated web archive collections was seen as essential by many cultural heritage institutions. From early 2020 onwards, web archiving professionals faced the challenge of how to capture the manifestations of the COVID-19 pandemic on the web in all their diversity so that these testimonies could be preserved and made available for future research. During the interviews conducted for the WARCnet oral history project, information was obtained about how operational decisions were made about scoping the collections, especially considering the global nature of the pandemic; the length of the collecting period; the biggest challenges faced; and access arrangements for the content. Each of these topics will be discussed in more detail in what follows.

Scoping the collections

When analysing the content included by each different initiative in its collections, it is worth noting that, except for the IIPC collaborative collection and that of the Library of Congress, all institutions focus solely on the national level. In many countries, legal deposit legislation provides a clear framework for how to scope collections, but this is not the only determining factor. Limitations related to technical infrastructure and know-how also need to be taken into account when considering the content that is included. This results in a wide variety of approaches.

When it comes to archiving a global event at the national level, the institutions in our sample use combinations of different criteria to delimit the national sphere. Some focus on information published in the national domain, such as .nl for The Netherlands, while others focus on information in the national language(s) of the country (although the fact that many countries share languages complicates this criterion). Still, others focus on information about the country, information published by national citizens or organisations, or information published in that country.

National web spheres encompass a wide variety of content types. First of all, there are different media to consider: websites, social media, or even apps. Moreover, within these different media, a huge range of topics exist. When comparing the media collected by each initiative, Table 8.1 shows that some did not collect social media and focused solely on websites. This is the case at the national libraries of Switzerland, Iceland, Hungary, and The Netherlands, the Library of Congress, and the IIPC collaborative collection.

The most popular social media platforms from which to collect data are Twitter and YouTube, followed by Facebook and Instagram. The Royal Danish Library offers a particularly wide selection of social media content, including Twitter, Facebook, Instagram, YouTube, TikTok, Reddit, Twitch and even some podcasts. Social media data can be collected through scraping or crawling, but it may also be harvested via a platform's Application Programming Interface (API). There is a wide literature addressing the challenges of working with APIs (see, for example, Pehlivan et al. 2021, Acker & Kreisberg 2020 and Littman et al. 2018), some of which are technical but, as illustrated most recently by the changes in access to the Twitter API, may also be commercial, legal and cultural. Even more than the live web, social media is a moving target for archiving institutions, and, as will be discussed later in this article, platform affordances significantly shape the nature of both special and broad collections.

Oral histories and scalable reading

Table 8.1 Organisation-type overview of types of web content collected by the initiatives

<i>Institution</i>	<i>Websites</i>	<i>Twitter</i>	<i>Facebook</i>	<i>Instagram</i>	<i>YouTube</i>	<i>TikTok</i>	<i>Podcasts</i>	<i>Reddit</i>	<i>Twitch</i>
Swiss National Library	Yes	No	No	No	No	No	No	No	No
National Library of Iceland	Yes	No	No	No	No	No	No	No	No
National Library of Luxembourg	Yes	Yes	Yes	No	No	No	No	No	No
National Library of France	Yes	Yes	No	No	Yes	No	No	No	No
National Library of Hungary	Yes	No	No	No	No	No	No	No	No
Royal Danish Library	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
IIPC Collaborative Collection	Yes	No	No	No	No	No	No	No	No
UK Web Archive	Yes	Yes	No	No	No	No	No	No	No
Institut National de l'Audiovisuel (France)	Yes	Yes	No	No	Yes	No	No	No	No
Royal Library of The Netherlands	Yes	No	No	No	No	No	No	No	No
Library of Congress	Yes	No	No	No	No	No	No	No	No
National Library of Australia	Yes	Yes	No	No	Yes	No	No	No	No

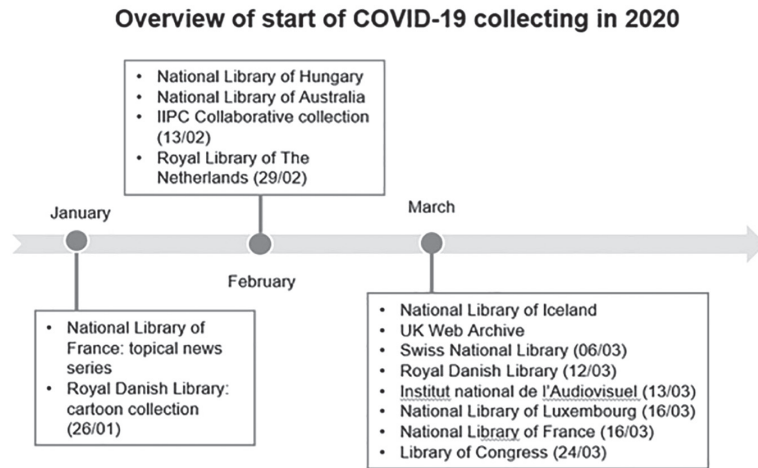


Figure 8.1 Timeline of the start of COVID-19 collecting (colour version in Zenodo repository).

Regarding the topics included in the collections, the elements that were most frequently mentioned during the interviews are official communications from the government and other public bodies, as well as news, media, and communications from key organisations and initiatives. But these collections contain material concerned with a wide range of topics: law, psychology, support services, charity, expat life and embassies, insurance, agriculture, medical research, sports, working from home, national protection measures, Indigenous and historically underrepresented groups, healthcare, politics, lockdown, controversial topics, trade unions, culture, education, professional organisations, domestic violence, food waste, business and the economy, everyday life, home-schooling, corona couture, weddings, volunteering, hospitality, protests, conservation, travel and tourism, religious communities, agriculture, and so on.

Another scoping dimension is temporal: when did the initiatives start or stop collecting information about a global pandemic? Figure 8.1 shows when the different initiatives started their COVID-19 collections. The first content was harvested in January 2020 at the National Library of France and at the Royal Danish Library. In both cases, it concerned a small preliminary collection. In Denmark, a collection around a COVID-19-related cartoon was created and at the National Library of France, a topical news series was curated. The first COVID-19 collection per se was the IIPC collaborative collection that started in February 2020. Hungary, The Netherlands, and Australia followed suit in the same month and the other initiatives started in March. In some cases, the lockdown that was imposed was the trigger. At the time the interviews were undertaken, all of the initiatives were still collecting content related to the pandemic. Other aspects of the collecting processes, for example, the available resources (human, financial, and technical), stakeholders, and consultation will be addressed in more detail later in this chapter.

Access

Access to the COVID-19 collections differs between institutions. Some of the COVID-19 collections are freely accessible online, which is the case for the IIPC collaborative collection

Oral histories and scalable reading

(via Archive-IT),² the Icelandic web archive³ and the COVID-19 collection of the National Library of Australia (via Trove).⁴ Other institutions have to work by requesting permission from the rights-holders to share the archived content freely online. Some collections, such as those of the Library of Congress⁵ and the UK Web Archive,⁶ are therefore partly available online and partly available only in the reading rooms of the institutions.

Other COVID-19 collections are solely accessible on site. This is the case at the National Library of France, the Institut National de l'Audiovisuel (France), the National Library of The Netherlands, the National Library of Luxembourg, and the Swiss National Library. In France, the collections are also available in the reading rooms of regional libraries so that interested parties do not need to travel to Paris to consult them. Netarkivet at the Royal Danish Library is an exception in the sense that remote access to the collections and data dumps are possible, but one has to be a Danish researcher or be associated with a Danish institute to take advantage of these options.

Given the limited access to some collections, which is perceived as a real barrier to promoting the research potential of web archives, some initiatives also make available the seed lists (lists of URLs that are fed to the crawler software) and/or metadata of their collections. The seed list of the COVID-19 collection at the National Library of Hungary⁷ is made publicly available, for example, and at the UK Web Archive, all the metadata of the COVID-19 collection is freely accessible online so that users can quickly determine if the resource they are looking for is readily available or not.

Following this broad overview of the collections, derived from the corpus of oral history interviews, the chapter now turns to exploring different modes of analysis — or analytical “reading” — of the interview dataset.

A scalable reading of web archive oral histories

Born-digital archives and scalable reading

One way the interviews can be explored is by using a scalable reading method. By scalable reading, we mean an analysis that combines what Franco Moretti has called close and distant reading (Moretti 2000). The advantage of scalable reading is that it combines methods with which humanities scholars are familiar, that is, close reading where every data point is read and analysed by the researcher, and methods with which they are usually less familiar, that is, distant reading, where a computer is used to read many documents and aid the analysis (Pedersen et al. 2022; Jensen et al. 2023).

The scalable combination of the two unites close reading of a single data point with computer-assisted distant reading in one analytical narrative. This could, for instance, be an analytical strategy where the distant reading provides the close reading with a wider context than would have been possible by reading only one case. It could also be used to select case studies for close reading in a systematic and reproducible manner based on an initial distant reading. Combining the two methods, scalable reading has the potential to overcome the gap between “the digital” and “the humanities” that has been lamented by researchers in the field, as it forces results generated by the two different sets of methods to remain in conversation with each other (Booth 2017; Hammond 2017).

What we get from applying a scalable reading to the interviews, of course, depends on the questions we want to ask. One researcher might want to use it to understand how web archiving worked in these extraordinary circumstances, for example, what were the general

challenges of building these collections; another might want to use it as a way to understand how a particular issue that they want to explore, for example, social media usage, is affected by the acquisition and curation of data.

Reading the oral history corpus with Voyant

The in-depth interviews with web archiving practitioners about curating COVID-19 collections are not a large corpus and could in principle be analysed using close reading only. However, using the graphic user interface (GUI) text mining program Voyant (Sinclair & Rockwell 2016), the researcher can quickly survey all of the sources and gain ideas for possible analytical entry points, as it provides an overview of common themes across the interviews.

The 12 interviews, downloaded from WARCnet's webpage on 5 October 2022, were imported into Voyant, to provide an overview of their similarities and differences at a glance. We see that the shortest is 1,963 words and the longest is 9,213 words. This means that we must expect a difference when, for instance, it comes to the level of detail in which the collection process is described.

Searching for the two words “challeng*” and “problem*”, and then visualising the frequency of these words across the 12 documents, we can see that all of the documents talk about the difficulties of making collections of this kind. A researcher can use these interviews to inform their research either into the difficulties of making such collections generally or into whether the particular challenges they are finding are unique.

To further understand the nature of the challenges and problems, the researcher can choose to see the context of up to 50 words to the right and left of each term. In the case of challeng* there are 45 occurrences across the 12 documents; in the case of problem* there are 29 occurrences. Exporting these to Excel enhances readability (and the potential for further processing) and the researcher can now make a close reading of all the data points to see how challeng* and problem* have been perceived in relation to the different collections. Seeing the data points in context allows for a detailed comparison of how they were talked about in the interviews and, indeed, of what the challenges and problems might be.

Another example of scalable reading is exploring how the interviewees talk about social media content in the collections. Searching for “social media” “Twitter” “Facebook” and “Instagr*”, the visualisation of the relative frequency of these terms across the 12 documents shows that it is a persistent theme. Both the term “social media” and the specific names of three of the biggest social media platforms are mentioned in the interviews.

Again, we can choose to see the context in which these terms occurred. Displaying the 50 words on either side of the term will enable a researcher to make a close reading of the different instances. On this basis, one can both do a comparison and see in which interviews something was the most prominent theme, but also know if this was a unique or unusual issue. This is the case for instagr*, for example, which only comes up in 5 of the 12 interviews of the cases and is only mentioned 9 times in total.

This exploratory investigation of a small corpus reveals the promise of scalable reading, not just for the content of web archive collections but for their documentation. The oral history interviews that are the focus of this analysis provide vital context for the collections themselves — they are de facto important documentation for a period of web archiving that was characterised by innovation and rapid response, adaptation, and adjustment of existing frameworks and processes. Voyant is an accessible tool that can be used in a browser and, as

Oral histories and scalable reading

such, it allows for an immediate form of scalable reading. But it necessarily has limitations and particular affordances, like all tools and software. More specialist tools, and a different method, can bring other kinds of insight and discovery, as will be described in the next section of this chapter.

A complementary approach: Zooming out with Iramuteq

Choosing to use a different tool, in this case Iramuteq (iramuteq.org), reveals different patterns and emphases in the data and suggests new avenues for close reading. Based on the theory of lexical worlds (“théorie des mondes lexicaux”; see Reinert 1983 and Reinert 1993), the texts and their metadata are imported into Iramuteq, which divides them into segments of text that correspond roughly to sentences. The software then calculates the proximity of each segment based on word co-occurrences. The result is clusters (or topics or profiles) of text segments from which the most relevant words are extracted for human interpretation. The relevance of each cluster for each interview can then be visualised. Lastly, Iramuteq’s interface allows users to return to each specific text segment, enabling distant reading to be combined with close reading, hence resulting in scalable reading (Clavert & Fickers 2021).

A distant reading of a subset of the corpus of 12 interviews was conducted using the Iramuteq software. Ahnert et al. (2023) have urged digital humanities projects “to think about how to interrogate the contours of their data” (p. 34), and it is here that a tool such as Iramuteq comes into its own. Some cleaning of the data is required before the analysis, based on clusters of segments of text, with each segment more or less equivalent to a sentence (around 40 words), can begin, but this is true for most kinds of reading at scale. The views of the data that result can help to identify the most representative words for each cluster grouped into topics. The method is a statistical one rather than involving artificial intelligence and is consequently both explainable and reproducible.

Figure 8.2 illustrates some of the themes that emerged from our dataset after this initial modeling of topics. For example, there is a discernible emphasis on collaboration and the human aspect of collecting in the third cluster from the left (headed “IIPC”). Here there are words such as “collaborative,” “stakeholder,” “department,” “colleague,” “contribute,” and “nominate.” The fifth cluster from the left (headed “crawl”), by contrast, seems much more concerned with the management and scheduling of the web crawling process, with words suggestive of time and process such as “stop,” “weekly,” “ongoing,” “update,” “end,” and “regularly.” These clusters are not necessarily revealing anything unexpected, but they are clear proof of a concept based on a relatively small dataset, and give us confidence that the method can be scaled up, for example to the collections themselves, where they are openly accessible.

Figure 8.3 illustrates how the different clusters do or do not connect to each other. What you might call the “person/team” cluster and the “collaborative, membership” cluster, where the IIPC features prominently, are quite closely related, while what you might call the more “technical” clusters are near to each other but relatively distanced from the more human-centered activity. This tells us something about both the emphases of the interviews and aspects of the collecting processes, but also potentially about institutional cultures and operations. This is a starting point for the kind of scalable reading described above.

In Figure 8.4 the “story” of each web archive and its COVID-19 special collection is represented. The first thing to notice is the huge degree of variation, both in the length of

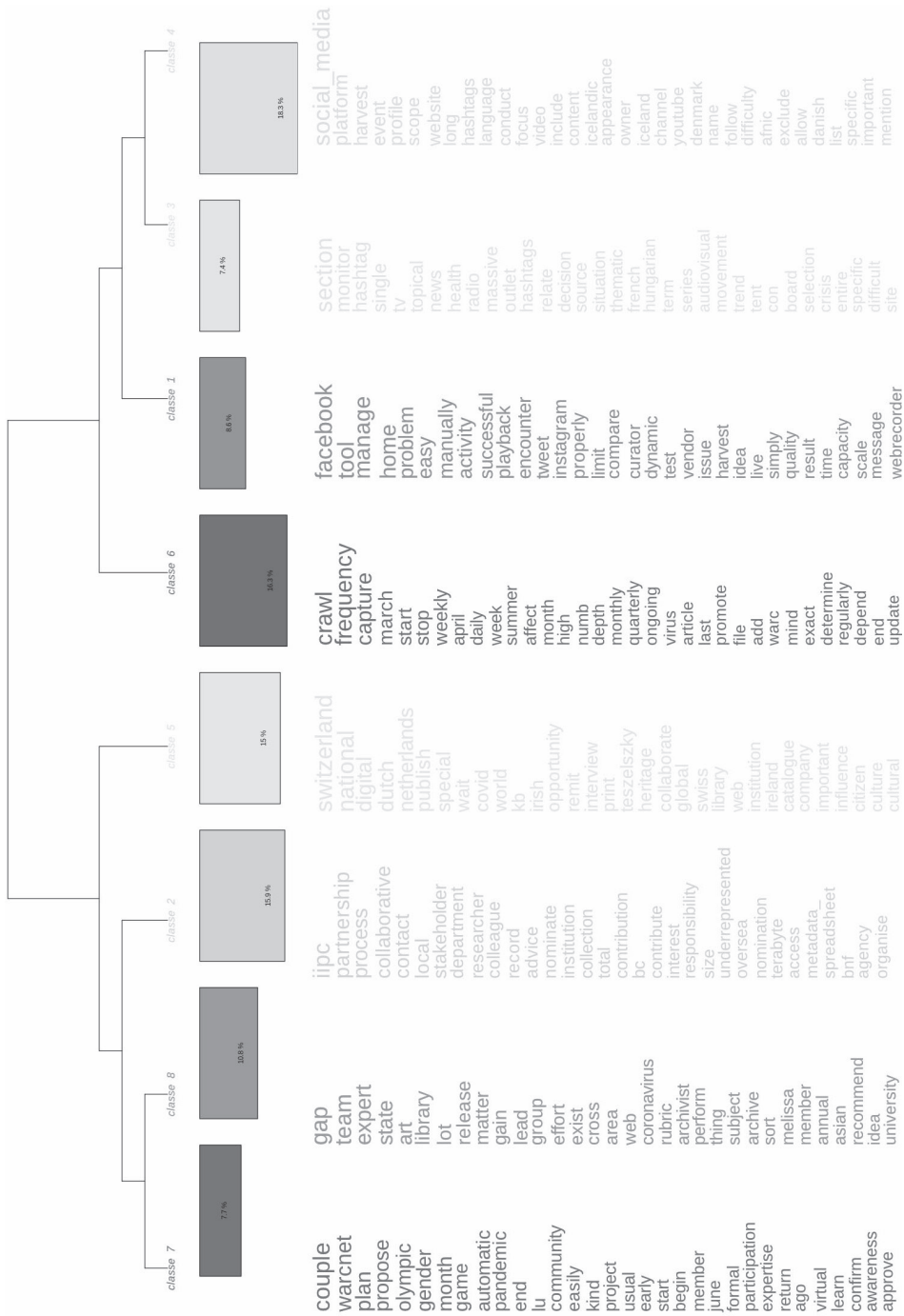


Figure 8.2 Themes in the oral history corpus, generated using Iramuteq (colour version in Zenodo repository).



the interview and in the degree of interest in particular topics of exploration (the thematic topics shown in Figure 8.2 above). The interview with Netarkivet, the Danish Web Archive (column 3), is the longest of the eight by some margin, and most topics receive equivalent attention. In the interviews about the Dutch, UK, and IIPC collections (columns 10, 12, and 6), single topics tend to dominate. The interview with the Icelandic Web Archive (column 5) is by far the shortest of the eight, and consequently, it is likely to be less “present” on the topics identified above.

131

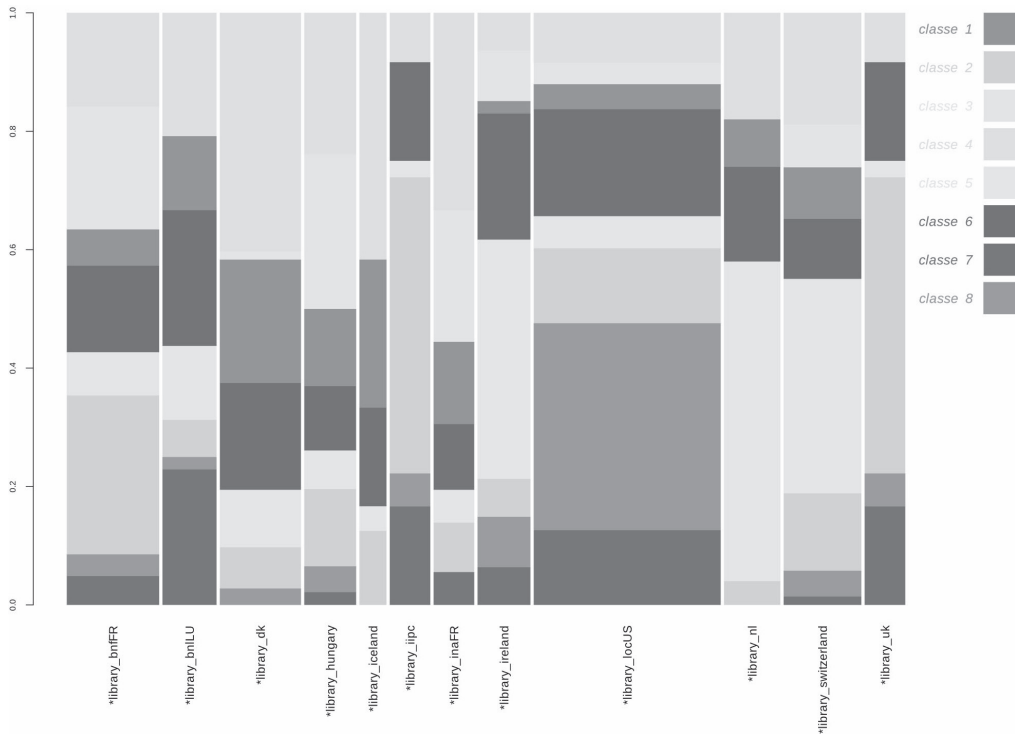


Figure 8.4 Visualisation of the structure and emphasis of the oral history interviews, generated using Iramuteq (colour version in Zenodo repository).

Challenges, innovation, and opportunity: A close reading

Challenges

Several challenges were highlighted in the interviews. Some of them are related to web archiving activities in general, such as technical challenges when capturing particular content types, limitations on financial and human resources, difficulty in defining the national sphere on social media, or obtaining permissions to capture and make available the content. However, a few challenges were specific to the COVID-19 collections. Most initiatives have a tradition of collecting about certain events, but these are all limited in time. Given the duration of the pandemic, managing the COVID-19 collections required a sustained effort over several years. Delimiting the collections was also seen as difficult because the pandemic touched upon every aspect of life and left traces in many different parts of the web. Deciding which parts were important or significant enough to include in the collections was therefore a problem for many institutions. Furthermore, creating collections that are representative of a nation as a whole and deciding how to handle controversial content such as COVID-19 denial websites also presented web archiving professionals with specific challenges. In what follows, more detail will be provided about the different approaches the institutions in our sample used to create representative collections and to handle controversial content.



133

01-Oct-24 22:34:10

The Routledge Companion to Transnational Web Archive Studies

Table 8.2 Overview of the ways in which selection of content was opened up

<i>Institution</i>	<i>Web archiving team</i>	<i>Internal network of curators throughout institution</i>	<i>External network of curators</i>	<i>Recommendations from the public</i>
Swiss National Library	Yes	No	Yes	No
National Library of Iceland	Yes	Yes	No	No
National Library of Luxembourg	Yes	No	Yes	Yes
National Library of France	Yes	Yes	No	No
National Library of Hungary	Yes	No	No	No
Royal Danish Library	Yes	No	Yes	No
IIPC Collaborative Collection	Yes	Yes	No	Yes
UK Web Archive	Yes	Yes	Yes	Yes
Institut National de l'Audiovisuel (France)	Yes	No	No	No
Royal Library of The Netherlands	Yes	No	Yes	No
Library of Congress	Yes	Yes	No	No
National Library of Australia	Yes	No	Yes	No

At the Library of Congress, a lot of consultation was done with experts across the Library who were not part of the COVID-19 project team. The team also undertook an analysis a year into collecting to identify gaps in their collection and, as a result, organised sprints of one week or a month to fill these gaps by focusing on specific groups:

Because there were some subject matter experts that were not on this team, we did a lot of consultations with experts across the Library. [...] We did general analysis after a year or so of collecting and were able to identify other gaps. Then we'd say: "This month, or this week, this is the gap area we're going to focus on, or let's focus on these groups".
(Holownia & Geeraert 2022, p. 11)⁸

Another way of diversifying input was to partner with an external network of curators, for example, partner libraries in Switzerland, Denmark, and Australia, specific government departments in Luxembourg, or networks of specialised institutions such as the Royal Colleges of Nursing in the UK. In The Netherlands the National Library, for example, requested help from other heritage institutions to create a national collection: "We also collaborated with the Digital Heritage Network in The Netherlands, together we initiated a call for heritage institutions to help build a national collection about the coronavirus and its effects in The Netherlands" (de Wild & Kyritsis 2021, p. 10).⁹

In the case of Luxembourg, the UK, and the IIPC collaborative collection, recommendations from the public were also invited to make the collections more inclusive. At the National Library of Luxembourg, web content about religious communities was added after the call for participation was launched:

We [...] launched a call for participation at the end of March and there too we got a very good response from the media and a few communities that we would not have otherwise thought of suggested their websites [...]. For example, the Muslim community and shoura. lu.¹⁰ I hadn't thought of looking for religious communities. The

Oral histories and scalable reading

Muslim community posted online information and recommendations for its members about services in mosques, religious holidays, etc. Based on this suggestion, we then looked more closely at other religious communities.

(Schafer 2020a, p. 5)¹¹

Another challenge identified by many institutions was how to handle fake news and other controversial topics. Different approaches exist. In some collections, this type of content was excluded, although a record is kept for the URL of the website. For the IIPC collaborative collection, for example, the decision was made to exclude an extreme right-wing website promoting anti-vaccination as it could constitute a danger to health. A record was kept of the website so that a trace of the nomination of the website still exists:

It was quite an extreme right-wing website that had an anti-vaccination policy [and said] that the coronavirus was a made up pandemic. It looked like it was quite factual and verified information. [...] So the decision was taken to not add this website to the collection, because there is a risk that somebody might look at that article [...] [and that] it could potentially cause danger to health. [...] We kept a record that the website had been nominated. We recorded what the content of the website was and why we decided to not include it in the collection.

(Geeraert 2020a, p. 10)¹²

Some institutions, on the other hand, made conscious efforts to include this type of content. At the National Library of France, for example, specific steps were taken to ensure the inclusion of highly controversial topics in the collection: “I was particularly concerned about representativeness, choosing sites that reflected all viewpoints. That was all the more important because there were highly controversial topics like chloroquine [...]” (Gebeil & Schafer 2020, p. 7).¹³

In other institutions, specific measures were adopted to contextualise the fake news. The Library of Congress wanted to include the content in the web archive but did not want to give these websites promoting misinformation authority by sending them a permission letter. Instead, they archived an aggregator of websites publishing misinformation instead, thereby, contextualising the content:

We’re not collecting them because of their authority; we are collecting them because they were an example of misinformation. The solution was to go to one of these big aggregators that listed them all — NewsGuard¹⁴ — so that is in the archive and so there is content that represents the misinformation.

(Holownia & Geeraert 2022, p. 17)¹⁵

Innovation and opportunity

There was plenty of evidence for innovation and opportunity demonstrated throughout the interviews. One of the foremost was the opportunity for the archiving institution to contribute to the public good in collecting material about COVID-19, especially when it was able to make its collection publicly available. Some even saw the collection as an opportunity to help reinforce the value of the national library. At the British Library, for example, it was pointed out that

we are creating a research collection that can support efforts to understand the social and economic impact of the virus. There is certainly a role for the Library to play in that researchers of the future might want to mine this collection in managing epidemics in the future.

(Geeraert 2020b, p. 5)¹⁶

The immediate societal value of web archives was evident in the thinking of the National Library of France, with Digital Legal Deposit perhaps offering a mechanism, early in the pandemic, for “providing tasks for people working from home who were unable to perform their usual job remotely, for example, warehouse workers” (Gebeil & Schafer 2020, p. 5).¹⁷ The importance of public access to and usage of the COVID-19 collections also inspired innovations in the structuring and documentation of data. At the Institut National de l’Audiovisuel, for example, work was undertaken to demystify the hashtags used for collecting social media content: “Before, you had to know the hashtags you wanted to search for, but now the list is accessible and we have documented the hashtags, there is a detailed description and descriptors from the INA thesaurus” (Schafer 2020b, p. 12).¹⁸ This speaks strongly to a focus on future use when hashtags that might seem obvious to contemporaries lose their familiarity as time passes.

In the case of the British Library, collecting around the pandemic helped to focus efforts on subjects that were not traditionally well represented in the web archive, such as Science, Technology, Engineering, and Medicine (STEM), which, in comparison to arts and humanities, are not as strongly entrenched in the traditional collecting expertise of the institution. The British Library web archiving team worked with the Royal College of Nursing and was, in turn, able to coordinate with the other Royal Colleges of Health and Medicine, each having their own specialism, and helpfully, their own archivist to nominate websites for inclusion. The National Library of Luxembourg similarly emphasised the importance of developing new partnerships to enhance the richness and diversity of collections: “This COVID-19 experience is also useful for the future: we aim to find national partners for thematic collections to expand our lists and collections” (Schafer 2020a, p. 8).¹⁹

Another opportunity that arose from the intense collecting activity around COVID-19 was an increased awareness of the importance of web archiving within the host institution, as well as within the library and archive community more generally. At the Library of Congress, web archive collections are usually “done by individual divisions and not library-wide, but this was meant to be a more library-wide initiative” requiring a new collection proposal process that could be adapted for similar efforts going forward. The result was that there was a more “holistic view” to approaching the collection (Holownia & Geeraert 2022, p. 7).²⁰

Similarly, the British Library web archiving team was part of an internal network of departments collecting different formats of material about the pandemic and used the opportunity of COVID-19 collecting to promote their work through internal networks as well as their externally facing press and social media teams. The team also participated in the Wellcome Trust Network and was able to take advantage of the opportunity to engage with a wider network of stakeholders to promote the role of web archiving in preserving contemporary history.

The Library of Congress experienced a reinvigoration of relationships with other archiving institutions due to common work on COVID-19 collecting, via a federal government web archiving interest group which previously met sporadically but was revived

Oral histories and scalable reading

during the pandemic. The group included organisations such as the National Library of Medicine National Archives, the Government Publishing Office (GPO), and the Smithsonian Institution who all took part in knowledge-sharing calls.

Another creative response that came from the Library of Congress as a result of building its COVID-19 collection was to create a rubric that informed the selection process and prioritised content. This was a vital innovation as the team was dealing with an unprecedented amount of content that potentially could be collected.

We scored the content on things like: “Is this a gap area?” and “What are the permissions?” [...] “Does it have a unique value?”, “Is this duplicated information?”, “What’s the evidentiary value?”, “Is this talking about a specific event or creative act?”, “What’s the informational value?”, “Are there statistics in this?”, “Are there embedded documents?”

(Holownia & Geeraert 2022, p. 8)²¹

Resources that scored highly were added to the collection.

Conclusion

The value of oral history for internet and web histories is well recognised (see, for example, Brügger & Goggin 2022), and in this chapter, we have adopted an oral history approach to contextualise a set of COVID-19 web archive special collections and to shed light on the collecting practices that underpin them. We have also compared different methods for analysing a corpus of interviews, choosing not to place distant and close reading in binary opposition, but to illustrate how they can be combined through scalable reading. Scalable reading has allowed us not to explore the content of the COVID-19 collections themselves but to tell stories about their creation. This approach has highlighted similarities of method and practice across cultural heritage institutions with a collecting remit, but also important differences based on resourcing, institutional context, legislative frameworks, language, and levels of engagement with the wider public. It has also demonstrated the extraordinary ability of web archiving teams to respond rapidly and imaginatively to a highly disruptive global event and to sustain that response over a period of years rather than weeks or months. These efforts will ensure that researchers now and, in the future, will be able at least partially to reconstruct the role of the web in sustaining people through the pandemic, from sites of official advice and guidance to forums for expressing loss and grief. They will also be able to reconstruct how these vital born-digital archives were curated, and what might be missing as well as present. As Alexandre Faye noted, during the interview with the web archiving team at the National Library of France,

This is an event that everyone lived through, one that was documented in real time. What happened needs to be discussed, debated, and accessible to citizens so that we can analyse the legacy of the event. Web archiving contributes to the memory of the event, including from the perspective of the general public and society at large.

(Gebeil & Schafer 2020, p. 11)

Care has been taken to ensure that the interviews conducted as part of the WARCnet oral history initiative will remain accessible to researchers and practitioners beyond the lifetime

of funding for the project. All of the interviews have been published openly online, but they have also been submitted to the Wayback Machine (it is these archived versions that are referenced in this volume). But it remains the case that this unique documentation of special collection processes exists in isolation from the collections themselves, as is the case for other similar interviews (for example, those conducted as part of the “ASAP: Archives sauvegarde attentats Paris” project, which explored the archiving of the 2015 terrorist attacks in Paris and Saint-Denis). Schneider et al. (2003) began to suggest processes for the co-creation of collection development policies, involving researchers and practitioners. We would suggest that oral history interviews, like those that form our corpus, might be a space in which this co-creation of documentation can occur. The WARCnet interviews brought researchers (a key group of end users) into conversation with archivists, curators, and technologists to explore and record a unique moment of web archiving creative practice. If such an approach were to be routinely integrated into collecting workflows, oral histories would provide invaluable context for both researchers using web archives and archivists concerned with developing best practices, as well as helping institutions to benchmark and evaluate their collecting activities. And, as this article has shown, these “interviews as documentation” soon become an invaluable object of study in their own right.

Acknowledgments

We would like to thank all the interviewers and interviewees for having taken the time to take part in this study. In alphabetical order: David Benoist (BnF), Nicola Bingham (British Library), Boris Blanckemane (INA), Niels Brügger (Aarhus University), Peter de Bode (National Library of The Netherlands), Karin de Wild (Leiden University), Ben Els (BnL), Alexandre Faye (BnF), Joanna Finegan (National Library of Ireland), Sophie Gebeil (Aix-Marseille University), Friedel Geeraert (KBR — Royal Library of Belgium), Abbie Grotke (Library of Congress), Jennifer Harbster (Library of Congress), Olga Holownia (IIPC — International Internet Preservation Consortium), Stephen Hunt (Royal Danish Library), Anders Klindt Myrvoll (Royal Danish Library), Paul Koerbin (National Library of Australia), Ismini Kyritsis (Istorima), Michael Kurzmeier (Maynooth University), Gulnar Nagashybayeva (Library of Congress), Márton Németh (National Széchényi Library), Kristjana Mjóll Jónsdóttir Hjörvar (The National and University Library of Iceland, Landsbókasafn Íslands — Háskólabókasafn), Caroline Nyvang (Royal Danish Library), Maria Ryan (National Library of Ireland), Valérie Schafer (C²DH, University of Luxembourg), Sabine Schostag (Royal Danish Library), Barbara Signori (Swiss National Library), Pascal Tanesie (BnF), Jérôme Thièvre (INA), Kees Teszelszky (National Library of The Netherlands). We would also like to thank the peer reviewers for this chapter, whose comments helped to strengthen and clarify the argument.

Notes

- 1 The thematic sections in regular type are consistent across the interviews, while those in italics are applicable to subsets of the corpus.
- 2 International Internet Preservation Consortium (2020). Novel Coronavirus (COVID-19). Archived at Wayback Machine February 2, 2024, <https://web.archive.org/web/20240221163711/https://archive-it.org/collections/13529>.
- 3 The Icelandic Web Archive (2023). Vefsafn.is. Archived at Wayback Machine May 10, 2023, <https://web.archive.org/web/20240221163958/https://vefsafn.is/is/>.

Oral histories and scalable reading

- 4 National Library of Australia (2024). Trove Coronavirus (COVID-19) Pandemic and Australia, 2020–2022. Archived at Wayback Machine February 21, 2024, <https://web.archive.org/web/20240221164415/https://webarchive.nla.gov.au/collection/17818>.
- 5 Library of Congress (2024). Library of Congress Coronavirus Web Archive. Archived at Wayback Machine February 21, 2024, <https://web.archive.org/web/20240221164848/www.loc.gov/collections/coronavirus-web-archive/about-this-collection/>.
- 6 UK Web Archive (2022). Coronavirus (COVID-19) UK. Archived at Wayback Machine December 5, 2022, <https://web.archive.org/web/20221205154037/www.webarchive.org.uk/en/ukwa/collection/2975>.
- 7 National Széchényi Library (2024). OSZK webarchívum Browse: Coronavirus Epidemic – 2020. Archived at Wayback Machine February 21, 2024, <https://web.archive.org/web/20240221165726/https://webarchivum.oszk.hu/en/webarchive/browse/browsing-in-the-event-based-subcollections/browse-coronavirus-epidemic-2020/>.
- 8 Interview with Jennifer Harbster, Library of Congress.
- 9 Interview with Peter de Bode, National Library of The Netherlands.
- 10 Shoura (2024). Welcome to Shoura Luxembourg. Archived at Wayback Machine February 21, 2024, <https://web.archive.org/web/20240221171122/https://shoura.lu/en/home/>.
- 11 Interview with Ben Els, National Library of Luxembourg.
- 12 Interview with Nicola Bingham, British Library.
- 13 Interview with David Benoist, National Library of France.
- 14 Newsguard (2024). Transparent Tools to Counter Misinformation for Readers, Brands and Democracies. Archived at Wayback Machine 21 February, 2024, <https://web.archive.org/web/20240221170358/www.newsguardtech.com/>.
- 15 Interview with Jennifer Garbster, Library of Congress.
- 16 Interview with Nicola Bingham, British Library.
- 17 Interview with David Benoist, National Library of France.
- 18 Interview with Boris Blanckemane, Institut National de l'Audiovisuel.
- 19 Interview with Ben Els, National Library of Luxembourg.
- 20 Interview with Abbie Grotke, Library of Congress.
- 21 Interview with Jennifer Harbster, Library of Congress.

References

- Acker, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science*, 20 (2020), 105–123. <https://doi.org/10.1007/s10502-019-09325-9>
- Ahnert, R., Griffin, E., Ridge, M., & Tolfo, G. (2023). *Collaborative Historical Research in the Age of Big Data: Lessons from an Interdisciplinary Project*. Cambridge University Press. <https://doi.org/10.1017/9781009175548>
- Bell, M., Storrar, T., & Winters, W. (2022). Web archives and the problem of access: Prototyping a researcher dashboard for the UK Government Web Archive. *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections*. Bielefeld University Press. <https://doi.org/10.1515/9783839455845-003>
- Booth, A. (2017). Mid-range reading: Not a manifesto. *PMLA/Publications of the Modern Language Association of America*, 132(3), 620–627. <https://doi.org/10.1632/pmla.2017.132.3.620>
- Brügger, N., & Goggin, G. (2022). *Oral Histories of the Internet and the Web*. Routledge. <https://doi.org/10.4324/9781003319191>
- Clavert, F., & Fickers, A. (2021). On pyramids, prisms, and scalable reading. *Journal of Digital History*, no. jdh001. <https://journalofdigitalhistory.org/en/article/jXupS3QAeNgb>
- Day, M. (2003). Preserving the Fabric of Our Lives: A Survey of Web Preservation Initiatives. In T. Koch & I. T. Sølberg (Eds.), *Research and Advanced Technology for Digital Libraries* (vol. 2769, pp. 461–472). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-45175-4_42
- De Wild, K., & Kyritsis, I. (2021). *Exploring Special Web Archives Collections Related to COVID-19: The Dutch Web Archive (KB)*. WARCnet Papers. Archived at Wayback Machine September 24, 2023, <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.

- Gebeil, S., & Schafer, V. (2020). *Exploring Special Web Archives Collections Related to COVID-19: The Case of the French National Library (BnF)*. WARCnet Papers. Archived at Wayback Machine September 24, 2023, <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.
- Geeraert, F. (2020a). *Exploring Special Web Archives Collections Related to COVID-19: The Case of the IIPC Collaborative Collection*. WARCnet Papers. Archived at Wayback Machine September 24, 2023, <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.
- Geeraert, F. (2020b). *Exploring Special Web Archives Collections Related to COVID-19: The Case of the UK Web Archive*. WARCnet Papers. Archived at Wayback Machine September 24, 2023, <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.
- Hammond, A. (2017). The double bind of validation: Distant reading and the digital humanities — trough of disillusionment. *Literature Compass*, 14(8), e12402. <https://doi.org/10.1111/lic3.12402>
- Holownia, O., & Geeraert, F. (2022). *Exploring Special Web Archives Collections Related to COVID-19: The Case of the Library of Congress*. WARCnet Papers. Archived at Wayback Machine September 24, 2023, <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.
- Jensen, H. S., Jensen, J. M., Thygesen, A. U., & Pedersen, M. O. (2023). Digital methods in memory studies: A beginner's guide to scalable reading of Twitter data. *Memory Studies*, e-pub ahead of print. <https://doi.org/10.1080/17506980.2023.1197126>.
- Littman, J., Chudnov, D., Kerchner, D., Peterson, C., Tan, Y., Trent, R., Vij, R., & Wrubel, L. (2018). API-based social media collecting as a form of web archiving. *International Journal on Digital Libraries*, 19(1), 21–38. <https://doi.org/10.1007/s00799-016-0201-7>
- Locatelli, E. (2017). Blogs as cultural products: A multidimensional approach to their diffusion in Italy (2001–2008). In N. Brügger (Ed.), *Web 25: Histories from the First 25 Years of the World Wide Web* (pp. 91–105). Peter Lang Publishing.
- Mackinnon, K. (2022). Databound: Histories of growing up on the World Wide Web. PhD thesis, University of Toronto. https://tspace.library.utoronto.ca/bitstream/1807/125246/2/Mackinnon_Katherine_202211_PhD_thesis.pdf.
- Maemura, E., Worby, N., Milligan, I., & Becker, C. (2018). If these crawls could talk: Studying and documenting web archives provenance. *Journal of the Association for Information Science and Technology*, 69(10), 1223–1233. <https://doi.org/10.1002/asi.24048>
- Moretti, F. (2000). Conjectures on world literature. *New Left Review*, 1(1), 54–69. <https://newleftreview.org/issues/i11/articles/franco-moretti-conjectures-on-world-literature>
- Nolan, S. (2017). Born outside the newsroom: The creation of the Age Online. In N. Brügger (Ed.), *Web 25: Histories from the First 25 Years of the World Wide Web* (pp. 107–122). Peter Lang Publishing.
- Pedersen, M.O., Møller Jensen, J., Harbo Johnston, V., Thygesen, A. U., & Strandgaard Jensen, H. (2022). Scalable reading of structured data. *Programming Historian*, 11. <https://doi.org/10.46430/phen0103>.
- Pehlivan, Z., Thièvre, J., & Drugeon, T. (2021). Archiving social media: The case of Twitter. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 43–66). Springer International Publishing. https://doi.org/10.1007/978-3-030-63291-5_5
- Reinert, M. (1983). Une méthode de classification descendante hiérarchique: Application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, 8(2), 187–198. www.numdam.org/item/CAD_1983__8_2_187_0/
- Reinert, M. (1993). Les “mondes lexicaux” et leur “logique” à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, 66(1), 5–39. <https://doi.org/10.3406/lsoc.1993.2632>.
- Schafer, V. (2020a). *Exploring Special Web Archives Collections Related to COVID-19: The Case of the BnL*. WARCnet Papers. Archived at Wayback Machine September 24, 2023, <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.
- Schafer, V. (2020b). *Exploring Special Web Archives Collections Related to COVID-19: The Case of INA*. WARCnet Papers. Archived at Wayback Machine September 24, 2023, <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.

Oral histories and scalable reading

- Schafer, V., Musiani, F., & Borelli, M. (2016). Negotiating the web of the past. *French Journal for Media Research, French Journal for Media Research*, 6. <https://frenchjournalformediaresearch.com/lodel-1.0/main/index.php?id=952>.
- Schafer, V., Truc, G., Badouard, R., Castex, L., & Musiani, F. (2019). Paris and Nice terrorist attacks: Exploring Twitter and web archives. *Media, War & Conflict*, 12. <http://dx.doi.org/10.1177/1750635219839382>.
- Schneider, S. M., Foot, K., Kimpton, M., & Jones, G. (2003). Building thematic web collections: Challenges and experiences from the September 11 Web Archive and the Election 2002 Web Archive. In J. Masanès, A. Rauber, & G. Cobena (Eds.), *Proceedings of the 3rd Workshop on Web Archives, Trondheim, Norway, in Conjunction with the 7th European Conference on Research and Advanced Technologies for Digital Libraries* (pp. 77–93). (n.p.). Archived at Wayback Machine September 4, 2024, <https://rb.gy/8ylatj>
- Sinclair, S., & Rockwell, G. (2016). Voyant tools. <http://voyant-tools.org/>. Archived at Wayback Machine September 4, 2024, <https://web.archive.org/web/20240904114710/https://voyant-tools.org/>
- Winters, J. (2019). Negotiating the archives of UK web space. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities: The Case of National Web Domains* (pp. 75–88). Routledge.

9

SURVEYING THE LANDSCAPE OF COVID-19 WEB COLLECTIONS IN EUROPEAN GLAM INSTITUTIONS

An explorative analysis

*Nicola Bingham, Karin de Wild, Friedel Geeraert, and
Caroline Nyvang*

Introduction

During the coronavirus (COVID-19) pandemic, cultural heritage institutions initiated large efforts to capture this historical moment for future generations. A sub-group of researchers from the WARCnet project was interested in doing an explorative comparative analysis of the web archiving practices surrounding the creation of COVID-19 collections in the European Galleries, Libraries, Archives, and Museums (GLAM) sector.

The World Health Organization (WHO) declared COVID-19 a public health emergency on 30th January 2020, escalating it to pandemic status on 11th March 2020 (World Health Organization 2023a), with the emergency status ending on 5th May 2023 (World Health Organization 2023b).

Research conducted by Burke et al (2020) in the context of museums and cultural institutions, attests to an increase in online content during the pandemic, attributing both to the closure of physical spaces and to the curation of collections that specifically address themes related to the pandemic. Furthermore, an empirical study by Lemenager et al. (2021) delves into the dynamics of web content consumption during the lockdown period in Germany. Their study provides compelling evidence of a marked escalation in internet usage concurrent with the COVID-19 lockdown.

This chapter provides an exploratory analysis rather than a full assessment of web archives, highlighting relevant studies. Vlassenroot et al. (2019) examined nine European web archives, offering insights into their infrastructures useful for digital scholars. Additionally, Ogden and Maemura (2021) analysed how socio-technical and organisational challenges in web archiving affect research use of national web collections, specifically the UK Web Archive at the British Library and the Netarchive at the Royal Danish Library.

Surveying the landscape of COVID-19 web collections

The two research questions underlying this chapter are (1) What are the main similarities and differences between European countries in creating COVID-19 web archiving collections, and (2) What, if any, is the impact of differences in the maturity of web archiving programmes? Maturity was interpreted as encompassing the duration of the web archiving programmes, how web archiving practices have evolved over time, and the scale on which web archiving is done.

A survey among European partners in the GLAM sector was decided upon to conduct the explorative analysis. The chapter focuses on three different groups of respondents based on their country of origin: the UK, Belgium, and other European countries. This division was made because of the relatively high number of respondents from the UK and Belgium which was due to the way the survey was distributed.

The first section will focus on the methodology that was used to set up and distribute the survey. In the second section, the results of the survey are presented, focusing on respondents and their institutions, the reasons behind collecting and collaborations that were set up, the scope of collections, technology and quality assurance, access, and challenges. Further analyses of the survey results are included in the discussion after which the conclusion is presented.

Methodology

Across Europe, web archiving is carried out by different institutions operating at local, regional, and national levels. The questions we have included and excluded in our survey are informed by a hypothesis about the decisive factors that form web archives. In contrast to physical archives, archiving the web is a novel endeavour characterised by technological advances and little international coherence. Local strategies are governed by specific national legislation and new tools for harvesting are continuously being developed as web archiving technology evolves.

Yet our choice of questions also leaves us with some blind spots that do not allow us to explore other considerations, such as funding opportunities and/or political priorities. The discussions and findings of this chapter should be read with that bias in mind.

The survey covered various topics, starting with questions about the respondents' backgrounds (Bingham et al. 2022). It then delved into understanding why people collected COVID-19-related information, the scope of their collections, when they started, how often they captured data, and the challenges they faced. The notable findings from each section are highlighted in the chapter.

Launched on 13th July 2022, it ran for approximately three months, closing on 5th September 2022. It was distributed to approximately 600 email addresses and distribution lists, such as the Archives-NRA JISC mail and IIPC distribution lists, which had been collated by the project team from their existing contacts and networks and by conducting desk-based research. There was a bias in terms of the distribution of contacts in favour of the UK and Belgium due to two members of the project team being based at National Libraries in those countries. In the UK, 99 organisations or individuals were contacted and 354 in Belgium.

Respondents

In total, 61 respondents from 12 countries filled out the survey, meaning that there was a completion rate of approximately 10%. Potential factors contributing to this relatively low

response rate include the distribution timing during the summer coinciding with the absence of university staff due to leave, ambiguity regarding the intended recipients within large organisations, and limited availability of resources in smaller organisations. Furthermore, despite efforts to include organisations that did not collect COVID-19 web archives, the survey's title and overall thrust may have inadvertently discouraged those individuals from responding. Acknowledging the low response rate as a potential limitation emphasises the need for a cautious interpretation of the survey results.

In what follows, the focus will be on comparing COVID-19 web archiving practices in three geographic regions: the UK, Belgium, and other European countries. In total, 33% of respondents were based in the UK, 25% in Belgium, and the remaining 41% from other European countries.

When looking at the size of the organisations, there are differences between the UK, Belgium, and other European countries. For ease of interpretation, we consider organisations that have 50 or fewer employees small, between 50 and 250 employees middle, and more than 250 large. It is assumed that smaller countries have fewer large organisations than bigger countries. In Belgium, only 8% of respondents work at a large organisation, 23% work at a middle-sized organisation and 69% work at a small organisation. In the UK, 29% work at a large organisation, 24% work at a middle-sized organisation and 47% work at a small organisation. In other European countries, 50% of respondents work in large organisations, 47% of respondents work in a middle-sized organisation and only 3% work in a small organisation. The percentages of small organisations are markedly higher in the UK and Belgium compared to other European countries.

Organisation types differ between countries. In the UK and other European countries, a mix of galleries (6% in the UK), archives (47% and 10% respectively), libraries (18% and 73% respectively), museums (18% and 3% respectively) and universities (7% in other European countries) replied to the survey (see Table 9.1). In Belgium, only two organisation types were indicated: library (23%) and archive (77%). This is telling given that the mailing list for Belgium included 354 addresses, many of which were museums. The near absence of answers from universities can also be noted. This might be because the survey ran over the summer when many universities were closed. The fact that in other European countries, there were more respondents from libraries than in Belgium and the UK is also because the survey was mainly distributed to European National Libraries and Archives which are also members of the IIPC. Looking at the level at which the organisations are active, there is again a marked difference between the UK and Belgium on the one hand and the other European countries on the other (see Table 9.2). Only 7% of respondents in other European countries work for a regional organisation, compared to 35% in the UK and 46% in Belgium. Furthermore, 93%

Table 9.1 Organisation type

	<i>UK</i>	<i>Belgium</i>	<i>Other European countries</i>
Archive	47%	77%	10%
Library	18%	23%	73%
Museum	18%	—	3%
Other	12%	—	7%
Gallery	6%	—	—
University	—	—	7%

Surveying the landscape of COVID-19 web collections

Table 9.2 Level of the organisation

	<i>UK</i>	<i>Belgium</i>	<i>Other European countries</i>
International	12%	—	—
National	47%	38%	93%
Regional	35%	46%	7%
Other	6%	15%	—

Table 9.3 Job profile

	<i>UK</i>	<i>Belgium</i>	<i>Other European countries</i>
Archivist	82%	69%	12%
Librarian	9%	—	36%
Curator	9%	—	20%
Engineer	—	8%	8%
Communication	—	8%	4%
Director	—	—	4%
Other	—	15%	16%

of respondents in other European countries work for a national organisation, compared to 47% in the UK and 38% in Belgium. These differences can again be attributed to the way the distribution list was compiled. Organisations working on the international level were only registered in the UK at 12%. Six per cent of UK respondents and 15% of Belgian respondents classified the level of their organisation as ‘Other’.

Job profiles also differ between the UK and Belgium on the one hand, and the other European countries on the other (see Table 9.3). In Belgium and the UK respectively, 69% and 82% of respondents describe their job profile as archivist, compared to only 12% of the respondents in other European countries. Thirty-six percent of respondents in other European countries describe themselves as librarians, compared to only 9% in the UK. None of the respondents in Belgium would describe themselves as librarians. Other profiles that are mentioned are curator, engineer, communication, director, and others.

Motivation and collaborations

Participants were asked to report whether their institution had indeed curated a special collection of web materials pertaining to the COVID-19 crisis.

Of the UK respondents, 63% answered that their institution created a special COVID-19 collection of web material and/or social media, compared to 77% of participating Belgian institutions. Among other European countries, 72% indicated that they had curated such a collection on the COVID-19 crisis.

Regarding the responsibility for taking the initiative, approximately half of the respondents from the UK (47%) and Belgium (54%) reported that the initiative came from the staff; about 30% overall stated that the initiative came from the management of the institution (UK: 32%; Belgium: 31%). Similar to the Belgian case, 54% of the respondents from other European countries reported that the initiative had mainly come from the staff.

Collaboration with partners for technical support was quite common among all respondent groups. Fifty percent of respondents from Belgian institutions confirmed they undertook such partnerships. In other European countries, 55% had been engaged in similar collaborations. Whereas with the UK group, it was almost a tie between institutions that had (44%) and had not (50%) collaborated with other institutions for technical support.

Belgian respondents indicated that recommendations from the broader population had been collected primarily using invitations via social media channels (30%). In the UK, 40% of respondents had employed social media to this end. Outreach via mailing lists was utilised by 30% of Belgian respondents, whereas this only applied to 10% of the UK and 3% of the other European countries' institutions.

The survey elicited responses about whether efforts were made to reach out to underrepresented groups. No respondents from Belgium responded affirmatively, however, 55% of UK respondents indicated that they 'agreed' or 'strongly agreed' with the statement "Efforts were made in my organisation to contact groups that are underrepresented in collections when creating the collection of archived web content related to COVID-19." For other European countries 25% 'agreed' or 'strongly agreed' with the statement. Only UK respondents provided information as to how outreach to underrepresented groups was managed. Several options were mentioned: contacting mental health charities or support groups via email, and social media writing to existing contacts at partner institutions, and making use of internal networks such as the LGBTQ+ staff network.

Selection scope

In terms of scoping the collection, most UK respondents indicated that they collected on a national scale (40%), while in Belgium most respondents (36%) indicated that they collected on a local scale. In the other European countries, most indicated that they collected on a national scale (68%). To make more sense of these results, we compared this with the number of respondents who indicated they worked for a local, regional, national, or global organisation. In the UK, 47% indicated they operate at a national level, 35%, at the regional level, and 12% at the international level. In Belgium 46% of the institutions operate at the regional level, 38% at a national level, and 15% at a local level. It seems, therefore, that there is a correlation between the level at which the institution operates and the level at which collecting was focused.

All respondents — including those who did not conduct a collection of web materials — were also asked whether they had been collecting other types of materials related to the COVID-19 crisis. The responses showed marked differences between the two countries. In the UK, video (22%), oral histories (16%), and archives (16%) constituted the primary forms of collected materials, whereas, in Belgium, archives (18%), publications (18%), and physical objects (13%) had been prioritised.

Regarding the type of website/social media content collected, in Belgium, the top five types of content collected were listed, Facebook as the most common (15%), followed by Twitter, Instagram, and YouTube (10%). In terms of websites, non-profit organisations were the most commonly collected (10%). In the UK, the top five were slightly different; Twitter was the most common (13%) followed by blogs, university websites, and non-profit websites (11%), and then YouTube, government agencies, and cultural institutions' websites (9%).

Various types of 'other content' were cited including 'LinkedIn', 'websites of student associations', 'academic research', 'the output of project sites', 'digital photos', 'video content',

Surveying the landscape of COVID-19 web collections

‘Flickr’, ‘and websites of individuals with relevant historical content’ and ‘websites of public inquiries’.

In terms of which criteria were used to scope the collections, in Belgium, the top three answers to this question were, firstly ‘based on subject/theme of web content’ (36%), followed by ‘based on nationality of the creators/owners’ and ‘based on specific social media profiles’ (21%). In the UK, the most common response was ‘based on subject/ theme of web content’ (32%), followed by ‘based on top-level domain’ (21%), and ‘based on nationality of the creators/owners’ (16%). In Belgium, no respondents indicated that the top-level domain or the postal addresses mentioned in the web content were a determining criterion; in the UK, none indicated that language was used as a selection criterion. In the other European countries category, the top response was also ‘based on subject/theme of web content’ (28%), followed by ‘based on nationality of the creators/owners’ (21%), confirming that the most common way to scope COVID-19 collections among all the survey respondents was based on subject/theme of the web content.

Other collection scopes that were mentioned were ‘creator-based collections’ and ‘local collections’. Specific themes or subjects around which web content was collected in Belgium comprised: ‘liberalism, freethinking and trade’, ‘solidarity and inequality’, and ‘student and campus life’. In the UK, one respondent said “websites were selected by searching on related keywords and relevant types of publication/organisations, e.g. news sites, medical institutes, government websites, etc.” One respondent indicated that as the call for participation in the collection was open, they “took everything that was sent to us.”

Temporality also affects the corpus generated from a special collection. Therefore, respondents were asked about the start and potential end date of the collecting process. In the UK (25%) and Belgium (45%), the most common time to start collecting was between the 16th and 31st of March 2020. In other European countries, the most common start date was between 1st and 15th March 2020 (35%) (see Table 9.4).

At the time of the survey (June–September 2020), 25% of respondents in the UK and 45% of respondents in Belgium had already stopped collecting, compared to 28% in other European countries (see Table 9.5). When asked what the reason was behind the decision to stop collecting, the following elements were mentioned: faded interest, staff time needing to be filled in differently, end of a project, end of pandemic-related measures, most of the relevant content was no longer online, and technical difficulties when capturing certain social media.

Table 9.4 Start date of COVID-19 collections

	<i>UK</i>	<i>Belgium</i>	<i>Other European countries</i>
Before February 2020	—	9%	4%
1–15 February 2020	6%	—	15%
16–19 February 2020	6%	9%	4%
1–15 March 2020	19%	9%	35%
16–31 March 2020	25%	45%	23%
1–15 April 2020	19%	9%	4%
16–30 April 2020	13%	9%	4%
After April 2020	13%	9%	12%

Table 9.5 Overview of relative number of institutions who have (not) stopped collecting content at time of survey

	UK	Belgium	Other European countries
Yes	25%	45%	28%
No	75%	45%	68%
I don't know	—	9%	4%

It is important to acknowledge that the survey was conducted between June and September 2022, a period during which many collecting institutions experienced a deceleration or de-prioritization of their collection activities related to the pandemic. This was in contrast to the heightened level of activity observed during the initial two years of the pandemic in 2020 and 2021. The unpredictable nature of the pandemic made it challenging for organisations to accurately anticipate when they would cease their collection efforts. Therefore, it is crucial to consider this contextual factor when interpreting the survey findings.

Technical aspects

With regards to what tools were used to collect the web content, answers differed between both countries. In Belgium, the two most common tools were: Webrecorder, followed by the Chrome SingleFile extension. In the UK, these were Webrecorder and Heritrix. In other European countries, Webrecorder and Archive-It are the two most common tools (see Table 9.6). The webrecorder is the most common tool in all the countries included in the study.

It is worth noting that there is a very long tail of tools that were only indicated once or twice by respondents in the survey (see Table 9.6). They are characterised by being open-source and simple to use. Furthermore, some respondents mentioned tools developed for replaying archived web content such as PyWB and Open Wayback.

In the UK, 14% of respondents reported not undertaking any quality control measures, followed closely by 17% in Belgium and 16% in other European countries. Conversely, patch crawling was embraced by a small portion of respondents, with 7% in the UK, 8% in Belgium, and 10% in other European countries engaging in this practice. A significant majority conducted a visual check of archived web content, with 64% of UK respondents, 58% in Belgium, and 55% in other European countries reporting this method. Additionally, the use of specific software or tools for quality control was reported by 14% of UK respondents, 8% in Belgium, and 10% in other European countries, indicating a heavy reliance on visual quality checks on archived content across Europe.

Regarding the long-term preservation of the collections, 47% of the UK respondents indicated that strategies are in development while 24% indicated that they do not (yet) have resources for long-term preservation. Eighteen percent of the UK respondents make use of a dedicated long-term preservation system or software in-house and 12% use preservation systems or software external to their institution. In Belgium, the picture regarding long-term preservation is different: 38% of the respondents use a dedicated long-term preservation system/software in-house, 31% indicated that long-term preservation strategies are in development, 23% of respondents use preservation systems or software external to the organisation and only 8% indicated that they do not have resources (yet) for long-term preservation (see Table 9.7). This is surprising given that web archiving in the UK is a more

Surveying the landscape of COVID-19 web collections

Table 9.6 Tools used to collect web content

	UK	Belgium	Other European countries
Webrecorder crawler	5	7	10
Brozzler	1	—	6
Archive-It	2	1	8
NetArchiveSuite	—	—	5
API of social media platform	2	2	5
Heritrix	3	2	6
Instaloader	1	2	—
WebCuratorTool	—	—	3
Chrome SingleFile extension	—	3	—
Firefox extension	—	1	—
Youtube-Dl	—	1	—
Twarc	—	1	—
Facebook scraper	—	1	—
HTTrack	—	1	—
SocialFeedManager	—	1	—
Conifer	1	—	—
PyWB	1	—	—
Browsertrix	1	—	—
Web harvesting tool in digital preservation software	1	—	—
FTP	—	—	1
RSS	—	—	1
Self-developed technology	—	—	1
Open Wayback	—	—	1
Web form	—	—	1
E-mail	—	—	1

Table 9.7 Long-term preservation

	UK	Belgium	Other European countries
Dedicated long-term preservation system/software in house	18%	38%	47%
Preservation system/software external to the organisation	12%	23%	13%
Strategies are in development	47%	31%	34%
We do not (yet) have resources for long-term preservation	24%	8%	6%

well-established practice than in Belgium. In other European countries, 47% use dedicated long-term preservation software in-house, 13% use preservation software external to the organisation, 34% indicated that strategies are in development, and only 6% indicated that there are no resources for long-term preservation.

Access

Respondents were surveyed about access arrangements to their COVID-19 collections. The most common response among respondents in the UK and Belgium was ‘onsite (e.g.

reading room)’ (UK, 33%, Belgium, 27%). In the UK, the next most common method was with ‘playback software’ (25%) and then ‘via the catalogue’ (21%). In Belgium, after ‘onsite’ access, the next most common access modes were listed as ‘metadata’ (23%) and ‘via the catalogue’ (19%). For the other European countries, the answers were slightly different, in that ‘playback software’ was the most common (24%) followed by ‘onsite’ (19%) and by ‘seed lists’ (17%).

‘Playback software’ was only the fourth most common answer in Belgium. No Belgian or UK institutions make the collections accessible by means of derived data sets or by sharing the seed lists with the public, whereas this is done in other European countries. Via ‘seed lists’ was the third most common response for the ‘other European countries category’ (17%) after ‘playback software’ and ‘onsite’ access and the figure for derived datasets was (7%).

Providing data-level access or offsite access to the collections is very rarely done in either the UK or Belgium. A relatively small proportion of respondents indicated that the data would not be made accessible at all, for Belgium the response was (8%) and for other European countries (3%).

Regarding restrictions applied to the accessibility of the collection data, the most important restriction mentioned across all groups, the UK, Belgium, and other European countries, was the fact that the collection may only be made available in the reading rooms of the collecting institution. In the UK, restrictions around accessibility also included the restriction to the reading room of the institution, but ‘intellectual property rights’, the ‘protection of personal data’, and the ‘inability to provide the corpora for text and data mining due to the Non-Print Legal Deposit Legislation, 2013’ were also mentioned. In other European countries, further access restrictions included that the collection could only be made available to academic researchers and in one case the lack of an interface meant that no access could be provided. One respondent indicated that there were no access restrictions for the collected data.

One of the main challenges that follows from access restrictions is that in Belgium, the UK, and other European countries, special collections were rarely used for research and teaching. The Belgian respondents indicated that the collections were used for research and teaching either ‘rarely’ (42%), ‘never’ (25%), or ‘I don’t know’ (33%), clearly indicating that insight into the use of the collections is lacking in about a third of the cases or that the collections are underused. In the UK, the majority of respondents indicated that they ‘did not know’ if the collection was used for research or teaching purposes (45%) and in the other European countries the figure was (48%).

The survey also canvassed information about how national legal deposit regulations were considered limiting regarding collecting or providing access to the collections. In the case of Belgium and the other European countries, the consensus was that legal deposit regulations were most limiting for ‘access’ (Belgium (44%), other European countries (34%)). Whereas in the UK, the legal deposit regulations were considered most limiting regarding ‘collecting websites/web pages’ (36%). It was also the case that in Belgium 25%, the UK 16%, and the other European countries 20% of the respondents considered that the national legal regulations were not limiting to collecting nor providing access to the collections.

Challenges

Respondents were asked what they deemed to be the most important challenges for the COVID-19 special collection. They were given several options that they had to rate on a scale from not important (1) to very important (4). When adding the answers that were

Surveying the landscape of COVID-19 web collections

Table 9.8 Most important challenges for the COVID-19 special collections

	UK	Belgium	Other European countries
Organising the web archiving from home	5%	7%	8%
Scoping the collection	15%	16%	15%
Collecting the content (technical set-up, etc.)	14%	14%	13%
Creating collaborations with partner institutions	6%	9%	7%
Providing access to the collection	6%	0%	3%
Quality control	9%	7%	12%
GDPR/privacy-related issues	11%	14%	4%
Long-term preservation	14%	17%	9%
Securing sufficient funding	4%	2%	9%
Securing additional curatorial staff members	5%	3%	10%
Securing additional technical support	3%	10%	7%
Other	9%	2%	3%

indicated as important or very important, the top four challenges are similar in Belgium and the UK, but only the first two appear in the top four of other European countries (see Table 9.8). ‘Scoping the collections’ (Belgium: 16%, UK: 15%, other European countries: 15%), ‘collecting the content (technical set-up, etc.)’ (Belgium: 14%, UK: 14%, other European countries: 13%), ‘long-term preservation’ (Belgium: 17%, UK: 14%, other European countries: 9%), and ‘General Data Protection Regulation (GDPR)/Privacy-related issues’ (Belgium: 14%, UK: 11%, other European countries: 4%) were deemed most important in Belgium and the UK. In other European countries, the third and fourth most important challenges are ‘quality control’ (12%), compared to 7% in Belgium and 9% in the UK, and ‘securing additional curatorial staff’ (10%), compared to 3% in Belgium and 5% in the UK. A significant difference between the countries is that in Belgium ‘securing additional technical support’ (10%) is seen as a bigger challenge than in the UK (3%).

Respondents who indicated ‘other’ were asked which other challenges they encountered. No responses were obtained from Belgium, but in the UK, respondents also indicated that encouraging participation and submission of recommendations, learning to deliver online training to staff from other teams who were adapting to working from home, managing the increase in workload, and seeking permissions from right holders to collect web content (in a non-legal deposit context) was challenging. In other European countries, ensuring that work is not duplicated, making the right choices under the stress of time, data cleaning, training, and coding amid a pandemic, not having a platform to share methodologies with other researchers, the long duration of the pandemic and the sheer size of the material were seen as challenges. The pandemic and the general uncertainty it caused clearly brought along changes in work practices.

Discussion

This chapter aimed to do an explorative comparative analysis of the web archiving practices surrounding the creation of COVID-19 collections in the European GLAM sector. The data collected through the survey gave insight into the motivations behind setting up the collections, collaborations between institutions, selection scopes, technical aspects, and

perceived challenges. Two research questions formed the basis of the analysis: (1) what are the main similarities and differences between European countries in creating COVID-19 web archiving collections and (2) what, if any, is the impact of differences in the maturity of web archiving programs?

Regarding the first research question, the main similarities and differences, only the differences and similarities that were deemed of most interest will be highlighted in this section. There are numerous similarities between the three groups of respondents. About half of the respondents in the three groups collaborate with partners for technical support. This indicates that the need for collaboration and knowledge sharing between web archiving institutions is continuous and essential.

The temporality of the collecting efforts was also similar in Belgium and the UK where most initiatives started collecting between the 16th and 31st of March. These dates correlate with the first lockdown: in Belgium, lockdown started on 18th March 2020 and in the UK lockdown measures legally came into force on 26th March 2020 (The Institute for Government, 2021; Belga, 2021).

Approaches to quality control are also similar between all countries. Between 50% and 65% of respondents of the three respondent groups visually checked the archived web content, compared to around 15% of the respondents who indicated there was no quality control.

In both Belgium and the UK, the most common way to access the collections is on-site, for example in the reading room of the institution. It is also important to note that in all response groups, insight into the use of COVID-19 collections for research or teaching purposes is too limited or non-existent, clearly illustrating the need for outreach and promotion of the unlocked research potential of collections of archived web content.

Regarding the challenges, the same four aspects were deemed most important in Belgium and the UK: scoping the collections, collecting the content, long-term preservation, and GDPR/privacy-related issues. As shown by Friedel Geeraert et al. (2024), delimiting the COVID-19 collections was seen as particularly difficult because the pandemic influenced almost all facets of society and therefore, left many traces on the web.

The survey data also highlighted differences between the three groups of respondents. The respondent data yielded interesting insights for example. In contrast to the UK and other European countries, no responses from Belgian galleries or museums were collected even though Belgian museums were featured on the distribution list. No web archiving activities undertaken at galleries or museums were discovered in Belgium during a recent evaluation of the state of the art of web archiving in Belgium (Messens, 2023). This can underscore the fact that web archiving is a relatively new practice in Belgium compared to most other countries in Europe and that the practice still needs to spread to galleries and museums (meemoo, 2023) (for more information about the early history of web archiving at European national libraries, see Brügger & Schroeder (2017) and Webster (2019)). It also indicates the need for raising more awareness about web archiving as a concept in Belgium.

In addition, relatively few Belgian respondents (8%) work at a large organisation (more than 250 employees), as compared to 29% in the UK and 50% in other European countries. This corresponds well with the findings in a report produced in 2022 that analysed existing web and social media archiving initiatives in Belgium. Seventy-eight percent of initiatives studied in the report described themselves as a thematic or a local archive (Messens & Geeraert, 2022). These institutions are all small and do not have many human or financial resources to invest in archiving websites and social media. This is also reflected in the choice

Surveying the landscape of COVID-19 web collections

of tools as they prefer tools that do not require a lot of technical expertise or infrastructure to install. Heritrix, according to the UK Web Archives' Technical Lead, for example, is a "large complex crawler framework" and is only used by 28% of institutions (Jackson, 2019).

In terms of collaboration, efforts to reach out to underrepresented groups were most common in the UK where 55% of respondents (strongly) agreed with the related statement, in contrast to 0 Belgian respondents and 25% of respondents from other European countries. Inclusion may have been a more important theme in the public debate in the GLAM sector in the UK than in other European countries. There are several examples of initiatives and research in the web archiving sector in the UK that aim to promote inclusive practices and collection development, particularly focusing on underrepresented communities. For example, the Inclusive Terminology Glossary (Cultural Heritage Terminology Network, 2024) seeks to inform decision-making around language use in the heritage sector. The focus on inclusive language reflects a commitment to respectful representation.

Alice Austin's (2023) work, presented at the RESAW Conference in 2023, suggests a focus on preserving and documenting the discourse around trans health in the UK. The collaborative work of Cui et al. (2023), as presented at the same conference, suggests a focus on involving communities in the web archiving process. Additionally, Huc-Hepher and Ma (2024) discuss challenges and strategies for inclusively representing diverse communities in the UK web archives in their chapter on transnational "Community Micro-archives." These participatory approaches align with efforts to create more inclusive and diverse web archives.

Regarding the collection scope, differences exist in the level of archiving. In both the UK and other European countries most respondents indicated that they collect on a national scale (40% and 68% respectively), whereas in Belgium most respondents (36%) collect on a local scale. A possible explanation is the way the web archival landscape is set up in both countries. In Belgium, several web archives are collecting locally (meemoo, 2023), while in the UK there is a network of collecting institutions, all contributing to the national scope, especially with the UK Web Archive which is a partnership of the six UK legal deposit libraries (The British Library, Bodleian Libraries, University of Oxford, Cambridge University Library, National Library of Scotland, National Library of Wales and the Library of Trinity College Dublin), and is contributed to by many other UK GLAM organisations.

Other differences could be noted in the types of other materials the GLAM institutions were collecting alongside web content. In the UK the primary forms of collected materials included video, oral histories, and archives, whereas in Belgium the focus was on archives, publications, and physical objects. This clearly indicates that the UK has a stronger tradition for collecting oral histories, videos, and accompanying materials than GLAM institutions in Belgium.

Marked differences were noted in the types of website and social media content that were collected. It is interesting to note that collecting Facebook is one of the least common options in the UK while it is the top choice for Belgian respondents. A possible explanation for this outcome could be related to the technology that is available in the web archives. It is possible that large UK web archives use Heritrix and therefore are unable to collect Facebook, as this is a known limitation of the Heritrix crawler. These limitations stem from the dynamic, interactive, and personalised nature of social media platforms, as well as the technical and legal barriers imposed by the platforms themselves. Just a few of the limitations of Heritrix in this regard are that it is limited in executing and interpreting JavaScript, it cannot operate user-specific sessions, and does not mimic user interactions in a browser

so cannot load dynamic content, meaning that a lot of social media content is missed with Heritrix (Brunelle et al., 2015).

Instagram is also a more common social media platform to capture in Belgium than in the UK. Differences in the use of social media in both countries could also be a possible explanation for this discrepancy. The statistical evidence suggests that Facebook enjoys greater popularity in Belgium compared to the UK, whereas the reverse holds true for Twitter. According to data obtained from the Statcounter GlobalStats (2023) service — a web traffic analysis website based in Dublin, Ireland — the average percentage of Facebook users in Belgium between January 2020 and July 2023 was 66%, whereas in the UK the corresponding figure was 55%. Twitter users accounted for 7% in Belgium, whereas it was 21% in the UK. Another significant difference is that the websites of government agencies are collected by very few Belgian institutions in contrast to the UK. The fact that the State Archives in Belgium do not yet systematically archive the websites of government agencies undoubtedly contributes to this fact.

The selection criteria for the COVID-19 collections also differed between countries. In Belgium, no respondents indicated that the top-level domain mentioned in the web content was a determining criterion, whereas in the UK it was the second most important selection criterion. This can be partly explained by the fact that there is no top-level domain collecting being done in Belgium as KBR, the Royal Library of Belgium, is waiting for a change in the legal deposit legislation that would make it possible to collect on this scale.

There was also an important difference in the temporality of collecting. In June–September 2020 45% of Belgian respondents had already stopped collecting, compared to 25% in the UK and 28% in other European countries. One possible explanation for the difference between the UK and Belgium is that in the UK web archiving is a more mature field than in Belgium, which could mean that funding for web archiving has become sustainable, whereas in Belgium funding could still rely more on short-term project-based funding.

Regarding the tools for capture, complex tools such as Heritrix were more common in the UK than in Belgium or other European countries. The difference in the use of tools may be explained by the fact that in Belgium 69% of respondents indicated they worked for a small organisation that lacked the necessary resources to install and maintain tools such as Heritrix that are quite complex (Jackson, 2019).

It is important to note that the choice of web archiving tools is a complicated decision influenced by multiple factors, and organisations may employ a combination of tools to meet their specific requirements. Large-scale initiatives, such as national libraries or dedicated web archiving institutions, require robust and scalable tools such as Heritrix which is capable of capturing a huge volume of web content. On the other hand, smaller organisations or research projects may opt for simpler tools that suit their more focused archiving needs and do not require extensive knowledge of web technologies, programming, or server management, while offering more user-friendly interfaces, such as Webrecorder. Legal mandates can also play a role in the selection of tools. Smaller organisations may be archiving web content without specific legal mandates such as national legal deposit laws enjoyed by large, national libraries or archives. Tools that operate through a platform's API, such as Instaloader for Instagram, operate within the platform's terms of service and ensure that users are archiving in compliance with applicable laws and regulations.

A surprising difference that was noted relates to the long-term preservation practices. It was expected that, given the relatively long tradition of web archiving in the UK, more respondents from the UK would indicate that they use a long-term preservation system or

Surveying the landscape of COVID-19 web collections

software (in-house or external). However, 30% of the UK respondents indicated that they do so, compared to 38% of Belgian respondents and 60% in other European countries.

Regarding access, there were a few interesting differences between the three groups of respondents. Providing access using derived datasets or sharing seed lists with the public is not yet done in Belgium and the UK but might be worth looking into. The analysis by Geeraert et al. (2024) showed that the insight into the scope of a collection that can be gained from a seed list is limited in the context of transnational research, but it can still be an entry point into a collection since web archive metadata is sometimes only available onsite. Another lesson that can be learnt from the respondents in other European countries is providing data-level access or offsite access to the collections. It is also interesting to note that ‘Playback software’ was only the fourth most common answer in Belgium compared to the second most common in the UK and the most common in other European countries. This might hint at the fact that archiving web content is a more mature field in the UK and other European countries than Belgium, and may also explain why Belgian institutions need to rely more on metadata to provide access to the collections than UK institutions.

The second research question was related to the impact of differences in the maturity of web archiving programs. Web archiving is a field that has seen varying degrees of establishment and development across different countries. From the three groups of respondents, the UK, Belgium, and other European countries, the level of maturity of web archiving initiatives in only the former two countries could be assessed in an approximative manner. If maturity is measured by metrics including the scale on which web archiving is done in the country, how long web archiving programmes have been running, and the evolution of web archiving practices, it is clear that web archiving at GLAM institutions is more mature in the UK than in Belgium.

Taking the example of the national libraries: the UK Web Archive started creating selective collections in 2005 and has been doing a top-level domain crawl since 2013. The British Library and other legal deposit libraries including the National Library of Scotland and the National Library of Wales collaborate to ensure comprehensive coverage and preservation of web content. Other UK organisations such as the National Archives and National Records Scotland contribute significantly to web archiving efforts. In Belgium, in contrast, the first pilot project related to web archiving at KBR (the Royal Library of Belgium) only started in 2017 and a sustainable web archiving program still has to be set up at the national level. Other cultural heritage institutions in Belgium have established web archiving programmes before 2017 but these only concern selective collections. In contrast to the UK (as also demonstrated by Simone Natale (2024)), the practice of web archiving has also not yet spread to galleries and museums in Belgium (Messens & Geeraert, 2022). While the UK has a well-established and recognised web archiving infrastructure, Belgium does not have the same level of infrastructure and national coordination as the UK because the web archiving landscape is more distributed and small-scale compared to the centralised efforts in the UK.

This difference in maturity is reflected in several aspects of the web archiving lifecycle. The tools for capture and replay that UK respondents indicated using are overall more technologically advanced than the tools Belgian respondents indicated. Furthermore, as mentioned above, the fact that Belgian initiatives stopped collecting earlier than their counterparts in the UK may also be related to differences in funding opportunities and available infrastructure, underlining the necessity for further promoting web archiving as an activity in the Belgian

GLAM sector. This difference in maturity is also reflected in the perceived challenges. Ten percent of Belgian respondents see ‘securing additional technical support’ as a challenge compared to only 3% in the UK.

These analyses and discussions point to many future avenues for research. It may be beneficial to revisit those institutions that explicitly stated their intention to halt collecting based on the World Health Organization’s (WHO’s) decisions to see if and how the declaration by the WHO on the 5th May 2023 that COVID-19 was over as a global health emergency influenced the collecting activities of European GLAM institutions. By examining whether such declarations were eventually made and comparing them against the survey responses, valuable insights could be gained to enhance the understanding of the relationship between collecting practices and the perceived trajectory of the pandemic.

The analysis could also be broadened by including initiatives outside of Europe to make a global comparison although it needs to be noted that web archiving is still mainly an activity in the Global North which also introduces an important bias. Within Europe, more information could be gathered regarding funding opportunities or political priorities that influenced the creation and/or maintenance of COVID-19 collections. Further qualitative research could also be done to make recommendations to set up rapid response collection strategies, gain further insight into how the main challenges identified during COVID differed from before or after the pandemic, and see how the research use of the collections has evolved over time.

Conclusion

This chapter explores how cultural heritage institutions have been collecting and documenting traces of the COVID-19 crisis on the Web by analysing the results of a survey among European GLAM institutions that ran between July and September 2022. The survey data shed light on the motivations behind setting up the collections, collaborations between institutions, selection scopes, technical aspects, and perceived challenges. The main focus of the chapter was on two sub-questions: (1) what are the main similarities and differences between European countries in creating COVID-19 web archiving collections and (2) what, if any, is the impact of differences in the maturity of web archiving programmes? It is important to acknowledge the presence of potential biases in the data due to a couple of factors related to the survey distribution. A deeper analysis of the results concentrated mainly on Belgium and the UK due to the scarcity of responses from other European countries.

Several key points emerged from the analysis of the data. It was observed that there are notable differences in approaches between Belgium and the UK regarding web archiving. Web archiving initiatives in the UK employ more technologically advanced tools as compared to Belgium. Playback software is also less commonly used for access in Belgium than in the UK. Moreover, web archiving in galleries or museums is more common in the UK and other European countries than in Belgium. In addition, in both the UK and other European countries most respondents indicate that they collect on a national scale, compared to the local scale in Belgium. Regarding the end of collecting COVID-19 content, almost half of Belgian respondents had already stopped collecting by the time of the survey, compared to only a quarter of respondents in the UK. Marked differences were also noted in the types of social media that were collected. Facebook, for example, is the most common option in Belgium, whereas it is the least common in the UK. The results indicate that web archiving efforts and the scope of collections are to a large degree determined by local factors such

Surveying the landscape of COVID-19 web collections

as infrastructure and national legislation. While all these factors point to the fact that web archiving is a more mature field in the UK and other European countries in general than in Belgium, it is important to note that web archives in different stages of maturity can still complement each other, collaborate, and share expertise. To gain further insights, however, a more extensive study and another research approach, including interviews and focus groups, are needed.

While the survey data was intended to further understand COVID-19 collections of web-archived content, this also allowed us to identify some priority areas that may guide further work and research. For instance, the survey data provided insights into the scope of COVID-19 Special Collections. The outbreak of the infectious virus caused a global pandemic, necessitating collections that cover or can aggregate web content from various parts of the world. However, most collections are regional, local, and, in some cases, national in scope. The Content Development Group of the International Internet Preservation Consortium, in collaboration with Archive-It, established collaborations with 130 collecting institutions to form a collection with more than 80.000 archived websites to document some of the impact of the COVID-19 pandemic on a global scale. To better comprehend events on the web that extend beyond national borders, web archives must embrace international collaborations and devise methods to support transnational collecting and research initiatives.

The survey data also revealed that quality control is not consistently performed on the COVID-19 Special Collections and that most respondents relied on visual checks rather than automated processes. Coupled with the uncertainty surrounding the long-term preservation of most collections, this will impact the quality of these collections and how we remember online phenomena on the web, as exemplified during the pandemic. While not all institutions may have the financial means to conduct extensive quality control on their collections, implementing automated quality control processes could enhance basic delivery standards. The survey data also underscored the necessity for strategies, dedicated systems, and infrastructures for long-term preservation.

To conclude, a final priority area is access. While the mission of web archives is not only to preserve but also to provide access to primary sources that document and illuminate our past, it is not yet fully established how access can be provided. Currently, the COVID-19 Special Collections can mainly be accessed on-site, meaning in the library itself. For most institutions, it is not yet possible to make the collections accessible through derived datasets or by sharing seed lists with the public. This presents challenges for researchers who wish to study research questions that would require combining datasets from various web archives and other heritage institutions. It also presents obstacles for the use of web archive datasets within teaching in universities and other educational institutions. A significant challenge in providing access to collections is related to questions around GDPR and privacy concerns, which could again be an area for further collaboration in the community.

This chapter refrains from undertaking a comprehensive assessment of web archives within their contexts. It is hoped that the survey may be of use in providing a deeper level of understanding about archiving special web collections within European archiving institutions. Potentially, the survey and the dissemination of the results will also be useful in motivating researchers to explore the vast web archive collections available throughout Europe, particularly those related to COVID-19. These diverse collections hold great potential for further research and analysis, and their accessibility can contribute to a deeper understanding of the pandemic's impact and response across different regions.

Acknowledgements

The authors extend their sincere appreciation to Susan Knight, Customer Insight Coordinator at the British Library, for her invaluable expertise and support in the setup, management, and configuration of the software that underpinned the survey conducted in the course of this research.

References

- Austin, A. (2023, June 5). Bridging the gap: Capturing UK trans health discourse in the Archive of Tomorrow [Conference presentation abstract]. *RESAW Conference 2023*, Marseille, France. <https://resaw2023.sciencesconf.org/434751> archived at Wayback Machine June 5, 2023, <https://web.archive.org/web/20230605132047/https://resaw2023.sciencesconf.org/434751>.
- Belga. (2021, March 17). *1 jaar corona — een chronologisch overzicht*. Medi-sfeer. www.medi-sfeer.be/nl/nieuws/1-jaar-corona-een-chronologisch-overzicht.html archived at Wayback Machine March 23, 2023, <https://web.archive.org/web/20230323024318/www.medi-sfeer.be/nl/nieuws/1-jaar-corona-een-chronologisch-overzicht.html>.
- Bingham, N., Geeraert, F., Nyvang, C., & de Wild, K. (2022). COVID-19 web collections survey. [Dataset]. <https://doi.org/10.23636/g5n0-pp30> archived at Wayback Machine February 22, 2024, https://web.archive.org/web/20240222091431/https://bl.iro.bl.uk/concern/generic_works/a6cbdb35-9897-4533-a420-f609fcacbfad.
- Brügger, N., & Schroeder, R. (Eds.). (2017). *The Web as History*. UCLPress.
- Brunelle, J. F., Weigle, M. C., & Nelson, M. L. (2015). *Archiving Deferred Representations Using a Two-Tiered Crawling Approach*. Old Dominion University Department of Computer Science. <https://arxiv.org/pdf/1508.02315.pdf> archived at Wayback Machine December 25, 2022, <https://web.archive.org/web/20221225194549/https://arxiv.org/pdf/1508.02315.pdf>.
- Burke, V., Jørgensen, D., & Jørgensen, F. A. (2020). Museums at home: Digital initiatives in response to COVID-19. *Norsk museumstidsskrift*, 6(2), 117–123. <https://doi.org/10.18261/issn.2464-2525-2020-02-05> archived at Wayback Machine February 22, 2024, <https://web.archive.org/web/20240222091611/www.idunn.no/doi/10.18261/issn.2464-2525-2020-02-05>.
- Cui, C., Pinfield, S., Cox, A., & Hopfgartner, F. (2023, June 6). Participatory Web Archiving: The path towards more inclusive web archives? [Conference presentation abstract]. *RESAW Conference 2023*, Marseille, France. <https://resaw2023.sciencesconf.org/433545> archived at Wayback Machine June 6, 2023 <https://web.archive.org/web/20230606152051/https://resaw2023.sciencesconf.org/433545>.
- Cultural Heritage Terminology Network. (2024). Cultural heritage terminology network. <https://culturalheritageterminology.co.uk/> archived at Wayback Machine February 2, 2024 <https://web.archive.org/web/20240202124336/https://culturalheritageterminology.co.uk/>.
- Geeraert, F., Bingham, N., Clavert, F., Strandgaard Jensen, H., & Winters, J. (2024). Oral histories and scalable reading: analysing born-digital collecting practices during the COVID-19 pandemic. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge Companion to Transnational Web Archive Studies*. Taylor & Francis (Forthcoming).
- Geeraert, F., de Wild, K., & Aasman, S. (2024). What can we learn from URLs? Understanding the scope of COVID-19 web archive collections for transnational analyses. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge Companion to Transnational Web Archive Studies*. Taylor & Francis (Forthcoming).
- Huc-Hepher, S. & Ma, X. (2024). The trouble with community: Constructing, deconstructing and reconstructing transnational “community micro-archives”. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge Companion to Transnational Web Archive Studies*. Taylor & Francis (Forthcoming).
- The Institute for Government. (2021, February 15). Timeline of UK coronavirus lockdowns, March 2020 to March 2021. www.instituteforgovernment.org.uk/sites/default/files/timeline-lockdown-web.pdf archived at Wayback Machine March 14, 2023, <https://web.archive.org/web/20230314233210/www.instituteforgovernment.org.uk/sites/default/files/timeline-lockdown-web.pdf>.

Surveying the landscape of COVID-19 web collections

- Jackson, A. (2019, February 19). A new release of Heritrix 3. <https://netpreserveblog.wordpress.com/tag/heritrix/> archived at Wayback Machine September 22, 2023, <https://web.archive.org/web/20230922112916/https://netpreserveblog.wordpress.com/tag/heritrix/>.
- The Legal Deposit Libraries (Non-Print Works) Regulations 2013*. (2013). www.legislation.gov.uk/ukxi/2013/777/contents/made archived at Wayback Machine August 23, 2023, <https://web.archive.org/web/20230803193758/www.legislation.gov.uk/ukxi/2013/777/contents/made>.
- Lemenager, T., Neissner, M., Koopmann, A., Reinhard, I., Georgiadou, E., Müller, A., Kiefer, F., & Hillemacher, T. (2021). COVID-19 lockdown restrictions and online media consumption in Germany. *International Journal of Environmental Research and Public Health*, 18(1). www.mdpi.com/1660-4601/18/1/14 archived at Wayback Machine May 21, 2023, <https://web.archive.org/web/20230521082824/www.mdpi.com/1660-4601/18/1/14>.
- meemoo. (2023). Best practices for social media archiving in Flanders and Brussels. <https://meemoo.be/nl/projecten/best-practices-voor-de-archivering-van-sociale-media-in-vlaanderen-en-brussel> archived at Wayback Machine January 25, 2024, <https://web.archive.org/web/20240503141733/https://meemoo.be/nl/projecten/best-practices-voor-de-archivering-van-sociale-media-in-vlaanderen-en-brussel>
- Messens, F. (2023). Belgian social media archiving initiatives mapped. *In Monte Artium*, 15. <https://journals.openedition.org/ima/322>, archived at Wayback Machine January 25, 2024, <https://web.archive.org/web/20240125170508/https://journals.openedition.org/ima/322>.
- Messens, F., & Geeraert, F. (2022). Country report: Web and social media archiving in Belgium. <https://orfeo.belnet.be/handle/internal/9980> archived at Wayback Machine June 8, 2023, <https://web.archive.org/web/20230608065000/https://orfeo.belnet.be/handle/internal/9980>.
- Natale, S. (2024). The many lives of WeChat: Curating histories of the web in museum environments. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge Companion to Transnational Web Archive Studies*. Taylor & Francis (Forthcoming).
- Ogden, J., & Maemura, E. (2021). ‘Go fish’: Conceptualising the challenges of engaging national web archives for digital research. *International Journal of Digital Humanities*, 2, 43–63. <https://doi.org/10.1007/s42803-021-00032-5>.
- Statcounter GlobalStats. (2023). Statcounter GlobalStats. <https://gs.statcounter.com/> archived at Wayback Machine February 2, 2024, <https://web.archive.org/web/20240202124657/https://gs.statcounter.com/>.
- Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, 1, 85–111. <https://doi.org/10.1007/s42803-019-00007-7>.
- Webster, P. (2019). Existing web archives. In N. Brügger & I. Milligan (Eds.), *The SAGE Handbook of Web History* (pp. 30–41). Sage.
- World Health Organisation. (2023a). Timeline: WHO’s COVID-19 response. www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline archived at Wayback Machine February 2, 2024, <https://web.archive.org/web/20240202124827/www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline>.
- World Health Organisation. (2023b, May 5). WHO Director-General’s opening remarks at the media briefing — 5 May 2023. www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing---5-may-2023 archived at Wayback Machine February 2, 2024, <https://web.archive.org/web/20240202125658/www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing---5-may-2023>.

10

WHAT CAN WE LEARN FROM URLS?

Understanding the scope of COVID-19 web archive collections for transnational analyses

Friedel Geeraert, Karin de Wild, and Susan Aasman

Introduction

As Astrid Erll observed in 2020, in the midst of the COVID-19 pandemic: “It is the first worldwide digitally witnessed pandemic: a test case for the making of global memory in the new media ecology” (Erll, 2020, p. 867). It is not hard to read this as a call to both researchers and archivists to think about which “mediated experiences, beliefs and narratives will make it into the dominant memory discourses” (ibidem). This chapter discusses how to make use of a collaborative effort of the International Internet Preservation Consortium (IIPC) and the national libraries of Luxembourg, Hungary, France, the UK, Denmark, and The Netherlands to document the presence of the pandemic on the web. It aims to explore the seed lists and accompanying metadata provided by the archives in order to give insights into the scope of the COVID-19 special collections (Aasman et al., 2021).

Since the pandemic provided an opportunity to analyse the way in which an unforeseen event is taken into account by European web archives in the face of an emergency, the aim was to illustrate the usefulness of sharing metadata about COVID-19 web archive collections. This chapter will demonstrate various use cases from the perspective of two fictional profiles: a researcher and a web archivist. Each of the profiles focuses on a use case related to their day-to-day work.

A number of assumptions were made regarding the profiles. Firstly, it was assumed that access conditions to the web archives in 2030 would be the same as in 2020. Most national library web archive collections are only available in reading rooms due to national legal deposit laws. Of the sample datasets included in this chapter, the IIPC Collaborative Collection is freely accessible via Archive-It and Netarkivet’s collections are accessible online to researchers or PhD students affiliated with a Danish research institution (Danish Royal Library, 2023). The other collections are only accessible on-site. Secondly, sharing the metadata that was obtained from the different institutions for this chapter was facilitated by membership of the WARCnet network. It was assumed that the fictional profiles would succeed in obtaining similar metadata from the institutions, even without such connections.

What can we learn from URLs?

Another assumption was that both profiles wished to avoid having to travel to different countries to consult the web archives and underlying WARC files and preferred doing a preliminary analysis based on metadata. Economic barriers can play an important role in how research is conducted. If the research budget does not allow for travel, potential research avenues may be left unexplored because of budgetary constraints. It was also assumed that both profiles had a similar approach to data aggregation and cleaning. The chapter, therefore, discusses the data aggregation and cleaning process, which entailed creating a test bed, undertaking a preliminary analysis of how a corpus could be created from the heterogeneous data, and providing feedback about and documenting the process.

The researcher is a postdoctoral student in social sciences at a European university wanting to study the collections in 2030. The use case this profile is interested in is to determine if and how the COVID-19 collections can be used to assess how governments communicated with their citizens about the COVID-19 pandemic and the response of the public to this communication. More specifically, the researcher wishes to compile a transnational corpus of websites of national institutions such as the National Health Service (NHS) in the UK and social media posts by private citizens that refer to these websites to gain insight into the mood during the pandemic. The researcher wishes to assess if such a corpus can be compiled by requesting metadata.

The web archivist, on the other hand, is a staff member of a national library in Europe. This profile is primarily focused on utilising the COVID-19 collections to analyse and compare the curation approaches of her institution's COVID-19 collection with those of similar foreign institutions as of the end of 2020, following 8 months of pandemic-related data accumulation. The web archive which the web archivist helps to curate is relatively new. A comparison with the collections of more experienced institutions is deemed useful to detect potential blind spots or overlooked categories of websites in the COVID-19 collection. Additionally, she also provided input for the collaborative collection of the IIPC and is interested in determining the overlap between national collections on the one hand and the IIPC collection on the other.

Three areas of analysis of interest to both profiles are discussed in detail in this chapter: the geographical spread of top-level domains in the collections, the overlap between the different collections, and the presence of social media in the collections.

In addition, the chapter reflects on the underlying methodological and theoretical implications. Traditionally, historians and archivists are considered natural partners, sharing a strong engagement with historical sources. The emergence of web archives does not change this principle, but it does challenge existing knowledge about record-keeping procedures, (trans)national datasets, and research agendas. This chapter provides reflections on issues related to processes of heritagisation of unforeseen events through born-digital sources and “living archives”.

Studying the usefulness of sharing metadata to gain insight into the scope of COVID-19 collections from the point of view of the two profiles provided a reality check. It allowed us to gain an overview of national and transnational corpora available in web archives and provided an opportunity to combine metadata, derived data, and seed lists to achieve a preliminary overview of the homogeneity as well as the diversity of the datasets and to reflect on the choices made by web archiving institutions. The chapter offers valuable insights for researchers seeking collections, broadening their perspectives with the discovery of datasets from other countries. It also serves as a resource for web archivists interested in comparing various methods of scoping emergency collections.

The heritagisation of COVID-19

By the time our fictional researcher requests the archived COVID-19 collections, it will have been 10 years since the collection was created. By then the global pandemic has become part of our collective memory but a process of heritagisation has also started. The COVID-19 collections show in almost real time how born-digital materials on the live web undergo a process of heritagisation. Musiani and Schafer (2016) pointed out that such a transformation provides a moment to reflect on crucial methodological issues such as the question of how to deal with the available metadata, since it ensures accessibility, interoperability, and contextual information. In addition, the authors stress the need for asking epistemological questions.

The term “heritagisation” comes from the field of archaeology and refers to a more “processual understanding of heritage making” (Thouki, 2022, p. 1036). It emphasises the process of how and why historical artefacts become heritage “filtered through collection, institutionalisation, commodification, and protection”. In our case, the heritagisation of the live web adds a new dimension, as it deals no longer with collecting materials long after an event has taken place. Because of the vulnerability of online materials, scholars and archivists need to discuss archival practices of the present and question how ongoing crises can or should be represented (Adams & Kopelman, 2022). This might not be so self-evident as Astrid Erll observed since it might be difficult to understand how an event emerges, “if people don’t know what the event is, where it started and how it evolved?” (Erll, 2020, p. 867).

Another level of complexity is related to how these potential web archival materials need to be understood within the broader memory culture. According to Hoskins (2011) we are dealing with a new media ecology wherein memory-making is shaped in new ways. Hoskins describes this new digital memory culture as a “connective memory”, thereby emphasising the element of real-time and instantaneous memory-making on a global scale. This also means that what previously counted as “boundaries of space, time and social experience” (Keightley & Schlesinger, 2014, p. 747), now become less dominant.

The IIPC Collaborative Collection demonstrates the deeper impact of the heritagisation process of an unforeseen event that took place on a global scale. These collections offer a unique opportunity for transnational research while at the same time, there need to be insights about specificities, such as: national policies with regard to collecting web materials (for instance, archiving top-level domains such as .uk but not .com), the technical state of the art of archiving the dynamic web, institutional practices with regard to metadata illustrating that the types of data (news sites or social media) and contextual information collected might influence transnational comparisons. Recently, important work has been done in gaining a better understanding of archival strategies for the contemporary collecting of web materials (Webster, 2017; Maemura et al., 2018; Vlassenroot et al., 2019, Bingham & Byrne, 2021; Maemura, 2022). These studies share the observation that web archival practices can only be understood in their “social and technical context that shapes those decisions and the ways in which these individual decisions interact” (Maemura et al., 2018). Bingham and Byrne (2021) analysed, in the British context, how processes of heritagisation and archivalisation, especially when it is done at scale, “disproportionately represents the more official voice” (p. 2). According to Bingham and Byrne, technical and legal affordances and restrictions determine how heritage is both captured and created.

Data collection

This study is based on a transnational corpus for which the data was provided by a selection of national Web archives. These organisations are primarily located in Europe, including Luxembourg (Bibliothèque nationale du Luxembourg), the United Kingdom (UK Web Archive at the British Library), Denmark (Netarkivet at the Royal Danish Library), France (Institut National de l'Audiovisuel and Bibliothèque nationale de France), the Netherlands (KB National Library), and Hungary (National Széchényi Library). Additionally, the IIPC Content Development Group provided data. This last collection contains seeds nominated by their members and the public and therefore includes data from a wide variety of partners around the world (including partners in Asia, Canada, and the United States).¹

All the datasets were requested and collected in November and December 2020. Due to the urgency caused by the COVID-19 outbreak, the archiving of web content had already begun in mid-February 2020. During these early stages, the primary focus was on collecting and preserving the data. The labour involved in acquiring web content was understandably prioritised and accessing the COVID-19 special collections for research engagement remained challenging. The data for this study was gathered through relationships fostered by long-term partnerships with institutions, established in part through the WARCnet research network, the European research network for studying web domains and events. Each institution was approached with a request to provide data about their COVID-19 special collections. The format and data fields were not predefined, and all data (metadata, derived data, remote access) were welcomed. The goal was to gain access to as much data as possible to obtain initial insights into the scope of these collections and the curation decisions. The result was heterogeneous datasets in various formats (such as Comma Separated Values [CSV] and JavaScript Object Notation [JSON]) and content (including tweets, seeds, and metadata). This approach was chosen because the copyright legislation in most European countries prevents national libraries from opening up their collections outside of their physical reading rooms.

The heterogeneity presented several challenges in cleansing the dataset to ensure the accuracy, completeness, and consistency of the data. The first difficulty was about maintaining consistency. Each of the received datasets had its own variables, which were not mutually exclusive. Ideally, complementary documentation would be available providing a list of variables (or labels) with accurate definitions. Problems can arise when attempting to combine data from various archiving institutions. How can one determine if data labels from multiple sources can be matched up? Even when the labels of the variables are the same, variables from different sources may differ (also known as the entity identification problem) (Lim et al., 1996, 2). Additionally, each data collection contained different variables (including “URL”, “domain URL”, “Page title”, “Top Level Domain”, “Curator”, “Crawl date”, “Crawl frequency”, “Actors”, “Themes” and “Keywords”). Web archivists were consulted to gather additional information regarding the provided variables, and it was found that only the variable “URL” and “domain URL” appeared consistently in each of the datasets. A second challenge involved ensuring the accuracy and completeness of the data entries. Ideally, when tracing errors, one would be able to refer back to the original source. However, due to legal reasons, access to the raw WARC files was not possible. In addition, other barriers to making WARC files available exist, such as identifying the relevant files in a large collection or technical difficulties encountered in sharing the files if an operational procedure for file-sharing with researchers does not exist. Instead, derived datasets, seed lists, and metadata

were provided. Although inaccuracies and inconsistencies were identified and cleansed, it was not possible to consult the raw data (from the WARCs) to evaluate the accuracy of the data or determine the percentage of data entry errors to improve data quality.

A third challenge was the absence of provenance data describing the origin of the data. Several scholars have indicated the crucial need for provenance (Ben-David & Amran, 2018; Maemura et al., 2018; Acker & Chalet, 2020), which is defined by Maemura et al. (2018, p. 3) as everything “[that] users need to know about how a collection was made as they use, analyse, and make inferences from these aggregations”. Although some information was provided, such as the institution that collected the data and the date when the web page was crawled, essential details about the collection efforts, curatorial decisions, and other contextual information that could help in further understanding the special collections were missing. Furthermore, it was unclear if the data in the datasets provided to the research team was extracted from the WARCs, as no documentation was available regarding the data extraction process, which might have introduced changes to the data along the way.

Despite these challenges, gaining a better understanding of data collection in the context of web archiving, even in the early stages of the collection process, is highly beneficial. This is particularly true for significant special collections, such as the capturing of web content during the COVID-19 pandemic. Most collections are curated by subject specialists who determine the seeds, decide on crawl frequencies, and perform quality assurance (Milligan et al., 2016). Analysing the scope of these collections can support curatorial decisions, assess the completeness of the collection, and identify gaps that may impact future use. In the long term, these analyses can provide valuable provenance information, including the context, scope, and creation of a special collection. The results can also facilitate more efficient collaboration between archiving institutions by highlighting how the efforts made by various institutions can complement each other. Furthermore, early research engagement can also help evaluate the suitability of the collected material and provide potential research corpora for future studies. Complementing these datasets with provenance information can, in turn, stimulate the reuse of datasets derived from the WARCs and support future research.

Data analysis

As a next step, we will discuss how the fictional profiles may wish to use the provided data to gain further insights into the scope of the COVID-19 special collections. The analysis focuses on three main areas that both profiles might deem interesting to examine further. The first section will focus on the international scope of the data collection. This will provide greater transparency regarding the countries (and national Web domains) represented within the data collections, as well as highlighting any gaps or missing data. The second section of the analysis concentrates on identifying potential overlaps between the different collections. Lastly, the analysis focuses on social media. In comparison to archiving web pages, capturing and storing social media content is still a developing field, so it is important to provide insights into how institutions were able to capture this type of data (Vlassenroot et al., 2019).

Geographical spread

The first analysis focuses on the transnational scope of the collection. The researcher aims to analyse the size of the COVID-19 collections and identify those with the most international scope. This information is crucial in determining which collections to prioritise for

What can we learn from URLs?

addressing the research question. The web archivist is interested in this analysis to see how large the other collections are and how strictly other web archives adhere to the top-level domain(s) (TLD) of their respective countries.

Other than the IIPC collection, the data is mostly provided by national web archives from five European countries. Of these five countries most unique domain names are provided by the Bibliothèque nationale de France (2,603), while smaller collections are provided by the national libraries of Luxembourg (538), the Netherlands (546), and Hungary (113). This means that not all countries are equally represented within the corpus.

The primary focus of national Web archives is on the national domain (Brügger & Laursen, 2019). However, to determine the transnational scope more accurately, additional details about each collection are needed, including whether some collections also include content beyond the national domain. Figure 10.1 reveals the (inter)national focus of each COVID-19 Special Collection included in the dataset. For this analysis, the TLD names are extracted from the unique domain names. The TLD represents the last segment of the domain name, including the final dot and the letters immediately following it. For example, in the domain name *www.nhs.uk*, the TLD is *.uk*. The country associated with a domain name is determined by the country code top-level domain (ccTLD) of the URL, which is an Internet top-level domain generally used or reserved for a specific country. However, it should be acknowledged that not all websites have a ccTLD (e.g., *.gov*) and that in some countries TLDs exist for specific regions of the country such as *.re* for Ile de Réunion in France, or *.frl* for the Frisian part of the Netherlands.

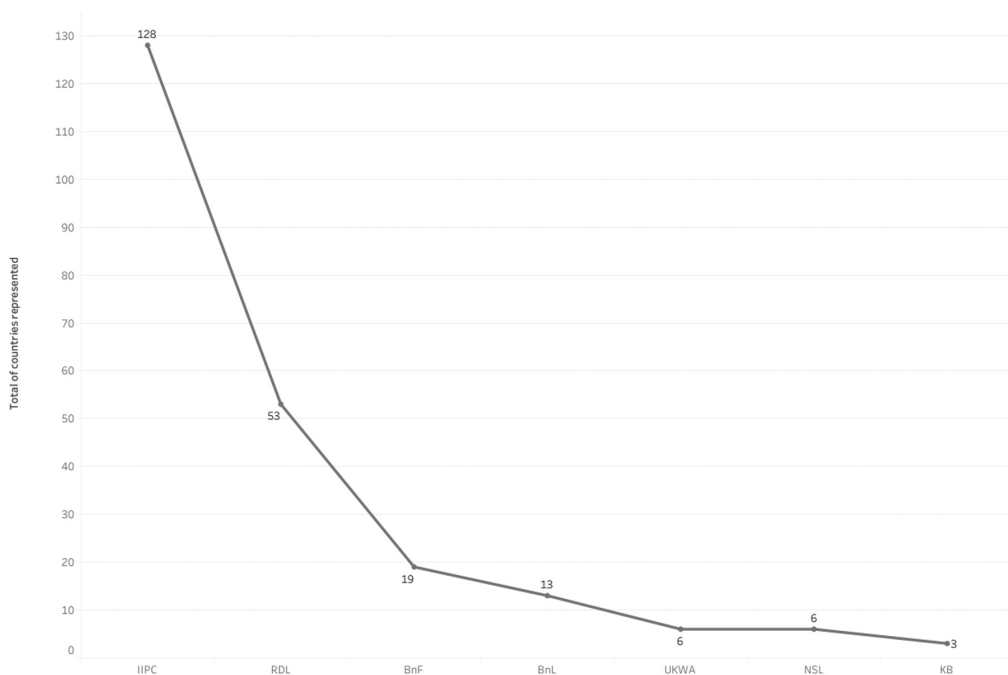
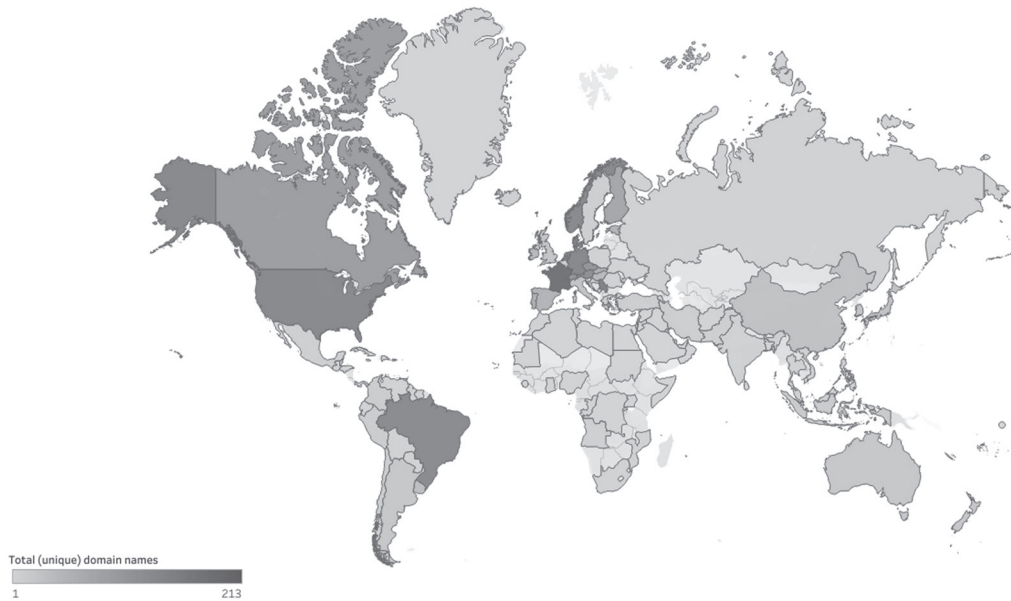


Figure 10.1 (Inter)national scope of COVID-19 special collections.

The figure presents the total of countries represented in each web archive.



Map 10.1 (Inter)national scope of the COVID-19 Special Collection of the IIPC in collaboration with Archive-It.

The figure presents the total of unique domain names per country.

Except for the IIPC, all the provided data was sourced from national web archives. However, not all web archives exclusively collect within their own national domain. As shown in Map 10.1, the IIPC collection represents the most countries (128), while the Dutch collection represents the least (3). The Dutch web archive exclusively collects within the Dutch domain (.nl). This is likely because the archive follows a selective approach, where each seed is manually selected, and the crawlers are configured to capture web pages only within the seed domain name.

Figure 10.2 shows that, while the scope of web archives may vary, most have a national focus. The exceptions to this are Denmark (56%) and, to a lesser extent, Hungary (88%). These differences can be explained by the fact that the national web is not only defined by the legal deposit legislations of the respective countries but also by the way in which the legislation is implemented by the institutions themselves. There are of course also countries without legal deposit legislation as is the case in The Netherlands. In the Danish example, the legislation states that: “Material published in electronic communication networks is considered Danish when: (1) it is published from Internet domains and the like specifically allocated to Denmark, or (2) it is published from other Internet domains and the like and is aimed at an audience in Denmark”, which can be considered as a broad definition (Retsinformation.dk, 2004). In comparison, the way in which the national web is defined by the national library in Luxembourg can also be considered rather broad: “1) all websites in the “.lu” Top Level Domain (TLD), 2) websites published in Luxembourg, 3) websites hosted on servers in Luxembourg and 4) websites published by non-resident Luxembourg citizens or websites in relation with Luxembourg”, but the scope remains much more within the ccTLD than is the case in Denmark (Bibliothèque nationale du Luxembourg, 2023).

What can we learn from URLs?

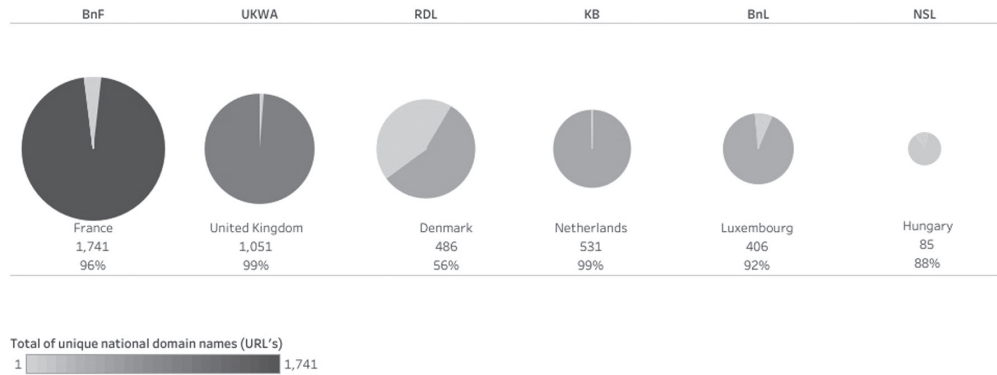


Figure 10.2 Total number of websites in the COVID-19 special collections of national web archives and the percentage of those websites having the same ccTLD as the archive's coverage area.

Insights into the size and transnational scope of collections are valuable to researchers. Our fictitious researcher is primarily interested in studying a dataset with a transnational scope, and would therefore show a preference for analysing the Danish and IIPC collections. However, working with more internationally oriented datasets can also pose challenges due to the multilingual nature of the data and the potential for bias and underrepresentation of specific regions and groups. Analysing such data also requires cultural expertise to accurately interpret the nuances and expressions within the corpus. While the Hungarian collection is relatively international in scope, it is also relatively small in size. This limited size may pose several constraints for conducting comprehensive international analyses. Nonetheless, delving deeper into the data can provide insights into the countries represented within the corpus. Given that this collection may include websites from Eastern Europe more prominently than other web archival datasets, it could be of particular interest to our researcher.

The web archivist, on the other hand, is surprised at how much approaches differ between institutions both in the size of the collections and their international scope. The evidence allows the web archivist to determine that the COVID-19 collection they curate is medium in size (based on the number of unique domain names at least) and has an international scope that is similar to the collection of Luxembourg.

Overlap between the different collections

Another aspect of these collections that was analysed was the overlap between different collections. The researcher is interested in this question to determine whether it would be sufficient to solely focus on the IIPC collection to answer his research question, as it would have the advantage that travel is not required to gain access to the data as the collection can be accessed remotely through the Archive-It web archiving service. In addition, the researcher wishes to see if there is any overlap between national collections. The web archivist, on the other hand, wants to see if other institutions mostly limit nominations they submit for the IIPC collaborative collection to content that is not present in their own collections, or if there is significant overlap between the national and collaborative collections. The web archivist also wants to determine whether there is any content included in more than one national collection.

Table 10.1 Overlap between the IIPC collaborative collection and national collections (based on the number of URLs)

	<i>Number of URLs</i>		
	<i>National collection</i>	<i>Overlap between IIPC and national collections</i>	<i>% of overlap between IIPC and national collection based on number of URLs</i>
Bibliothèque nationale du Luxembourg	27,787	18	0.06%
Netarkivet	17,180	271	1.58%
Bibliothèque nationale de France	4,598	240	5.22%
UK Web Archive	2,672	6	0.22%
National Library of the Netherlands	603	4	0.66%
National Széchényi Library	128	4	3.13%

Of the institutions that shared the seed lists of their COVID-19 collections, five also contributed to the IIPC collaborative COVID-19 collection. The analysis concerned both URLs and domain names. URLs refer to web pages whereas the domain is the name of a website. The occurrence of URLs and domain names in relation to the different COVID-19 collections was mapped in Table 10.1. All URLs that overlapped were found between the IIPC Collaborative Collection and the collections from the contributing institutions. There was only one exception: a news item published by the UK government about a travel ban following the outbreak of COVID-19 in Denmark that was included in the collections of both the UK Web Archive and Netarkivet (Department for Transport, 2020). In absolute terms, most overlap existed between on the one hand Netarkivet (271 URLs) and the web archive of the Bibliothèque nationale de France (240 URLs) and the IIPC collaborative collection on the other. However, in relative terms, most overlap existed between the collections of the Bibliothèque nationale de France (5.22%) and the National Széchényi Library (3.13%) on the one hand and the IIPC collection on the other.

This analysis reveals that, in general, the overlap between the IIPC collection and the national collections is limited. As a result, there exists substantial potential for the various collections to enhance and enrich each other. The web archivist notes that the percentage of overlap between their collection and the IIPC collection is low based on this analysis and therefore decides to add more seeds to the collaborative IIPC collection.

The same comparison between collections was done for domain names. Both the researcher and the web archivist expected that, as domain names are less specific than URLs, more overlap between national collections would be detected. The analysis shows that this is indeed the case. Table 10.2 shows domain names that are included in three or more collections. The most popular domain names relate to social media platforms, news media, and to a lesser extent, government information. As the analysis of the geographical spread already led us to expect, Netarkivet in Denmark shares the most unique domain names with the other national collections. The National Library of The Netherlands' collection, on the other hand, does not contain any domain names that overlap with other national collections, meaning that they have a very focused collection scope, which the geographical spread also showed.

When comparing the domain names that only occur in two collections, the majority of the overlap is between the IIPC and individual national collections. However, there are a

What can we learn from URLs?

Table 10.2 Overlap between the IIPC collaborative collection and national collections (based on domain names)

<i>Count of domain name</i>	<i>Netarkivet</i>	<i>Bibliothèque nationale de France</i>	<i>National Széchényi Library</i>	<i>IIPC</i>	<i>Bibliothèque nationale du Luxembourg</i>	<i>UK Web Archive</i>
facebook.com	333	360	5	26	62	3
twitter.com	2,271	948	1	37	203	528
youtube.com	428	109		166	28	1
docs.google.com	2			3	1	1
ec.europa.eu	4	1	1	9		
instagram.com	275	20	1			2
theconversation.com	1	2		3		4
thesun.co.uk	6			4	1	1
vimeo.com	37	1		1	2	
bbc.co.uk	10			4		7
bbc.com	17			88		5
bmj.com	2			1		1
change.org		1		2		6
courrierinternational.com	1	2		3		
dailymail.co.uk	16			3		2
de.wikipedia.org	1			2	1	
diplomatie.gouv.fr	2	2		3		
drive.google.com		1		5		1
ecdc.europa.eu	2			5	1	
en.wikipedia.org	5			16	3	
forbes.com	12			8	1	
fr.euronews.com	1			1	1	
fr.wikipedia.org		4		4	1	
gov.uk	2			5		143
independent.co.uk	6			5		1
irishtimes.com	6			3		1
latribune.fr	1	1		1		
lavoixdunord.fr	1	1		1		
lemonde.fr	4	14		5		
leparisien.fr	1	4		2		
lepetitjournal.com	2	4		1		
lexpress.fr	1	2		2		
liberation.fr	1	3		3		
medium.com		1		5		1
metro.co.uk	5			1		1
nouvelobs.com	1	3		3		
parismatch.com	1	2		2		
reddit.com	1,290			6		1
sicnoticias.pt	1			10	2	
soundcloud.com	393	8			2	
spiegel.de	3			2	1	
sueddeutsche.de	5			4	2	
telegraph.co.uk	6			2		4
theguardian.com	23			15		14
thejournal.ie	2			1		1
who.int	6	1		22		
zdf.de	1			1	1	

The Routledge Companion to Transnational Web Archive Studies

Table 10.3 Overlap between two national collections (based on domain name)

<i>Count of domain name</i>	<i>Netarkivet</i>	<i>Bibliothèque nationale de France</i>	<i>Bibliothèque nationale du Luxembourg</i>	<i>UK Web Archive</i>
amazon.co.uk	1			1
ar.wikipedia.org	2		1	
chathamhouse.org	1			1
express.co.uk	2			1
ft.com	7			3
itv.com	1			4
lefigaro.fr		4	3	
mirror.co.uk	1			1
n-tv.de	2		2	
rtbf.be	1	1		
sites.google.com		2	1	
standard.co.uk	2			1
theartnewspaper.com	1			1
thetimes.co.uk	1			2
uk.reuters.com	5			1

few exceptions, included in Table 10.3. Most of the domains that appear in two national collections are news media publications. It is interesting to note that most of these domains are shared between Netarkivet in Denmark and the UK Web Archive, totalling 10 overlapping domain names.

These analyses allow the researcher to see that there is relatively little overlap between national collections, and because almost no government websites are included in this overlap, examining the overlapping content further to answer his research question is ruled out. For the web archivist, it is reassuring to see that the overlap existing between the COVID-19 collection of her institution and other national libraries can be considered reasonable.

Presence of social media

Other than the geographical spread and overlap between the different collections, another element that was mapped was the presence of social media in different collections. For the researcher, this is of great importance as the reactions of citizens to the information published on the websites of the NHS and other national counterparts are a key part of his desired research corpus. For the web archivist, this analysis allows her to determine how much social media is present in the other collections and to compare this with the ratio of social media content and websites in her own COVID-19 collection. This analysis was done based on the count of different domains related to social media platforms. The National Library of the Netherlands does not appear in the table as none of the social media platforms included in Table 10.4 are included in their collection.

Table 10.4 shows that the inclusion of social media content differs widely amongst institutions. Netarkivet in Denmark is the most prolific platform both in terms of the range of social media platforms included in their COVID-19 collections and the number of domains related to social media content. Their collection holds data from Twitter (2,271

What can we learn from URLs?

Table 10.4 Overview of the presence of social media in the different COVID-19 collections (based on count of domain names)

<i>Count of domain name</i>	<i>Netarkivet</i>	<i>Bibliothèque nationale de France</i>	<i>National Széchényi Library</i>	<i>IIPC</i>	<i>Bibliothèque nationale du Luxembourg</i>	<i>UK Web Archive</i>
twitter.com	2,271	948	1	37	203	528
facebook.com	333	360	5	26	62	3
youtube.com	428	109	0	166	28	1
instagram.com	275	20	1	0	0	2
reddit.com	1,290	0	0	6	0	1
old.reddit.com	1,118	0	0	1	0	0
dailymotion.com	498	0	0	0	0	0
tiktok.com	377	0	0	0	0	0

domain names), Facebook (333 domain names), YouTube (428 domain names), Instagram (275 domain names), Reddit (1,290 domain names), Dailymotion (498 domain names), and TikTok (377 domain names). The IIPC holds data from five different social media platforms, while (sometimes very) limited social media content stemming from four social media platforms can be found in the COVID collections of the Bibliothèque nationale de France and the UK Web Archive. On the other hand, the COVID-19 collection of the National Széchényi Library holds very limited Facebook, Twitter, and Instagram data. In terms of the number of domains included in the archives, Twitter is by far the most popular social media platform included across all collections, followed by reddit.com, Facebook, YouTube, Dailymotion, TikTok, and Instagram.

This analysis allows the researcher to see which web archive collections contain sufficient social media content to potentially warrant a visit to their web archive. The researcher is considering going to France, Denmark, the UK, and Luxembourg. The IIPC collection was also deemed interesting, but when searching through the collection on Archive-It, it quickly became clear that only the landing pages of the social media platforms were included, and little user-generated content. For the web archivist, it is interesting to see the diversity in social media content contained in other collections, especially in the Danish collection. The web archivist decided to schedule a call with a representative of Netarkivet to obtain more information about the tools they use to capture the different types of content and to see if enough resources are available at her library to implement a similar approach for including social media content in emergency collections as in Denmark.

Discussion

The aim of the chapter was to determine the usefulness of studying the URLs included in the seed lists of COVID-19 special collections at a number of European national libraries and the IIPC from the point of view of two profiles: a researcher and a web archivist. The analyses allowed the researcher to see that:

- 1 The collections of the IIPC and Netarkivet have the most international focus.
- 2 There is relatively little overlap between the IIPC collections and the respective national collections, totalling no more than 5.3% at the most. Furthermore, there is minimal

overlap between national collections, and the existing overlaps do not encompass the government websites of interest to the researcher.

- 3 The collections of Bibliothèque nationale de France, UK Web Archive, Bibliothèque nationale du Luxembourg, and Netarkivet are particularly interesting in terms of further investigating social media data. In contrast, the IIPC collaborative collection, despite its international scope, does not contain social media data and therefore appears less relevant than initially anticipated.

The researcher will need to decide whether he has enough time and budget to travel to four countries to look at the respective COVID-19 collections or if he wants to reorient his research question.

For the web archivist, on the other hand, the analyses showed that:

- 1 The collection she helped curate is of average size and shares a similar international scope to the collection of the Bibliothèque nationale du Luxembourg.
- 2 Her institution may need to add more seeds to the IIPC collaborative collection to align with the approach of other European national libraries. Moreover, the overlap between her institution's collection and other national collections can be considered reasonable.
- 3 Netarkivet's collection holds a wide range of social media content. The web archivist is interested in obtaining more information about the underlying infrastructure to assess the feasibility of implementing a similar approach for her own web archive.

The process has also revealed to the web archivist that offering seed lists with URLs as a means to access a web archive, which is only available in the reading rooms of institutions, yields very limited insights. She therefore decides to convince her management of the need to invest in the creation of as many indexes and derivative files as possible to allow researchers to obtain as much knowledge about their collection as possible before they set foot in the reading room.

Given the considerable investment of time that the compilation and cleaning of the dataset require, comparing collections based on URLs can be considered rather ineffective. For the web archivist, it would have been easier to consult documents such as the WARCnet papers containing interviews with web archivists or to reach out to colleagues at other national libraries. For the researcher, exploring the scope of a potential research corpus is typically something that would occur before submitting a research proposal, meaning that theoretically, he has no funding to undertake the labour-intensive task of compiling a cleaned dataset. It can therefore be argued that making available this type of metadata is insufficient and that web archives would be better striving towards making as much metadata and derived data available as possible. It is also important for heritage institutions to put time and effort into the creation of documentation and metadata even when the priority is on collecting during an emergency. If not, important metadata or contextual information may not be available to researchers.

It is important for researchers to gain insight into collections even before they enter the reading rooms. This can be achieved by making use of indexing programs such as Solr. Offering dashboards or other types of derivative data would also be beneficial. It needs to be noted, however, that the creation of such data requires sufficient resources, which is not the case for many smaller institutions. Even for large institutions, Solr indexing is

What can we learn from URLs?

resource-intensive because of the larger scale of the collections. Prioritising which elements to index may therefore be necessary. However, even in this situation, larger and better-funded institutions can help smaller ones with Solr indexation for example. In addition, the very diverse nature of metadata that was received at the beginning of the study and the fact that URLs were the only common metadata element between institutions, also underlines the need for further standardisation and harmonisation of metadata practices within the field of web archiving to facilitate transnational studies.

It should also be underlined that there are significant discrepancies between (1) the intention of the web archivists related to scoping the collection, (2) the URLs that end up in the seed lists, and (3) the actual content included in the web archive. This insight was obtained through privileged access to additional research findings of the WARCnet network, which would not have been easily accessible to the researcher and web archivist profiles. In the interviews with web archivists who were involved in curating COVID-19 web archive collections for example, it was mentioned by the representative of the National Széchényi Library that no social media was collected, but limited traces of social media data were nevertheless discovered in the collection (Németh & Geeraert, 2020). The same can be said for the IIPC Collaborative Collection where the collection description stated that social media content was to be excluded (Bingham & Geeraert, 2020a). At the UK Web Archive, only Twitter data was supposed to be included in the collections, but traces of other social media platforms were also discovered. This shows that there is a discrepancy between the scope of a collection at the start of the collection and the resulting seed list (Bingham & Geeraert, 2020b).

It is also important to note that an even bigger discrepancy is possible between what can be found in the COVID-19 collections of these institutions and the seed lists analysed in this study. Firstly, crawler settings have a significant influence on the collection. If for example, it was stipulated that for one seed, the crawler should also crawl the first level of the pages that the seed linked to, it is easy to see how more social media content can be found in the collections than the seed list on its own would indicate. It is crucial to keep in mind that a seed list does not describe the content of an entire collection in detail. Secondly, the fact that the seed lists included domains of social media platforms does not mean that social media data is present in the collections. In the case of the IIPC Collaborative Collection, for example, further analysis of the content of the collection clearly showed that only the login pages were archived, not the underlying social media data. A seed list can therefore only be seen as a starting point for further analysis; insight into collections solely by means of seed lists needs to be cautiously formulated.

The analyses have also stimulated a reflection on which research questions could be answered using only the URLs. It could be worthwhile to identify which URLs belonging to the national web sphere of the countries involved in this study only appear in the IIPC Collaborative Collection and not in any of the other national collections. It would also be interesting to analyse which second-level ccTLDs (for example, gov.uk or ac.uk) are archived. An analysis of which collections contain information published on the .eu domain is also a possibility. The analyses have also allowed for the exclusion of a number of research questions. Studying content that overlaps between national collections could be interesting for researchers for example, but the analyses clearly show that overlaps between national collections are minimal, thereby excluding a number of potentially interesting research questions such as: who are considered important actors in the COVID-19 web sphere as they are included in more than one national collection?

Overall, research questions that have a transnational component require more comparable metadata than those based only on URLs. For an analysis of which actor categories were archived (state, commercial news outlets or broadcasters, [...]) a website typology would need to be provided by the web archives, for example.

Conclusions

The aim of this chapter was to assess the usefulness of URLs in understanding the scope of COVID-19 web archive collections from the point of view of two different fictional profiles: a researcher and a web archivist, each of them interested in a specific use case. Three analyses were done to illustrate how they could be used by both profiles, in the case of the researcher a decade after the original datasets were created. Analysing the geographical spread of the collection allowed to determine what the international scope of each collection was. The analysis of the overlap between the different collections showed that there was not much overlap; furthermore, there was not enough material to warrant formulating specific research questions to analyse the content of this overlap. The wide variety of approaches towards including social media in the COVID-19 special collections show that it is still a developing field. The different analyses allowed the web archivist to compare the web archiving practices in her institution with those at other national libraries and enabled the researcher to decide which corpora would be most useful for him to look further into.

However, given the significant amount of time that was necessary to compile and clean the dataset, it is important to underline the need for web archives to make as much data as possible available to researchers before they enter the reading rooms. It would also be beneficial to develop sustainable data curation workflows that would allow for the extraction and publication of datasets within dedicated “labs” in libraries. Effective access, through tools and data standards, is therefore essential. For example, the search engine Solr Wayback provides the opportunity to extract 60 data fields from WARC files and export this data in both CSV and JSON formats. Each data field is precisely defined within the documentation of Solr Wayback, leaving no ambiguity regarding the meaning of the data. This could allow the datasets of different national web archives to be interoperable, enabling the formation of transnational corpora.

As a result of early access to the COVID-19 special collections, this study had to overcome some constraints in accessing and engaging with the data. Due to their historical importance, special collections about unforeseen events have the potential to be widely used, but to study them effectively, it is necessary to raise awareness of new forms of heritagisation, such as that which takes place on the live web. Early access to special collections presents an opportunity to further develop dashboards that can support curatorial decisions. These dashboards could provide insights into the scope of collections, including the usability of the collected data, and identify any important gaps for future research. They could also facilitate collaborative approaches to archiving web content on a transnational scale. Regular progress and data quality reviews are essential not only to ensure transparency regarding the scope and creation of the collection but also to enable informed decisions and generate consensus on actions required to enhance the data quality of these collections.

Acknowledgements

This analysis would not have been possible without the collaboration of a number of organisations who were gracious enough to share the seed list of their COVID-19 collections.

What can we learn from URLs?

We would like to thank the web archiving teams of the Royal Danish Library (Netarkivet), Bibliothèque nationale de France, Institut National de l'Audiovisuel (France), IIPC, Bibliothèque nationale du Luxembourg, National Library of The Netherlands, National Széchényi Library and the British Library.

We would also like to thank the other members of the WARCnet working group “Analysing Transnational Events” for their help and support: Nicola Bingham, Niels Brügger, Frédéric Clavert, Sophie Gebeil, Claude Mussou, Federico Nanni, Caroline Nyvang, Henrik Smith Sivertsen, Helle Strandgaard Jensen and Jane Winters. A special thank you to Valérie Schafer, the chair of the working group, for her excellent managerial and organisational skills.

Note

- 1 Participating IIPC members include Arquivo.pt; Biblioteca Nacional de España; Bibliothèque et Archives nationales du Québec; Bibliothèque nationale de France; Bibliothèque nationale du Luxembourg; The British Library; Columbia University Libraries; Deutsche Nationalbibliothek; Eesti Rahvusraamatukogu; Εθνική Βιβλιοθήκη της Ελλάδος; Internet Archive; Kansalliskirjasto; Koninklijke Bibliotheek; Library and Archives Canada / Bibliothèque et Archives Canada; Library of Congress; Library of Congress, Rio de Janeiro Office; Nacionalna i sveučilišna knjižnica u Zagrebu; Narodna Biblioteka Srbije; Národní knihovna České Republiky; Nasjonalbiblioteket; National Diet Library, Japan; National Library of China; National Library of Ireland; National Library of Korea; National Library of New Zealand; National Library of Scotland; NetArchive.dk; Országos Széchényi Könyvtár; Österreichische Nationalbibliothek; Schweizerische Nationalbibliothek; Univerzitná knižnica v Bratislave.

References

- Aasman, S., Bingham, N., Brügger, N., de Wild, K., Gebeil, S., & Schafer, V. (2021). *Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections*. WARCnet Paper. https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Aasman_et_al_Chicken_and_Egg.pdf. Archived at Wayback Machine September 24, 2023. https://web.archive.org/web/20230924000852/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Aasman_et_al_Chicken_and_Egg.pdf.
- Acker, A., & Chalet, M. (2020). The weaponization of web archives: Data craft and COVID-19 publics. *Harvard Kennedy School (HKS) Misinformation Review*. <https://doi.org/10.37016/mr-2020-41>.
- Adams, T., & Kopelman, S. (2022). Remembering COVID-19: Memory, crisis, and social media. *Media, Culture & Society*, 44(2), 266–285. <https://doi.org/10.1177/01634437211048377>.
- Ben-David, A., & Amram, A. (2018). The Internet Archive and the socio-technical construction of historical facts. *Internet Histories*, 2(1–2), 179–201. <https://doi.org/10.1080/24701475.2018.1455412>.
- Bibliothèque nationale du Luxembourg. (2023). *Luxembourg Web Archive*. www.webarchive.lu/. Archived at Wayback Machine February 22, 2024. <https://web.archive.org/web/20231016084352/https://www.webarchive.lu/>.
- Bingham, N., & Byrne, H. (2021). Archival strategies for contemporary collecting in a world of big data: Challenges and opportunities with curating the UK web archive. *Big Data & Society*, 8(1). <https://doi.org/10.1177/2053951721990409>.
- Bingham, N., & Geeraert, F. (2020a). *Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative Collection*. WARCnet paper. Archived at Wayback Machine September 24, 2023. <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.
- Bingham, N., & Geeraert, F. (2020b). *Exploring special web archives collections related to COVID-19: The case of the UK Web Archive*. WARCnet paper. Archived at Wayback Machine September 24, 2023. <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.

- Brügger, N., & Laursen, D. (Eds.). (2019). *The historical web and digital humanities. The case of national web domains*. Routledge.
- Danish Royal Library. (2023). *Research access to Netarkivet*. www.kb.dk/en/find-materials/collections/netarkivet/research-access. Archived at Wayback Machine February 22, 2024. <https://web.archive.org/save/https://www.kb.dk/en/find-materials/collections/netarkivet/research-access>.
- Department for Transport & the Rt Hon Grant Shapps MP. (2020, November 7). *Travel ban implemented to protect public health following Denmark COVID-19 mink outbreak*. Gov.uk. www.gov.uk/government/news/travel-ban-implemented-to-protect-public-health-following-denmark-covid-19-mink-outbreak. Archived at Wayback Machine February 22, 2024. <https://web.archive.org/web/20240222103751/https://www.gov.uk/government/news/travel-ban-implemented-to-protect-public-health-following-denmark-covid-19-mink-outbreak>.
- Erl, A. (2020). Afterword: Memory worlds in times of Corona. *Memory Studies*, 13(5), 861–874. <https://doi.org/10.1177/1750698020943014>.
- Hoskins, A. (2011). Media, memory, metaphor: Remembering and the connective turn. *Parallax*, 17(4), 19–31. <https://doi.org/10.1080/13534645.2011.605573>.
- Keightley, E., & Schlesinger, P. (2014). Digital media — social memory: Remembering in digitally networked times. *Media, Culture & Society*, 36(6), 745–747. <https://doi.org/10.1177/0163443714532985>.
- Lim, E., Srivastava, J., Prabhakar, S., & Richardson, J. (1996). Entity identification in database integration. *Information Sciences*, 89, 1–38. [https://doi.org/10.1016/0020-0255\(95\)00185-9](https://doi.org/10.1016/0020-0255(95)00185-9).
- Maemura, E. (2022). *Towards an infrastructural description of archived web data*. WARCnet paper. Archived at Wayback Machine September 24, 2023. <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.
- Maemura, E., Worby, N., Milligan, I., & Becker, C. (2018). If these crawls could talk: Studying and documenting web archives provenance. *Journal of the Association for Information Science and Technology*, 69(10), 1223–1233. <https://doi.org/10.1002/asi.24048>.
- Milligan, I., Ruest, N., & Lin, J. (2016). Content selection and curation for web archiving: The gatekeepers vs. the masses. *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 107–110. <https://doi.org/10.1145/2910896.2910913>.
- Musiani, F., & Schafer, V. (2016). Digital Heritage and Heritagisation. *RESET*, 6. <https://doi.org/10.4000/reset.806>.
- Németh, M., & Geeraert, F. (2020). *Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary*. WARCnet paper. Archived at Wayback Machine September 24, 2023. <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.
- Retsinformation.dk. (2004). *Lov om pligtaflevering af offentliggjort materiale*. www.retsinformation.dk/eli/lta/2004/1439. Archived at Wayback Machine February 23, 2024. <https://web.archive.org/web/20231207175156/https://www.retsinformation.dk/eli/lta/2004/1439>.
- Thouki, A. (2022). Heritagization of religious sites: In search of visitor agency and the dialectics underlying heritage planning assemblages. *International Journal of Heritage Studies*, 28(9), 1036–1065. <https://doi.org/10.1080/13527258.2022.2122535>.
- Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, 1, 85–111. <https://doi.org/10.1007/s42803-019-00007-7>.
- Webster, P. (2017). Users, technologies, organisations: Towards a cultural history of world web archiving. In N. Brügger (Ed.), *Web 25: Histories from 25 years of the world wide web* (pp. 175–190). Peter Lang.

11

THE CHALLENGES OF SEARCHING FOR WOMEN IN THE COVID-19 WEB ARCHIVE COLLECTIONS

Promises, achievements, and pitfalls

*Valérie Schafer, Susan Aasman, Frédéric Clavert,
Karin de Wild, and Joshgun Sirajzade*

Introduction

In early 2020, shortly after the launch of the WARCnet project (Web ARChive studies network researching web domains and events), the first signs of the COVID-19 crisis began to emerge, in March–April 2020, by drastic measures leading to lockdowns in various countries in view of the scale and effects of the pandemic and its rapid circulation.

As the WARCnet project went online, its Working Group 2, dedicated to transnational events and their searchability on the Web, was not only keenly aware of the disruptive and unprecedented nature of this crisis but also of the efforts initiated by web archivists in many institutions to begin documenting the digital traces of this event by means of special crawls and collections. The aim of this chapter is to highlight one aspect of the research initiated in WARCnet and expanded in parallel in a year-long research project, AWAC2 (“Analysing Web Archives of the COVID Crisis” through the International Internet Preservation Consortium (IIPC) Novel Coronavirus dataset). The latter was conducted in partnership with the IIPC, Archive-It, and the Archive Unleashed Team, using the Archives Research Compute Hub (ARCH; see ARCH, 2023) interface and its tools.

This research allowed us to experiment with the methods used to approach large, transnational, multilingual corpora, but also to question the strategies and limits of this collaborative research. This chapter focuses not so much on the results relating to women and gender issues in the face of the health crisis, but rather on the challenges of research-driven questions employing methods for searching, accessing, and processing data and web archives in an international dataset and an interdisciplinary research environment.

As a result, the chapter sheds light on the challenges raised by the IIPC collection, including multilingualism, the size of the dataset, access and searchability, silence and noise, and loss of information. It also presents a number of approaches, tools, and methods that we

used on this corpus, for instance, the creation of samples and sub-corpora, trial and error, iterative processes, comparisons of several algorithms, datathons, etc., while hybridizing close and distant reading, as well as data-driven and research-driven methods. It highlights the debates and technical challenges we faced, principally in selecting computational tools. Three major algorithms were applied when carrying out a semantic search of the corpus.

Therefore, this chapter explores issues relating to transnational studies, digital methods, and scalable reading, while also reflecting on collective methods, asymmetries, and inclusiveness in web archives and, last but not least, on transnational events and studies of live events by means of web archives.

The first section provides an explanation of the research context: we describe the conditions under which access to the IIPC corpus was made possible, its constitution, the documentation work performed within the WARCnet project and the partnership with the Archives Unleashed Team. The second section explores the research strategies and the challenges raised by the dataset and our research questions, while also providing some recommendations for the analysis of large datasets and for transnational studies in web archives.

Pushing the boundaries

The initial goal of the WARCnet project was to expand the boundaries of methods and analysis and to advance web studies, especially in the field of transnational research (Brügger, 2020). It began with the idea of testing and fostering transnational case studies by building on existing collections, but the subsequent health crisis sparked a desire to document and study an international event in real time. It energized the study of a corpus, which has a number of specific characteristics that need to be promptly highlighted.

Documenting web archiving in real time

Conducted at the same time as the project, the COVID-19 collections of national web archives offered an opportunity to monitor and widely document the initiatives launched on a European level. However, the corpora were not immediately indexed nor available: in the first half of 2020, the emphasis was clearly on preservation and not necessarily on indexing or providing immediate access.

Moreover, the choice of a living collection of web archives, which were compiled in a disruptive time of crisis (Rhodes, 2013), is a particular feature of the study. Indeed, it is not based on annual crawls and regular collections conducted by libraries and heritage institutions, but on an urgent process targeting specific content. However, in many countries, regular collections also include COVID-related content from 2020, as is the case for the French BnF's *Actualités* (news) collection. As a result, hybrid collections (e.g., incorporating both targeted crawls and content collected throughout regular crawls) exist.

Each country implemented its own strategy, depending on its human, data, and technical resources. The differences between countries include social media and calls for participation (within the institution or to the general public). The time frames also differ depending on the institution: some mainly preserved the first wave of the pandemic, while others also proceeded with crawls and special collections in the following months. While formats are largely WARC files, an institution such as the French INA (Institut National de l'Audiovisuel) is an exception because of its choice of JSON files, which are better suited to its audio-visual scope, the inclusion of social media content and its methods.

Searching for women in the COVID-19 web archive collections

Therefore, the working group's first steps in the study of pandemic preservation focused on national collections and on selection and curation methods. Oral interviews (WARCnet, 2024) were conducted to document ongoing efforts to capture the COVID-19 crisis in several countries and to compare strategies. In parallel, we asked a number of institutions (Royal Danish Library, Bibliothèque nationale de France, National Széchényi Library, Hungary, IIPC, Bibliothèque nationale du Luxembourg, Koninklijke Bibliotheek, the Netherlands, UK Web Archive — The British Library) to make their derived data and metadata available (see Geeraert, de Wild, and Aasman, 2024). In effect, access to content and web archives themselves is often limited to on-site access only by copyright or legal deposit legislation. The presence of many web archivists in the WARCnet project and long-established connections, as well as the project's framework, enabled these datasets to be received.

The datasets showed a certain degree of homogeneity (except for INA) in terms of the formats used, as the majority of institutions use the Heritrix robot for crawling and WARC files (Aasman et al., 2021b). There were often common metadata (Pomerantz, 2015) but also some differences in the entries, which required the data to be cleaned before envisaging comparisons (see Table 3 in Aasman et al., 2021a). Some institutions also provided detailed explanations of their collection methods, along with the provision of datasets, which complemented the interviews we conducted with web archivists. Indeed, it was by combining the testimonies of archivists with the datasets that it was possible to understand the extent of the effort made by institutions (Geeraert, Winters et al., 2021). This approach sheds light on their collection strategies, which involved varying degrees of public participation, as well as diverse time frames, different scopes or archiving policies, and a focus on different types of content.

Within these datasets, the IIPC collection clearly stands out for the number of TLDs (Top Level Domains) and URLs (Uniform Resource Locators). As a result of a collective work led by the Content Development Group of the IIPC, this collection transcends national (and European) boundaries because of a joint effort. It was an interesting case study for our research in Working Group 2, which aimed to assess the possibility of transnational studies by means of web archives relating to events, as it provided an existing complete transnational corpus, which is extremely unusual and rare.

The IIPC transnational collection as a new case study

The IIPC was formally established in July 2003 by 12 participating institutions. It now consists of organizations from over 45 countries and its members participate in working groups, one being the Content Development Working Group (Bingham, 2020), which pursues collaborative collections based on themes or events (IIPC, 2023). The novel coronavirus collection is just one of the collections released by this working group, which was created in 2014. In 2015, under the leadership of the founding co-chairs Abbie Grotke (Library of Congress) and Alex Thurman (Columbia University Libraries), it took out a subscription to Archive-It (Geeraert & Bingham, 2020) and maintains collections relating to the Olympic Games, Climate change or the European Refugee Crisis on Archive-It (2023).

The IIPC collection stands out from other collections, because of the international participation that allowed the working group to gather a large amount of data totaling around 5Tb¹, when the data was downloaded. The team was granted access to the collection because of an agreement between three stakeholders: the IIPC, our team — or to be more precise

a small part of Working Group 2, which made a special application² by means of the call for Archives Unleashed Cohorts, and finally the Archives Unleashed Team (AUT).

One of the interviews, with Nicola Bingham (National British Library), the current co-chair of the IIPC Content Development Working Group with Alex Thurman (Columbia University Libraries), who was also part of our Working Group 2 within the WARCnet Project, helped the team get a sense of the IIPC collection (Geeraert & Bingham, 2020). For example, Nicola Bingham provided details of the way in which the crawled URLs were selected and of the curation process:

For this collection, we had a Google sheet for members of the CDG and we had a separate form for public nominations. The members were able to contribute directly to the Google sheet and complete it. They were also able to see nominations that other members had submitted. For the public nominations, we set up a Google form [...].

As mentioned, the basic metadata is also submitted, such as title, language, description, top-level domain name, and we also ask members to specify the crawl scope as well. The nominations on the sheets are overseen by myself and Alex, and we manually review each seed before adding it to Archive-It to be crawled.

(*ibid.*)

This human curation is important in understanding the scope of the collection, as are the choices that were made. From the interview, we learned, for example, that video and social media content were not included, mainly because the IIPC was concerned about the data budget³ (Geeraert & Bingham, 2020). Web archives are always a complex intertwinement of technical and human agencies, negotiations, choices, and contextual and institutional requirements (Schafer et al., 2016; Maemura et al., 2018).

The balance between countries, languages (although English is clearly prevalent in the corpus), and some data limits were also considered when framing the process:

We initially didn't apply any limits to the number of seeds that any particular institution or region could nominate, but we did find quite early on that we had overwhelming nominations from some agencies. For example, one of our colleagues had implemented an automatic retrieval of content that was related to coronavirus. This returned a couple of thousand seeds. This is brilliant, but due to the data limits and the fact that we want to represent all regions and countries, we did some filtering based on that list of a thousand seeds.

(Geeraert & Bingham, 2020)

The latter was also at the core of national archiving, as National institutions needed to deal with their internal data limits and to balance archiving of the COVID-19 crisis with other crawls and events, once it became obvious that the pandemic was a prolonged situation and would experience several waves.

However, the 3.3 terabytes that were already collected at the time of the IIPC interview (2020), comprising 21,600,104 documents, including HTML, Word files, and PDFs, were already quite impressive.

It is interesting to note that, while filtering is not visible in the collection, the IIPC kept a record of all URLs nominated, making it possible to clearly document the URLs that were nominated but not crawled. This is undoubtedly important for documenting the collection

Searching for women in the COVID-19 web archive collections

and offers a clear example of data that may complement a collection, providing transparency regarding selection, while also avoiding the illusion of an exhaustive collection. It should also be borne in mind, when studying this collection, that only a proportion of national “web spheres” and their content was nominated and that, in many cases, national collections have preserved more websites and content. Indeed, Friedel Geeraert, an expert in web archiving at KBR (Belgium) and a member of the WARCnet Working Group 2, performed a comparison of several countries regarding their datasets and how they are represented in the IIPC collection, in an attempt to measure overlaps. Her comparison shows that there is clearly a rather small overlap, which underlines the need to cross-reference the data from the IIPC collection with national collections, to obtain a better view of where the focus is on specific countries (Geeraert, de Wild, & Aasman, 2024).

*“Lowering barriers to access”: A step forward, because of the
AUT and ARCH*

As outlined above, a number of stakeholders contributed to the research: part of the WARCnet Working Group 2, the IIPC, as well as Archive-It and the Archives Unleashed Team. The IIPC dataset was not provided directly to the team but was displayed and preserved by means of Archive-It and made available to the team via the ARCH interface created by the Archives Unleashed Team (AUT).

Since 2015, the AUT has been part of the work on “Lowering Barriers to Access Through Community and Infrastructure” (Milligan, 2020) and has organized regular workshops and datathons, while at the same time developing tools to help analyze a vast quantity of web archives. Its experience in the field has allowed it to better understand the diversity of users and the need for several levels of access and user experience, as well as for mentoring and supporting researchers. As explained by Ian Milligan (2020), the AUT has progressively moved from command-line-based-tools into WARC files to a whole infrastructure (Ruest et al., 2021) and developed an interface known as ARCH (The Archives Research Compute Hub):

Our initial expectations were that humanists and historians would be able to use the Toolkit by taking examples in our robust documentation “cookbook” and changing variables in order to work with it on their own data (i.e. changing the input path, the output path, and the specific filters or extractors being used). Unfortunately, we discovered that this required more technical competency than we could expect of the average scholar considering all the skills they already needed: users needed not only to know how to tweak code, but how to use a command line, follow the flow of programming scripts, and then have patience for open-source documentation and projects. In order to engage historians, it became clear that we would need to do something more than just provide command-line-based tools. [...] Thinking about infrastructure, we realized that we needed to move away from WARC files and towards derivative file formats. While WARCs are fantastic, standards-based files, the research user community is too small to use them effectively.

(Milligan, 2020, 9)

In effect, ARCH provides users with datasets, each time with the option of accessing the full dataset or a sample: domain frequency statistics, different kinds of network graphs

(based on links between pages or domains, file formats, etc.), text-based files (full-text, named entities, text file information), and a large quantity of file format information. It also produces some illustrations for exploring collection content. The ARCH design process was iterative, from the initial prototype to several stages of refinement based on internal and external UX Testing, by means of interviews, surveys, etc. (Archive-It, 2021). As such, the project demonstrates the need for a collaborative team effort when it comes to exploring and experimenting with the challenges of born-digital archive materials at scale.

A (too) ambitious topic: Women, gender, COVID

The study of the IIPC collection ran from mid-2021 to mid-2022 in two key stages: the first phase of the experiment aimed to simply appropriate and understand the dataset; thereby being a question, for example, of exploring multilingualism or the origin of content. The second phase of the study aimed to dig deeper into a specific topic. A number of topics were put to IIPC members for a vote (Aasman et al., 2021a). The topic of women was chosen, in particular, because of the ongoing observation of social media and web content during the pandemic. In fact, several central themes, which clearly involved gender issues, and in particular women, have emerged since the first wave of the pandemic. Among the themes that appear to attract popular attention, is that of the mental, as well as daily, burden devolved to women because of the closure of schools and the combination of home schooling and home working, as well as topics relating to domestic violence, nurses, and their working conditions, etc. Several reports and a quantity of web content specifically addressing the case of women were produced during the pandemic, for instance, by UN Women (UN Women, 2021), the European Commission (Blasko et al., 2020), or Eurofound (Nivakoski et al., 2022) to name a few.

An initial search of the French INA (Institut National de l'Audiovisuel) archives, to test the feasibility of such a study, allowed us to quickly identify additional themes and, in particular, that of pregnancy during COVID-19. The latter topic was also strongly reflected as a key topic for academic research in Google Scholar. Some of the findings in this corpus of web archives and videos relating to audio-visual material also revealed content that was more closely linked to the scope of the INA collection, for example, the reverberations of the hospitalization of Tom Hanks and his wife in the early days of the pandemic (March 2020).

The question of the keywords to be applied in search strategies quickly emerged as an issue in this first test. For instance, in French, it is necessary to search for “infirmière” and not “femme” to obtain feedback on (female) nurses. Feminist content is often not tagged as such and may rely on sentences or images that are not directly identifiable by a simple request for “Women + COVID”.

The topic raises interesting questions in terms of multilingualism, gender, power, and asymmetries, as well as inclusiveness in web archives while permitting discourse analysis, topic modeling, and a study of the participation and players in the field. However, being a complex and intersecting topic, it also increases the challenge.

Confronting the dataset: What we learned (and confirmed)

The first steps with ARCH during the academic year 2021 were both very exciting and somewhat frustrating for the AWAC2 cohort. It quickly became apparent that this vast dataset required different investigation methods, whereas we had been accustomed to working quite

Searching for women in the COVID-19 web archive collections

easily with Excel sheets. To give just one example, the domain-frequency.csv.gz listing the domains in the corpus had 161,757 lines.

Other challenges related to the fact that, initially, the ARCH interface only allowed one user at a time to download the data, and for the larger CSV files, the download time on a regular computer was lengthy, or even impossible. Once the files were downloaded, most operations that we needed to perform required computing time and power that exceeded personal computers' capacities, prompting us to seek assistance from the University of Luxembourg's High-Performance Computing (HPC) facility. In addition, the corpus was initially relatively unstable and continued to regularly expand via Archive-It.

A computer scientist joined the team when it became apparent that such a massive amount of data could not be analyzed solely by social sciences and humanities researchers, even with advanced skills in digital humanities. However, a gap quickly emerged in our AWAC2 team between research-driven and data-driven approaches, between those who dived into algorithms and those who were more accustomed to qualitative or hybrid methods and were less comfortable with big data. The latter group struggled to find its place in this immense dataset, which raised its own challenges due to the international nature and creation process.

The challenge of multilingualism

As outlined above, more than 30 IIPC members contributed to the IIPC collection, along with hundreds of individuals and institutions by way of public nominations. This diversity of participation and output should be noted; it reflects both the willingness to create an international collection as well as a desire for more inclusiveness (Schafer & Winters, 2021). This is evident in the languages represented in the final collection (*see* Table 7 in Aasman et al., 2021a), as there are 51 languages, which can be retrieved on the Archive-It collection page. It also provides a view of the prevalence of languages, with English at the top, followed primarily by Portuguese, Spanish, Japanese, French, German, and Dutch. Besides the diversity of languages, it is also important to bear in mind that crawl frequency by language varies over time, depending on the pace of nomination and crawling, as well as the regularity of monitoring the URLs to be archived.

The team needed to draw on language-specific sub-corpora, to construct tailored corpora for each language, which capture the nuances and specific context of each linguistic sphere. This allowed us to explore the intricacies and complexities of these topics through a linguistic lens and to understand the similarities and differences in experiences across various cultural and linguistic contexts. It thus enriched our analysis and fostered a more inclusive understanding of the topic, while also making analysis more manageable and the datasets smaller. However, it is crucial to also acknowledge the presence of countries with more than one official language, such as Luxembourg and Canada, within several sub-corpora.

Noise and duplicates in the archive and a need for sub-datasets

One of the difficulties of this international dataset lies in the full-text keyword search, as even a simple search requires an account to be taken of a multitude of varying formulas and cultural references. As outlined above, the word “nurse” in English does not have a gender, whereas in French it does, distinguishing between “infirmiers” and “infirmières”, “soignants” et “soignantes” (male and female caregivers), etc.

This search issue is compounded by the rapid discovery of content that appears to link women and COVID, but which belongs to two different contexts. This is particularly the case for news websites where, during the COVID pandemic, the MeToo movement was also in the headlines (see Figure 2 in Aasman et al., 2022). Thus, the proximity of keywords may suggest a connection between two topics, when in fact they are different stories. This scenario is quite common, and the IIPC collection contains numerous press sites and general websites, which create such noise and can distort analysis and generate biases, notably when developing topic modeling. Furthermore, many websites display dynamic content (including news, sponsored content, ads, etc.) that can relate to “women” in general in a context other than the COVID-19 crisis, adding to the noise problem.

Another factor that contributes to the presence of noise in the corpus relates to the duplicates found in the collection. We found very few (less than 1%) strict copies (i.e. even the web page’s crawl date and time were identical). Most copies were web pages that did not change between two crawls. We also spotted versions — web pages with variations between two crawls. The presence of copies and versions can skew the analysis of the data. Therefore, it would be important to effectively detect and manage duplicates.⁴

Finally, as part of our comprehensive web archive research on the impact of COVID-19 on women and the relationship between the two, it became evident that relying solely on these generic search terms is inadequate. While they provide a broad overview, they fall short when it comes to investigating specific and nuanced issues, such as home-schooling and domestic violence. To delve deeper into these intricacies, our search methodology relied on the creation of targeted and more precise sub-corpora.

Selecting and testing tools, algorithms, methods

Machine learning models were used as part of the experiments to build topic modeling analyses. This quantitative analysis and approach seemed obvious but became more complex than expected when digging into our sub-corpora. Information retrieval, text mining, and deep learning allowed us to progress in several directions, which could be combined to better grasp the topic of women and COVID-19. Firstly, this involves a semantic search, i.e., searching for words relating to feminism, such as “women”, “female”, or “girl” etc., and searching for broader concepts, such as “feminism”, etc., secondly, this involves topic extraction, i.e., looking for the topics where these words occur — with topics usually being defined as words that occur together as the crux of a document, finally, this involves looking for similar documents (i.e., which websites thematize women? What is often said about women?).

In effect, three major algorithms were applied, to be able to carry out a semantic search of the corpus: Latent Dirichlet Allocation (LDA), Word2vec, and Doc2vec (Aasman et al., 2022). LDA is an extension of Probabilistic Latent Semantic Analysis (PLSA), which is a probabilistic formulation of Latent Semantic Analysis (LSA). LSA is a dimensionality-reduction technique where documents in a corpus are compressed to a very small number of documents, which are readable by a human. These compressed documents are called topics. In essence, they contain the words that are shared by many documents and probabilistically occur more often together. Our experiment provides a way of identifying topics that contain keywords relating to women, but also of looking at how these topics are distributed across time and web pages (Aasman et al., 2021b).

Wor2vec and Doc2vec are further formulations of previous algorithms, based on neural networks, and offer more flexible use. Word2vec is very similar to LSA. However, Word2vec

Searching for women in the COVID-19 web archive collections

creates vectors from the so-called window or neighborhood of words, such as five words to the left and right of the searched word, operating on a more syntactical level, while LSA is purely based on a document-term matrix (a description of the frequency of terms in a collection of documents). With Word2vec it is not only possible to find words that relate semantically to the searched word, but also to connect the vectors of all the words in the document. This goes beyond the so-called bag-of-words approach, to which all the previous algorithms belong, as it enables accounts to also be taken of word order. In terms of implementation, this is performed with an additional algorithm, such as a Long-Short-Term Memory (LSTM) or a ready-to-use version with Doc2vec (Aasman et al., 2022). In our experiment, we trained a Word2vec algorithm on our corpus, which enabled us to find keywords relating to women or feminism. We also searched for where these keywords occur.

The advantages of the combination of these tools and algorithms are obvious: LDA-identified words that are shared by numerous documents and probabilistically occur more often together. For their part, Word2vec and Doc2vec allowed us to find words relating semantically to the searched word, while the vectors of all words in the document can be connected and account taken of word order. Comparing these results to those provided by LDA helped us analyze the efficiency of the algorithms, as well as offering a way of ensuring that the search covers our corpus in as much detail as possible.

These experiments paved the way for a number of case studies and initial visualizations (see Figure 11.1); for example, a time series and the top 20 domains for one particular topic: “children health mental child support family parents anxiety families care people social kids resources stress home time violence young **women**”.

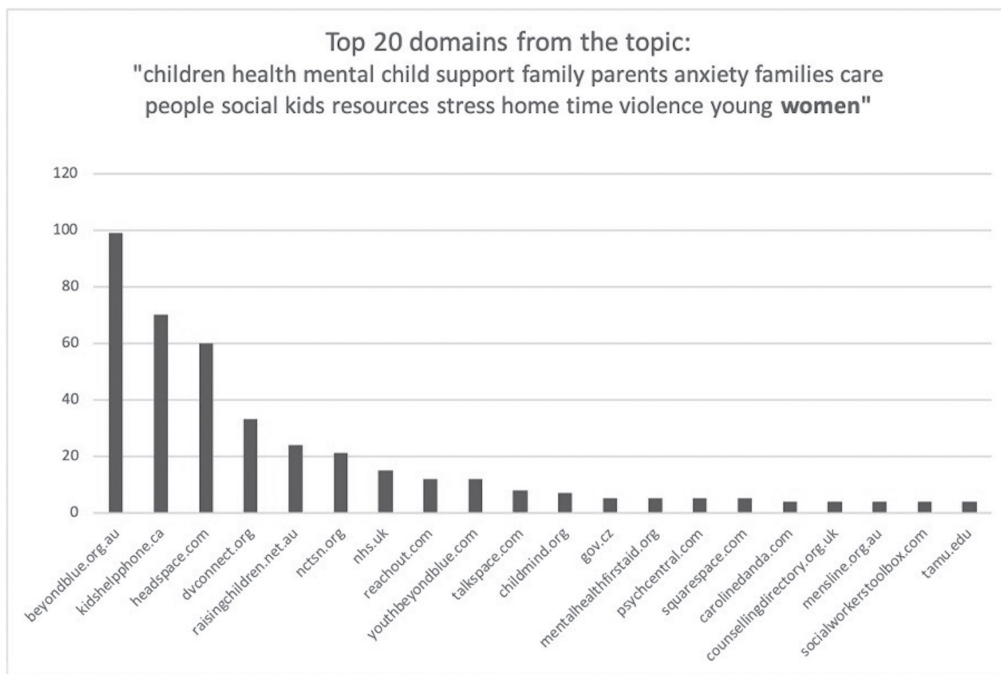


Figure 11.1 Top 20 domains in the IIPC collection from one topic (by Joshgun Sirajzade).

The visualization shows a clear dominance of a web sphere rooted in mental health issues, sometimes funded by governments (Beyondblue.org.au), and often by commercial organizations (Headspace.com). It also demonstrates that children were an important focus point. As Karsdorp et al. (2021: 30) suggest: “[...] producing elementary visualizations are important for building intuitions about a particular data collection, and, when used strategically, can aid the formulation of new hypothesis and research questions”.

Context and scalable reading stay key for analysis

The study was characterized from the outset by the intertwining of close and distant reading, which is becoming standard in Digital Humanities (Clavert & Fickers, 2021). The oral interviews conducted within the WARCnet Working Group 2 (Geeraert, Bingham, et al., 2024) were part of a more qualitative approach, while accessing the IIPC dataset via the ARCH interface quickly led to a quantitative approach and some distant reading (Moretti, 2007). As a result, the team constantly moved back and forth between these approaches, applying scalable reading (Mueller, 2014) to the corpus. However, the technical challenges encountered when accessing the IIPC collection necessarily prompted a quantitative turn and a focus on computational processing. The time needed to fully understand the diversity of the collection, as well as issues of multilingualism, for example, should not be underestimated. Therefore, a significant part of the study was devoted to appropriating and understanding the data, and there was a fundamental tendency to move away from the core of the subject to first grasp the inherent qualities and limitations of the datasets. Similarly, once this initial analysis had been conducted and sub-corpora relating to women and gender had been created, it was, once again, the technical issues that took precedence over more qualitative analysis. However, the latter remained necessary, especially for identifying noise, duplicates, and emerging themes, which cannot be addressed solely by word co-occurrence analyses or topic modeling.

Contextualization is extremely important, both for understanding the datasets and for analyzing quantitative results. Thus, a double level of scalable reading is essential, both with respect to the corpus itself and its content. There is a clear need to find a balance between a research-driven and data-driven approach.

In fact, contextualization applies on multiple levels. The diversity of countries and national “web spheres” represented poses an analytical challenge, due to variations in the measures adopted by different countries and their responses to the COVID-19 crisis. They followed different timelines for entering or exiting lockdowns, and they do not have the same healthcare systems in place or face the same home-schooling constraints. Delving into the issue of home-schooling, it becomes evident that even on a European level, there is a significant difference in educational systems and parental support. Similarly, when studying the role of women and their employment as nurses, this diversity persists.

Additionally, there is a wide range of “web spheres” (Brügger et al., 2015), which inevitably influences the analyses that can be conducted. Whether this involves analyzing news websites, government portals, or health-related websites, obtaining an accurate perspective is incredibly challenging: offering a nuanced analysis of the context in which content is created, as well as its intended audience, becomes extremely complex. Factors such as language barriers, legal regulations, cultural and media differences, and varying levels of internet penetration further complicate the analysis.

In view of these challenges, it is essential to adopt an interdisciplinary and multi-faceted approach. This approach should take account not only of the diversity of countries, “web

spheres”, and socio-economic contexts but should also encompass linguistic and cultural perspectives. Collaboration across disciplines and the combination of quantitative data analysis with qualitative interpretation and contextual understanding are the only ways to obtain a deeper insight into the global landscape.

A need for collective work and interdisciplinarity

Web archive research has undergone a significant shift toward interdisciplinarity, highlighting the need for collaborative work and innovative methodologies. Social Sciences and Humanities (SSH) scholars alone can no longer perform the task of analyzing web archives when it comes to “big data”⁵ like the IIPC collection, due to its vast scale and technical complexity. As highlighted by Nick Ruest et al. (2021), this challenge involves grappling with “petabytes of data, billions of words, tens of thousands of images, all accompanied by murky metadata, provenance, and accessibility issues”. The advent of this kind of “big data” has created a pressing need for expertise from computer scientists, beyond their engineering roles. Their contribution extends far beyond providing support or technical assistance: they play a pivotal role in not only translating research into computational approaches but also enriching it by means of innovative proposals based on digital methods and tools. Their data analytics, machine learning, and topic modeling expertise allow them to extract valuable insights from vast datasets, uncovering patterns, trends, and correlations, that may elude traditional research methods, and they can highlight new angles and uncover hidden relationships. The arrival of a computer scientist in the team was beneficial once the SSH researchers had already framed hypotheses. The latter were also able to offer contextualization: their understanding of the social, cultural, and historical contexts underpinning the data they were working with and enhancing the interpretation and analysis of big data, based on expertise in qualitative research methods, the interviews they had conducted, their deep knowledge of certain “web spheres” and the history of the web.

To address these challenges, the team selected datathons as a model for effective transdisciplinary collaboration and group learning. It had already used this kind of process within WARCnet (Aasman et al., 2021b) when embarking on studying national datasets. The first datathon had been conducted online, whereas, for the AWAC2 project, a 2-day intensive onsite meeting was possible, as the lockdown had ended and academic life was returning to normal in the spring of 2022. The datathon approach clearly offers a number of advantages for studying web archives (Fritz et al., 2021). For example, the continuous peer review process enhances the reliability of the datasets created. With multiple experts reviewing the data, we were able to identify inconsistencies in content and limitations in the collections. Our team consisted of individuals with advanced technical skills who led different stages of the project, while others had an in-depth knowledge of web archiving practices and qualitative study.

In addition to the datathon approach, our work with the AUT used a mentorship model. This was equally important, as it provided us with bi-monthly deadlines that drove us to produce valuable outcomes. Working closely with the AUT reinforced the importance of collaboration and the iterative nature of web archive research.

Conclusion

During this study, several points emerged regarding the exploration of an international phenomenon and a thematic corpus created with the participation of numerous heritage institutions.

First and foremost, it is important to emphasize how this study was able to overcome several barriers, particularly those relating to accessing international datasets. It is also important to highlight the collaboration, both with web archivists and the IIPC, as well as with Archive-It and the Archives Unleashed Team. The latter provided not only a simplified interface but also technical and intellectual support by means of regular exchange sessions, simultaneously enriching the development of the interface and our data analysis. This created a virtuous triangle between the creators and providers of datasets, the developers of tools for accessing them, and the research team. While, in the case of our study, the Archives Unleashed Team obtained feedback on the use of its tools, it is also worth noting how important the sense of being part of a collective effort was. This was the case in the bi-monthly meetings with the AUT team but also in the fact of belonging to a cohort, which allowed participants to meet each other at project launches, specifically for two other projects relating to COVID (but exploring more targeted datasets), or during exchanges with the IIPC or at a conference organized by Internet Archive. The sense of a collective is one of the aspects that guided the whole experience.

The potential of URLs, seed lists, metadata, and derived data for analysis is clear, but it is important not to underestimate the chain of transformations that data may have been subjected to. First, the data was subject to a curation process by national institutions based on the IIPC's criteria. The IIPC then curated the proposals and performed crawls, while the data was then further processed in the ARCH interface to enable access via URLs and mime types, among other things. Finally, the research team created several sub-corpora. This data transformation chain is extremely important and must be borne in mind when analyzing the data, as there may be some losses and biases created along the way.

Moreover, there are areas where we could have improved. For example, we set out via topics, but we could have targeted a specific language or a type of website (news, culture, etc.), or specific domains, such as those ending in .org, which would have changed our approach and the questions we were able to answer. Another potential area for exploration is the use of images and other multimedia in analysis. Our approach was mostly text-driven.

Finally, it is important to highlight the time required to study a collection as vast as that of the IIPC. It takes time to become familiar with the data before a researcher can even begin analysis. Researchers need sufficient time to delve into the intricacies of the collection, unravel its complexities, and understand its historical and cultural context. The one-year program conducted with the AUT was a crucial step, but it is not sufficient for a fully comprehensive exploration of the corpus' richness. It would require continuity and further effort, which was not possible as both the AWAC2 and WARCnet projects came to an end. The true potential of the corpus can only be fully realized with a sustained, long-term commitment, but account must be taken of the realities of academic research, including funding, staff mobility, and short-term positions, especially for young researchers.

In addition, this remark also highlights the significance of reproducibility in such an endeavor. The data from the IIPC and ARCH are readily available, enabling researchers who wish to continue or redirect this research to benefit from a corpus and from tools that not only facilitate reproducibility but also promote good scientific practice. In this respect, both the IIPC and the AUT contribute to FAIR Data principles (Mons, 2018).

Data availability

The Novel Coronavirus (COVID-19) Special collection was created by the Content Development Group of the IIPC in collaboration with Archive-It. The data collection that

Searching for women in the COVID-19 web archive collections

supports this study was downloaded on 24 January 2022 and is available via Archive-It. <https://web.archive.org/web/20230402105321/https://archive-it.org/collections/13529/>.

Code availability

The code written for our analyses is partly available at: <https://github.com/inactinique/awac2>

Acknowledgments

This research was conducted as part of the WARCnet project and the Archives Unleashed Cohort program (2021–2022). The authors would like to thank the Archives Unleashed Team for their feedback, mentorship, and support, and in particular Ian Milligan, University of Waterloo (CA), Nick Ruest, York University (CA), and Sam Fritz (Archives Unleashed), as well as the other members of our AWAC2 cohort, Niels Brügger, Aarhus University (DK), and Sophie Gebeil, University of Aix-Marseille (FR). They would also like to thank Archive-It and the IIPC, and in particular Olga Holownia, Nicola Bingham, and Alex Thurman for providing access to the data and for their support. Funds for the Archived Unleashed Cohort Team were provided by The Andrew W. Mellon Foundation (ref. no. 2001-07348). In 2020–2023, the WARCnet research network was funded by the Independent Research Fund Denmark, Humanities (grant no. 9055-00005B).

Notes

- 1 The IIPC COVID corpus is still expanding, and its current size is 6.3Tb. The files we used were generated on 24 January 2022.
- 2 The AWAC2 team comprised the authors of this chapter, along with Niels Brügger and Sophie Gebeil. Launched in 2020, the Cohort program we participated in is engaging with researchers in a year-long collaboration and mentorship with the Archives Unleashed Project and the Internet Archive, to support web archive research.
- 3 However, some content from social networks is part of the collection, and some of the social media websites were captured quite frequently (i.e., Twitter, Facebook, and YouTube). Yet this content must be treated cautiously, as foremost, the homepage or the login page was captured, but not the social media content itself. Therefore, we decided to exclude all social media websites from our dataset to prevent that this would influence the results.
- 4 Some members of the team are currently preparing an article about the management of versions and copies.
- 5 As underlined in Clavert et al., 2022, big data in history does not exactly correspond to the original definition that refers to the “3V”: Velocity, Volume, and Variety of data. Indeed, historical data logically often lacks velocity, meaning a constant and real-time renewal of the corpus. However, the other two “V” — volume and variety — relate directly to historians.

References

- Aasman, S., Brügger, N., de Wild, K., Clavert, F., Gebeil, S., & Schafer, V. (2021a). Analysing Web Archives of the COVID-19 Crisis through the IIPC collaborative collection: Early findings and further research question. IIPC netpreserve.org. Retrieved May 31, 2023, from <https://web.archive.org/web/20230322212643/https://netpreserveblog.wordpress.com/2021/11/02/analysing-web-archives-of-the-covid-19-crisis-through-the-iipc-collaborative-collection-early-findings-and-further-research-questions/>.

- Aasman, S., Bingham, N., Brügger, N., de Wild, K., Gebeil, S., & Schafer, V. (2021b). *Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections*. WARCnet Paper. Retrieved May 31, 2023, from https://web.archive.org/web/20230902214737/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Aasman_et_al_Chicken_and_Egg.pdf.
- Aasman, S., Brügger, N., de Wild, K., Clavert, F., Gebeil, S., Schafer, V., & Sirajzade, J. (2022). Studying women and the COVID-19 crisis through the IIPC coronavirus collection, IIPC, netpreserve.org. Retrieved October 16, 2023, from <https://web.archive.org/web/20230207134247/https://netpreserveblog.wordpress.com/2022/12/20/studying-women-and-the-COVID-19-crisis-through-the-iipc-coronavirus-collection/>.
- ARCH. (2023). Archives Research Compute Hub (ARCH). Website archivesunleashed.org. Retrieved February 26, 2024, from <https://web.archive.org/web/20230224212953/https://archivesunleashed.org/arch/>.
- Archive-It. (2023). International Internet preservation consortium. Archive-it.org. Retrieved February 24, 2024, from <https://web.archive.org/web/20230402054011/www.archive-it.org/home/IIPC>.
- Archive-It Team. (2021). Archives Unleashed and Archive-It's ARCH Program Update. Archive-It blog. Retrieved October 16, 2023, from <https://web.archive.org/web/20230428025730/https://ait.blog.archive.org/post/arch-program-update/>.
- Bingham, N. (2020). IIPC content development Group's activities 2019–2020. Retrieved October 16, 2023, from <https://web.archive.org/web/20230323130322/https://netpreserveblog.wordpress.com/2020/07/01/iipc-content-development-groups-activities-2019-2020/>.
- Blasko, Z., Papadimitriou, E., & Manca, A. R. (2020). *How Will the COVID-19 Crisis Affect Existing Gender Divides in Europe*. Publications Office of the European Union. Retrieved October 16, 2023, from <https://web.archive.org/web/20230401141007/https://publications.jrc.ec.europa.eu/repository/handle/JRC120525>.
- Brügger, N. (2020). *Welcome to WARCnet*. WARCnet Paper. Retrieved February 24, 2024, from <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.
- Brügger, N., Laursen, D., & Nielsen, J. (2015). Studying a nation's websphere over time: analytical and methodological considerations. *IIPC Conference: Innovation, Connection and Co-Operation in Web Data*, San Francisco, USA.
- Clavert, F., & Fickers, A. (2021). On pyramids, prisms, and scalable reading. *Journal of Digital History*, no. *jdh001*. Retrieved May 31, 2023, from <https://journalofdigitalhistory.org/en/article/jXupS3QAeNgb>.
- Clavert, F., Mahroug, S., & Schafer, V. (2022). Préservation et distorsion: L'espace-temps des réseaux socio-numériques et du web archivé. *Revue d'histoire culturelle*. Retrieved May 31, 2023, from <https://web.archive.org/web/20231004091659/https://revues.mshparisnord.fr/rhc/index.php?id=2791>.
- Fritz, S., Milligan, I., Ruest, N., & Lin, J. (2021). Fostering community engagement through Datathon events: The archives unleashed experience. *Digital Humanities Quarterly*, 15, 1. Retrieved October 16, 2023, from <https://web.archive.org/web/20230929031010/www.digitalhumanities.org/dhq/vol/15/1/000536/000536.html>.
- Geeraert, F., & Bingham, N. (2020). *Exploring Special Web Archives Collections Related to COVID-19: The Case of the IIPC Collaborative Collection*. WARCnet Paper. Retrieved May 31, 2023, from <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.
- Geeraert, F., Bingham, N., Clavert, F., Strandgaard Jensen, H., & Winters, J. (2024). Oral histories and scalable reading: analysing born-digital collecting practices during the COVID-19 pandemic. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *Companion to Transnational Web Archive Studies* (pp. xxx–xxx). Routledge.
- Geeraert, F., de Wild, K., & Aasman, S. (2024). What can we learn from URLs? Understanding the scope of COVID-19 web archive collections for transnational analyses. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *Companion to Transnational Web Archive Studies* (pp. xxx–xxx). Routledge.
- Geeraert, F., Winters, J., et al. (2021, 18 juin). Representation, participation and inclusivity: European web archives collecting the digital traces of COVID-19. Talk during the 4th RESAW Conference at

Searching for women in the COVID-19 web archive collections

- the University of Luxembourg. Retrieved October 16, 2023, from <https://web.archive.org/web/20230918164635/www.resaw2021.net/programme/>.
- IPC. (2023). Content Development Working Group. Netpreserve.org. Retrieved February 24, 2024, from <https://web.archive.org/web/20230323033003/https://netpreserve.org/about-us/working-groups/content-development-working-group/>.
- Karsdorp, F., Kestemont, M., & Riddell, A. (2021). *Humanities Data Analysis: Case Studies with Python*. Princeton University Press. Retrieved February 24, 2024, from <https://web.archive.org/web/20240217171742/www.humanitiesdataanalysis.org/introduction-cook-books/notebook.html>.
- Maemura, E., Worby, N., Milligan, I., & Becker, C. (2018). If these crawls could talk: Studying and documenting web archives provenance. *Journal of the Association for Information Science and Technology*, 69(10), 1223–1233.
- Milligan, I. (2020). *You Shouldn't Need to Be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure*. WARCnet Paper. Retrieved May 31, 2023, from <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.
- Mons, B. (2018). *Data Stewardship for Open Science. Implementing FAIR Principles*. CRC Press.
- Moretti, F. (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso.
- Mueller, M. (2014). Shakespeare his contemporaries: Collaborative curation and exploration of early modern drama in a digital environment. *DHQ: Digital Humanities Quarterly*, 8(3). Retrieved October 15, 2023, from <https://web.archive.org/web/20231003050948/www.digitalhumanities.org/dhq/vol/8/3/000183/000183.html>.
- Nivakoski, S., Calò, X., Mencarini, L., & Profeta, P. (2022). COVID-19 pandemic and the gender divide at work and home. Eurofound. Retrieved May 31, 2023, from <https://web.archive.org/web/20231016070730/www.eurofound.europa.eu/en/publications/2021/covid-19-pandemic-and-gender-divide-work-and-home>.
- Pomerantz, J. (2015). *Metadata*. MIT Press.
- Rhodes, T. (2013). *A Living, Breathing Revolution: How Libraries Can Use "Living Archives" to Support, Engage, and Document Social Movements*. IFLA WLIC.
- Ruest, N., Fritz, S., Deschamps, R., Lin, J., & Milligan, I. (2021). From archive to analysis: Accessing web archives at scale through a cloud-based interface. *International Journal of Digital Humanities*. Retrieved October 16, 2023, from https://web.archive.org/web/20231016071015/https://link.springer.com/article/10.1007/s42803-020-00029-6?wt_mc=Internal.Event.1.SEM.ArticleAuthorOnlineFirst&ArticleAuthorOnlineFirst_20210108#Sec7.
- Schafer, V., Musiani, F., & Borelli, M. (2016). Negotiating the web of the past. *French Journal for Media Research*, 6. Retrieved October 16, 2023, from <https://web.archive.org/web/20230918164203/http://frenchjournalformediaresearch.com/lodel-1.0/main/index.php?id=952>.
- Schafer, V., & Winters, J. (2021). The values of web archives. *International Journal of Digital Humanities*, 2, 129–144. Retrieved May 31, 2023, from <https://web.archive.org/web/20220507025030/https://link.springer.com/article/10.1007/s42803-021-00037-0>.
- UN Women. (2021). Explainer. How COVID-19 impacts women and girls. UN Women website. Retrieved October 16, 2023, from <https://web.archive.org/web/20230923145003/https://interactive.unwomen.org/multimedia/explainer/covid19/en/index.html>.
- WARCnet. (2024). WARCnet Papers and Special Reports. Retrieved February 24, 2024, from <https://web.archive.org/web/20240102160933/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.

12

CONVERSATION 2

Events

*Valérie Schafer, Saskia Huc-Hepher, Frédéric Clavert,
Helena Byrne, Juan-José Boté-Vericad, Sharon Healy,
Friedel Geeraert, Nadezhda Povroznik, Moisés Rockembach, and
Beatrice Cannelli*

What are the main challenges in establishing, supporting, and sustaining transnational collaborations related to the study of transnational events in web archives?

Valérie Schafer: We could undoubtedly find many challenges and sometimes obstacles to transnational collaborations in terms of legal, technical, or other matters. It is a fact that collections are not all organised in the same way, institutions do not all have the same flexibility with regards to access, to research support, etc., and one must also deal with the General Data Protection Regulation (GDPR and other regulations. But all these obstacles are surmountable. Currently, the main challenge, in my view, is to make these collaborations sustainable. However, this implies having (human and financial) resources, and the resources given to research are today oriented towards innovation at the expense of maintenance and capitalising on what already exists. We are dealing with short-term support; the collaborations are there, dynamic, intense, but they are supported for three to four years, and after that, a new project, a new theme must be deployed, and this often goes against the goal of sustainable collaborations.

Saskia Huc-Hepher: I completely agree. For a sustainable transnational collaborative model, we require a funding model(s) that sees beyond the short term and beyond the “shiny and new”.

What are the challenges in archiving crisis events?

Frédéric Clavert: I will answer this question as an historian. Based on my experience of collecting tweets on the COVID-19 crisis, the first challenge is to be able to rely on an infrastructure, that is material (*i.e.* no server, no archive), but also human. Archiving crisis events is strenuous, the human aspect is hence strategic.

Conversation 2: Events

*Helena Byrne,
Juan-José
Boté-Vericad,
Sharon Healy:*

Countries that do not have the necessary legal deposit legislation for archiving national web estates, have few options but to engage in collection development based on a fair use approach, or through permissions-based collection only. Permissions-based collection approaches present challenges for archiving crisis events on several levels. First, not all websites provide contact details and even if a contact is found there is no guarantee that a website owner will respond (Byrne, 2020; Ryan et al., 2022; Healy et al., 2022). Second, even when contact details are identified and permission requests are sent to website owners, there is a high failure rate (Byrne, 2020; Healy et al., 2022). Third, permissions-based approaches is a resource intensive process (Costa, 2021), and therefore at the time of crisis, some institutions simply do not have the resources to respond quickly enough to archive these events.

Friedel Geeraert:

I would like to add a point that Helena Byrne has mentioned during one of the meetings related to the present edited volume. It is also important to acknowledge the psychological effects of working on archiving crisis events and potentially traumatic content as it can contribute to mental health problems among staff members who are working on the collections.

I concur with Helena Byrne, Juan-José Boté-Vericad, Sharon Healy, and Frédéric Clavert that it is important to have a legal framework that allows crawling web content without having to obtain prior permissions from the right holders and also have access to existing web archiving infrastructure to get started immediately.

Another challenge is how to deal with an overload of content. The COVID-19 pandemic for example touched upon almost all aspects of our daily lives and therefore left many traces on the web. It is important to have a clear collection scope to navigate this challenge and also have enough human resources to dedicate to the selection and quality assurance of archived content. In addition, social media content poses a lot of technical problems when it comes to data capture and it is also difficult to exclusively scope in national social media content and not capture content from other countries that share the same language(s) for example. Lastly, dealing with fake news such as archived web content that promotes COVID-19 denial is another challenge: How can you present misinformation in a web archive with the necessary caveats for (future) users?

Nadezhda Povroznik:

Another challenge is to select and prioritise what sort of content will be selected for such a collection. In general, the selection and curation processes refer to a classical question of the value and significance of the resources. For example, the statistics about the COVID-19 spread, the medical-related information, reactions and opinions of the public and so on are greatly important for the subsequent analysis and reflection. However, for me, as a cultural historian, it is equally important to capture the cultural “layer” related to people’s perception of the crisis and reaction to related circumstances. There are many different examples including memes that were created

during the pandemic (and some of them went viral), artworks and performances, and exhibitions in the (virtual) museums that could be a valuable contribution to the COVID-crisis collection. There are so many curated collections by GLAM institutions, private initiatives and resources, and also specific projects, aimed at capturing the specific of COVID-19 crisis (for example, MuseiPandemic¹).

How well does the web archiving related to an unforeseen event like COVID-19 prepare web archives and researchers for collecting the next unforeseen transnational event?

Friedel Geeraert: The interviews done with web archivists (published as WARCnet papers) about the COVID-19 web archive collections have shown that in a number of institutions nominating web content for inclusion in the COVID-19 collections was a task that staff who normally work with physical objects could do during the confinement and telework. Should another pandemic occur, such practices can be more easily rolled out based on the experience gained during the COVID-19 pandemic.

Which methods are best suited to analyze the COVID-19 event? — and which particular skills are needed, if any?

Moisés Rockembach: To retrospectively analyse the COVID-19 event, it is crucial to examine how it was portrayed in the press and on social media, provided these data were also collected. Observing the trajectory of event coverage by press — how it intensified rapidly and then gradually diminished as vaccination coverage expanded and cases decreased — offers interesting insights. Additionally, the actions taken by governments in response to COVID-19 are represented in these data. Analysing and visualising data contained in web archives is key to gaining a better understanding of this period. Skills in data analysis, information visualisation, and an understanding of web archiving tools are essential for effectively carrying out this task, allowing us to reflect on the representation of and response to the event across various platforms and mediums.

What are the challenges in understanding the role that social media have played during the pandemic across countries?

Valérie Schafer: Analysing the role of social media networks in the context of COVID-19 presents multiple challenges, particularly because national institutions do not necessarily collect the same social networks and data, with the same frequency or regularity. While some focused on Twitter, others turned to Facebook or have explored Instagram, not to mention the audio-visual platforms. This results in extremely heterogeneous datasets, that are sometimes poorly representative, and that are difficult to compare quantitatively. Moreover, analysing social networks with the same rigour is challenging without a deep understanding of the specific platform or

Conversation 2: Events

network in question, as well as its users and audience. It is important to recognize that user behaviour can vary significantly across different countries.

Of course, it is still possible to identify general trends, for example, by focusing on a specific issue, or a hashtag, such as “lockdown”, but the researchers have to stay critical with the data and the limits of their results.

Frédéric Clavert: The main challenges in understanding the role that social media have played during the pandemic across countries are first to be able to collect those data — and since 2016, we all speak about an APiCalypse in the wake of the Cambridge Analytica Affair (the term “APiCalypse” describes the phenomenon of large web platforms closing off access to data to researchers) — and second, once data is collected, to be able to face its massiveness.

Beatrice Cannelli: I concur with Valérie Schafer’s point. Additionally, it is also important to consider that the platforms most commonly archived may not necessarily align with the most popular platforms at a national level and, especially, across countries. Therefore, it is crucial to consider that the type of conversations and individuals that are active, for example, on Twitter in a country where this platform is not widely used, may not fully represent the diversity of experiences and opinions of the wider public and its various groups and cultures. An example in this sense is probably the discrepancy existing between the role that platforms like TikTok played during the COVID-19 pandemic and the little amount of content actually archived due to a variety of legal and technical issues.

Note

- 1 Musei Pandemis. Retrieved April 4, 2024, from <https://web.archive.org/web/20240404131330/https://pandemis.fom.ru/>.

References

- Byrne, H. (2020, September 10). Launching the UK Web Archive 2020 Annual Domain Crawl [Blog post]. UK Web Archive Blog. Retrieved from <https://web.archive.org/web/20221001174748/https://blogs.bl.uk/webarchive/2020/09/launching-the-uk-web-archive-2020-annual-domain-crawl.html>.
- Costa, M. (2021). Full-Text and URL Search Over Web Archives. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 71–84). Springer.
- Healy, S., Byrne, H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., & Jansma, R. (2022). *Skills, Tools, and Knowledge Ecologies in Web Archive Research*. WARCnet Special Report. Retrieved from https://web.archive.org/web/20230902214716/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Skills_Tools_and_Knowledge_Ecologies.pdf.
- Ryan, M., Keating, D., & Finegan, J. (2022). Managing and accessing web archives: Irish practitioners’ perspectives. *AI & Society*, 37, 975–984. <https://doi.org/10.1007/s00146-021-01364-0>.

PART III

Methods and skills in web archive studies

13

INFORMATION ECOSYSTEMS THROUGH THE LENS OF WEB ARCHIVES

Matthew S. Weber

Understanding information ecosystems

Information ecosystems are a central component of modern society, and increasingly a critical component of our social fabric. In the past 50 years, however, these networks of media and other information sources have become incredibly complicated by the growth of digital media (Castells, 1999). Today, the concept of a global information ecosystem is sometimes used to refer to the interconnected nature of digital media today; elsewhere, this concept has been referenced as the networked information society (Castells, 2004). Regardless of the specific terminology, the broad concept references the notion that individuals produce and consume information in an increasingly complicated network of information sources.

The growth of digital technology has arguably democratized access to media globally (Bimber, 2003). In turn, several recent efforts have emerged seeking to better understand the scope and scale of today's information ecosystems. Non-profits and for-profits alike have made efforts dedicated to mapping information ecosystems. Internews¹ maps critical information ecosystems for the United Nations High Commission for Refugees. The Institute for the Future² has an ongoing effort to map social information ecosystems that impact individuals' well-being.

To better understand the shape of today's information society, these efforts are part of ongoing work to map and analyze the nature of information ecosystems, and are often transnational in nature as the examples demonstrate. By many accounts (e.g. Usher, 2023), however, extant efforts to map information ecosystems tend to be fragmented or capture a relatively narrow subset of media systems. In many cases, existing efforts capture a small fragment of a given ecosystem or capture a single topic. The scale of the global information society makes it nearly impossible to capture broader snapshots.

Despite the challenges of this scholarship, there is a pressing need to better understand our modern information ecosystems. Contemporary research on misinformation shows that to better identify sources of misinformation it is necessary to have a comprehensive understanding of the ecosystem within which misinformation is disseminated (Allen et al., 2020). A study of the dissemination of health-related misinformation found that the siloed nature of various mediums and various hubs makes it harder to ascertain the specific source

of erroneous health information (Chou et al., 2018). As rumors, fake news, hoaxes, and clickbait become more prevalent, the complex webs of misinformation become even more challenging to disentangle (Zannettou et al., 2019). Developing approaches to improve researchers' ability to map and analyze information ecosystems has the potential to make significant contributions to an array of scholarly inquiry.

This chapter focuses on the promise of web archives to support an emerging line of scholarly inquiry mapping and analyzing information ecosystems, which may be critical to examining the transnational flow of news media on the Web. Several different examples from contemporary research are used to demonstrate the importance of this research challenge, including two recent projects. The first study was a research initiative examining local media in 100 communities across the United States. The second study was a more focused effort to map the news ecosystem within a particular region of the United States. While these examples emphasize contexts in the United States, effort is taken to illustrate the potential for transnational scholarship. Further, this work points to similar efforts in related research to demonstrate the validity of this research approach. The following sections first explore the concept of an information ecosystem, and then entangle this concept with extant web archive research. Subsequently, specific examples from the authors' own work on the subject are used to illustrate some of the ways in which web archive research has the potential to unlock answers about the nature of our modern information ecosystems, especially with regard to better understanding the global flow of news media in a transnational media environment.

Information ecosystems

The concept of an information ecosystem is a polysemous term that has been used in a variety of contexts. In general, information ecosystems refer to systems of social media and online communities that interact in a common space. Studies of misinformation frequently use the term "information ecosystem" to refer to the space within which misinformation is disseminated (Lazer et al., 2018; Pertwee et al., 2022). For instance, an example of an information ecosystem could be the news media ecosystem in France (on a country-by-country level), or it could be bounded topically or contextually around information on a given topic, such as all the relevant information and websites produced about the Coronavirus disease 2019 (COVID-19). Some domains, such as journalism studies, have focused more explicitly on the concept of a media ecosystem. The primary difference is that information ecosystems focus beyond the specific context of news media.

Information ecosystems have come to refer to several different specific concepts, but broadly capture the notion of identifying and mapping information within a bounded and predefined ecosystem. For instance, the United Nations High Commission for Refugees (2020) uses the term in literature and through their reporting acknowledges the central role of information ecosystems in informing critical populations, as well as the importance of mapping information ecosystems. This work is used by the United Nations to understand the impact of information, the role of influencers within a network, and the development of social trust. Elsewhere, library and information scholars use information ecosystem as a term to capture the relevant scholarly domain within which information flows (Kuehn, 2023). In other examples, information ecosystems have been mapped to analyze policymaking processes (Weber & Yanovitzky, 2021), or to better understand the connect between information access and health outcomes (Tanner & Friedman, 2011).

Several examples exist demonstrating approaches to mapping information ecosystems. Anderson (2010) focused on a narrow but in-depth analysis of four communities to better understand the specific dynamics of a subset of community media outlets. Abernathy (2020) built a map of news outlets in the United States, and tracked it over a period of years to identify communities that lost their local news coverage. Indeed, several scholars have applied mapping of information ecosystems to better understand where gaps in media coverage maps exist (Ferrier et al., 2016). Elsewhere, scholars have used mapping tools in the context of misinformation. By mapping the information ecosystem of misinformation, it is then possible to track where misinformation spreads, and where it does not spread, to develop better approaches to combatting its dissemination (Shu et al., 2019; Waldrop, 2017).

Creating a bounded system

In modern media environments, information ecosystems are essentially boundless and global in nature; with a few notable exceptions, information generally flows across website and country boundaries (closed systems such as North Korea are notable examples of places where information does not flow as freely, although arguably digital technology has enabled some movement of information in those areas). Despite the interconnected, boundless nature of these systems, the scope and scale dictate a natural tendency to want to demarcate boundaries of some fashion.

Thus, one of the core questions in thinking about information ecosystems is how the boundaries of an ecosystem are defined, and how those boundaries are often particular to a given study or initiative. To this end, recent discussion of misinformation, for instance, has shifted discussions in research domains to developing a better understanding of the scope and scale of information ecosystems. As noted by Allen et al. (2020), consumers today increasingly navigate across a wide swath of information sources, traversing the web, but also accessing content on their mobile phones, from radio, and from television, among other sources. This means that in defining an information ecosystem it can be challenging to bound oneself to a given platform or medium; rather, it is increasingly critical to traverse across devices and mediums.

To categorize boundaries for information ecosystems, Anderson (2016) reviewed recent research and classified information ecosystems as either rhizomatic or environmental. The metaphor of rhizomatic boundaries focuses on media forms and outlets as organisms that grow and evolve over time. This metaphorical approach to conceptualizing an information ecosystem focuses on the notion that there is an equilibrium to which a system will return over time — the emphasis here is on information flows as opposed to entities. An environmental approach emphasizes the mapping of existing assets and the relationships between those assets. This approach calls for a cataloging of resources that exist within an information ecosystem. In other words, the rhizomatic approach focuses on the context and flow of information, whereas the environmental approach focuses on a cataloging of specific sources and resources. While Anderson demarcates between the two types, he also suggests that there may be an intersection that captures both the general context and the specific sources. Thinking of boundaries to an ecosystem in line with these dynamics provides a basis for examinations that cross national boundaries. This approach provides a metaphor for thinking about different strategies for capturing an information ecosystem, and in the context of this chapter, for thinking about potential strategies for web archiving.

Archiving and mapping a changing landscape

Considering the specific web context, Brügger (2018) observes that the web is a broader semiotic system of text; archived web data provides a record of the modern information society in a form that few researchers could have imagined in prior generations. Practical examples such as efforts by the U.S. Library of Congress as well as the British Library (and other national libraries), and work by the Institute for the Future demonstrates non-academic examples of organizations aiming to map information ecosystems for a variety of purposes — but also provide examples of organizations that develop their assessment of information ecosystems based on web archiving practices. Notably, corporate entities such as Facebook retain records of their media sources, although this information is rarely released for public or research consumption (Weber et al., 2019).

Whether or not the data used is referred to as an archive, in most cases researchers conducting this type of work utilize some form of archived digital information to support mapping efforts. Indeed, archived, digitized records have been critical to understanding global history since the late 1990s; omission of the web from that history would be an incredible gap in global history. More than history, however, web data today represent a fundamental part of the societal fabric, from information about health, communities, transportation, politics, and almost every other element of social interaction. Much like web archiving, attempts at information ecosystem mapping are variable; most efforts focus on evaluating text, images, and other forms of digital content (video, audio), but it would also be possible to examine the distribution of data formats or variations in broader topical themes.

Thus, capturing information about ecosystems requires researchers to surmount numerous challenges. Data aggregates quickly, and the velocity of information flows needs to be considered. Similarly, the dynamic nature of data flows requires researchers to think about temporality (Armstrong et al., 2019). Unlocking modern web archives is central to the ongoing success of efforts to map modern information ecosystems. In part, web archives generally afford researchers access to relatively coherent records of Web-based social activity, and in part, web archives provide a rare opportunity to examine the change of media over time (as opposed to cross-sectional samples, of more ad-hoc analyses, which may provide smaller samples of web data).

Recent advances in web archives and large-scale data science

Broadly, the term web archive is used herein to reference stored and preserved digital data, including records of web pages, digital media, and social media content. A persistent challenge of research in the context of web archives has been the problem of researcher access to data. As of 2021, the Internet Archive's web archive of international content contained approximately 45 petabytes of data — and the broader Internet Archive of film, music, games, and other digitized content is estimated to be nearly double that (Kramer, 2021), representing a broad record of transnational web content.

The sheer scale of web archives is well documented. According to archive.org, as of May 2023, there are more than 808 billion web pages archived within the Internet Archive's repositories. While the sheer scale is impressive, it creates a fundamental challenge for research. For perspective, as Milligan (2016) notes, GeoCities.com, one of the early web domains from the mid-1990s through the early 2000s, consisted of approximately 38 million web pages — a mere fraction of the current archive.

Information ecosystems through the lens of web archives

As Marciano et al. (2018) note, archivists have been building repositories of digital records for decades based on archival approaches. In recent years, however, the rise of computational methods has enabled new methods for accessing these troves of data. While some online platforms are increasingly difficult to access (e.g. social media), web archives are powerful tools for examining transnational social and cultural activity online (Hockx-Yu, 2014). As increased attention is focused on the issue of the information ecosystem, it is reasonable to expect more scholarly attention to focus explicitly on mapping extant information sources to better understand a given ecosystem.

Understanding breadth and depth

As noted, information ecosystems can be defined based on context or a catalog of sources. Arguably, even if one seeks to identify information based on context, it is still necessary to identify the relevant information sources. Thus, a starting point in mapping information ecosystems is the identification of a comprehensive list of relevant entities (e.g. geographic regions, web domains, media types), but this is hardly an easy task.

Consider, for example, the challenge of creating a comprehensive map of media outlets in each country or a given state. For the most part, lists of this type are not preexisting. There are pay-to-play services such as Meltwater, Nielsen, and Cision which maintain listings of media outlets in the United States, but these lists are often incomplete and generally focus on the largest media outlets in each region. Niche and hyperlocal outlets are often overlooked. Further, the lists maintained by these information sources are often outdated and rarely account for newer information sources such as social media sites and mobile applications.

Further, some critical community information sources, such as private Facebook groups and community information sources such as Nextdoor.com, can be hard to search making them hard to detect and identify. Thus, it is often left to the individual researcher to cobble together a list of relevant outlets as a starting point for identifying a given information ecosystem (Stonbely, 2023). While the focus is often on a community or region, the specific boundaries matter little. It is easy to conceptualize a mapping process that captures information flowing across transnational borders.

From scale to practice

Once a researcher identifies the relevant information sources, the next challenge becomes the collection and preservation of the desired content. As researchers, we are rarely able to collect data and simultaneously analyze those data. Thus, even if storage is temporary a researcher will need to create an archive of some kind to preserve information in support of future research. Web archiving is a relatively standardized practice for collecting and storing a wide array of digital content.

The field of web archiving has grown dramatically over the past two decades. The standardization of web ARChive (WARC) files in 2009 created a common format for the storage and preservation of archived web pages. WARCs are now recognized as an International Organization for Standardization (ISO) standard file format, with data stored in a JSON format that can easily be ingested into scripts in common languages such as Python and R. Several custom packages such as the WebRecorder Player and OpenWayBack have been created to enable users to read WARC files.

The volume of archived web data available means that even with a standardized format issues of access and scalability still abound. Initiatives such as Archives Unleashed have created robust libraries of code to support broader researcher access to web archives. The Archives Unleashed Project (Ruest et al., 2022) helped to create web archive search and data analysis tools to enable a wide array of inquiries, including text analysis and social network analysis. A core component of the toolset was a cloud-based set of analytical tools that allow researchers to import, analyze, and extract metadata from subsets of archived web data (Ruest et al., 2021; Schafer et al., 2024). As of 2023, the toolkit created by the Archives Unleashed Project has been integrated directly into the Internet Archive's umbrella of research services — under the Archive-It platform as the Archives Research Compute Hub (ARCH).

A critical outcome of the Archives Unleashed project and the incorporation of the Archives Unleashed toolkit into the Internet Archive's network, is the development of a workflow that establishes a process for moving from the scale of archived Internet data to a scale achievable by a broader network of researchers. Specifically, extant researcher activities working with archived Internet data suggest a process of filtering, extracting, aggregating, and visualizing to enable scholarly inquiry (Ruest et al., 2020).

Archived media

News media, from radio to television to newspapers, has consistently been a core component of a variety of web archives. In part, news media is frequently archived because it tells an important component of our societal narrative. Gaff (2017) referred to digital news archives as the archeological survey site for the twenty-first century.

In early work within this space (Weber & Napoli, 2018), researchers demonstrated that web archives could be used to analyze the health of local news within communities. In that work, the researchers created a custom archive to track the local news production in 100 communities across the United States. That work focused on identifying a random sample of communities and performing a stepwise web search of each community to identify the relevant local news outlets. That search was relatively restricted by its focus on news media within a community's geographic boundaries. Some 733 outlets were identified as relevant to the 100 communities; those outlets were then crawled over a period of months to create a comprehensive sampling of local news content. Even that limited scope generated a web crawl that captured 1.6 million documents and 2.2 terabytes of data. Weber and Napoli (2018) note the potential of web archiving techniques as a means for better understanding the diversity of information within a community or set of communities, but the focus on sampling and the bounded definition of an ecosystem within a given town creates some limitations.

Building a seed list

A common challenge is the construction of a seed list of URLs as a starting point for building a web archive — or for retrieving data from established archives. For instance, the Chronicling America project at the Library of Congress contains a vast record of news content, but nevertheless, a seed list is still needed to determine the appropriate data for analysis. Thus, when seeking to map an information ecosystem, a core challenge remains identifying the relevant resources for analysis (what has previously been referenced as a

“register” of websites (Brügger, 2011)). To that end, Stonbely (2023) conducted research to construct a list of local news outlets within the state of New Jersey. That work included a multipronged effort to identify local newspapers, radio stations, television stations, and minority publications, using web search (Google, Bing, etc.) and records requests from the Federal Communications Commission API. The resulting list includes 779 media outlets: 308 newspapers, 276 websites, 100 television stations, 66 radio stations, and 29 magazines. This work established a clear workflow for cataloging information sources within a given resource space and provides a starting point for mapping an information ecosystem. This challenge is notably more complicated in a transnational setting.

One challenge in building seed lists continues to be the issue of identifying relevant social media sources (Gulyas et al., 2019). While many consumers of local news access their local outlets via social media, much of this information dissemination occurs through community groups on social media outlets such as Facebook and Nextdoor.com. A minimal approach is to identify a minimal subset of local social media sources to include in each seed list. This could occur through targeted searches of social media outlets (Ogden, 2022), particularly as API access becomes increasingly limited. Thus, merging the approach established by Stonbely with ancillary and complementary data on social media provides a path for building a comprehensive list that represents a given information ecosystem.

Analyzing data

The ability to extract metadata from WARC files is just one core advantage that enables researchers to advance important avenues of social inquiry using archived web data. For years, one of the central problems has been that secondary analysis remains difficult even if one could extract the necessary metadata from web archive files. Secondary analysis references the examination of the extracted metadata, as well as subsets from archives, including the analysis of large sets of images and audio files.

There are some graphic interfaces that allow researchers to visually analyze and review archive files. The Internet Archive’s Archive-It platform, for instance, allows users to review their crawls in a format that is nearly identical to playback on the Wayback Machine. The WebRecorder project, for instance, continues to support the playback of single WARC files or multiple WARC files across its suite of tools. Similarly, the SolrWaybackMachine³ is a powerful tool for creating custom visualization packages and for analyzing archives.

Analysis at scale, however, continues to require some understanding of computer programming. Programming languages such as R and Python are generally considered easier to learn than more traditional computer programming languages, and multiple recent packages in both languages support access to and analysis of WARC files. For instance, the Sparkwar package (Luraschi, 2022) was designed to allow individuals to access WARC files from Common Crawl and to analyze those archive files in R (all code packages current as of 2023). Several packages in Python offer similar functionality, including the Python packages warc, pylibwarc, and warcio (see further work by Bevendorff et al., 2021). Leveraging these types of tools, it is then possible to build new applications. For instance, Khamis et al. (2023) built a tool that helps with building web archives around crisis events via Python. Another advantage of working in R and Python is that once data are extracted from the WARC file format into other data structures (such as large-scale databases and data tables) there are a multitude of additional packages that can be used for subsequent analysis.

Transnational information ecosystems

There are several extant examples of transnational efforts at mapping information ecosystems which exemplify the varied ways in which this approach intersects with ongoing web archiving initiatives such as WARCNet. In a joint initiative, the Internet Archive, JISC, and the UK Web Archive partnered to create a UK Web Domain Archive. This type of aggregate dataset provides a robust starting point for information ecosystem mapping. Scholars were able to use this work to map the topology of the British Web and to show connections to a variety of outcomes, including regional productivity and Internet adoption (Tranos, 2013; Tranos et al., 2021). In a more niche example, interview data was used in complement to web data to create a more specialized analysis of the changes in local news in the UK (Firmstone, 2016). Ongoing efforts in France similarly demonstrate attempts within the EU to map and assess local information ecosystems through the systematic collection of local news content (Delaunay, 2008). While quite varied in nature, there is clearly a transnational interest in the ongoing assessment of and mapping of local information ecosystems. The challenge of moving from collecting to mapping to analyzing remains a significant hurdle.

Mapping at scale: A regional example from New Jersey

The work by Stonbely (2023) established a seed list for New Jersey and created a baseline for mapping information contained within a given region. An artificial boundary exists in that the focus of this example is all information resources within a state's borders, but the scale is such that this example extends well beyond a single town or community and may serve as a proof of concept for expanding collection efforts further. Thus, that list was used as a starting point for data collection; of the initial 779 media outlets, 100 did not have crawlable URLs at the time that this proof of concept was initiated in 2019. The initial work by Stonbely and her team was conducted in late 2017 and early 2018, and even in that period, several URLs had become defunct. In addition, from a crawling perspective, some URLs were contained within a common domain, such as Patch.com. Thus, while the number is lower, the project largely captured the same websites as the seed list.

The crawl that supports this research was performed from September 2019 through December 2020, and included the home page of every URL in the seed list as well as each page linked to the home page (a crawl depth of one). The crawling occurred once a month, given four-time snapshots of the targeted URLs. The crawl generated 4.7 million individual archived web pages and 30 million documents (text, video, images, etc.).

In the case of this news ecosystem mapping, the initial extraction of metadata was conducted using Archive-It's ARCH services, but a similar output could have been obtained using Python locally and processing the data using existing scripts as noted above. Admittedly, the compute time would have been longer, and would likely require the use of a pay-for-use service such as Amazon Web Services Elastic Compute (EC2) to process the data. Broadly, Archive-It is the Internet Archive's partner program that supports custom web archiving on the part of libraries, researchers, and other community partners (Jones et al., 2018). The Archives Research Compute Hub was designed to allow for a range of analytical actions — namely the generation of datasets and visualizations of data (Holzmann et al., 2022). In this case, ARCH services were used to reduce the WARC files to the text contained within HTML documents — producing something akin to a database of the textual content of the web pages.

Secondary analysis of web pages

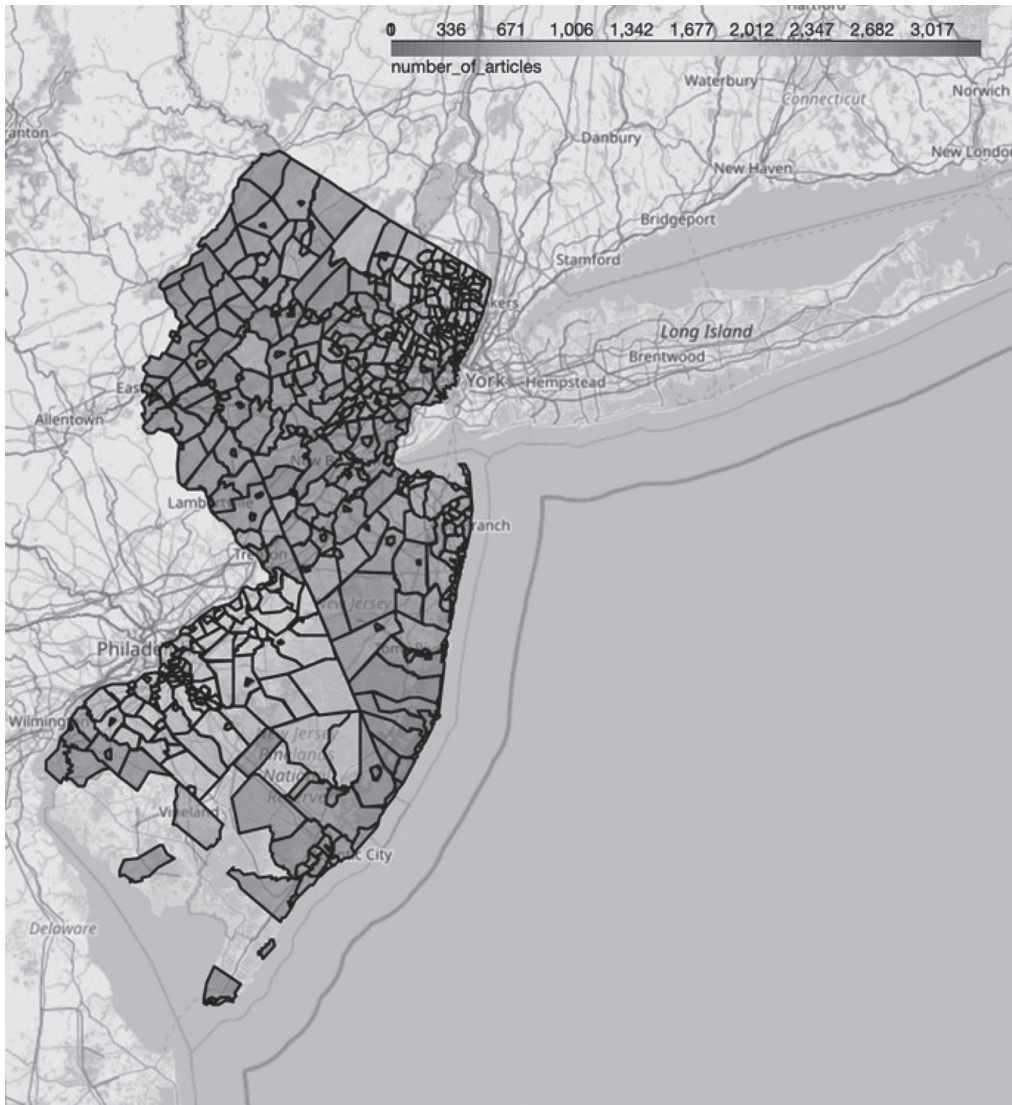
Given a set of WARC documents, or more broadly, given the set of documents that constitutes the archive used in this project, there are countless different analyses that one could perform (inter-organizational networks, an analysis of themes in images, an examination of media coverage of key events, etc.). The goal of this study was to map, as accurately as possible, the media information ecosystem of New Jersey as it stood in late 2019 and early 2020. Thus, the focus of this work was on examining the news articles published by television stations, radio stations, newspapers, and websites.

Thus, to capture the news articles, the Python package Newspaper3k is used to extract news-focused text from the raw ARCH text output. Newspaper3k (Ou-Yang, 2021) is a package that has been developed in Python to assist with the scraping and curation of news articles. In prior research, Newspaper3k has been used to examine a variety of media-oriented research questions, such as an examination of how newspaper articles are spread on Twitter (Brena et al., 2019) and to examine how health misinformation was disseminated in the media during the heart of the COVID-19 pandemic (Cui & Lee, 2020). The package was designed to crawl newspaper websites and to save the text of newspaper articles, but specific components of the package can be used to examine text and to extract newspaper content from raw text. The initial filtering from ARCH to reduce the text from WARC files to raw text resulted in 514,620 candidate blocks of text. Further review and analysis using Newspaper3k reduced the count to 117,524 newspaper articles covering four months of media coverage within New Jersey.

From news articles to maps

With a corpus of news articles in hand, the next step in the research was to move from textual articles to a mapping of those articles. The scale of data was still relatively large at this point, which meant that much of the analysis could not be done on a laptop computer. The data were moved to an Amazon Web Services Elastic Compute instance to help speed up the work. At this point, Python was used to clean the text — specifically the pandas package for cleaning text and preparing a corpus for textual analysis (McKinney, 2011). The cleaned data was then stored in a database format matching text to the respective media outlets. Finally, BERTopic modeling was used to classify stories. BERTopic is a variant of neural network topic modeling that is particularly effective in classifying text (Grootendorst, 2022); we trained the data to classify based on predefined terminology. Following earlier work on news media mapping, we used the idea of “critical information needs” to classify the news articles in our dataset (Friedland et al., 2012; Napoli et al., 2018). The primary critical information needs of a community are emergencies and risks, health and welfare, education, transportation, economic opportunities, the environment, civic information, and political information.

Stories were classified into a single best-fitting category. The lead researcher and an undergraduate research assistant hand-verified 400 news articles and agreed with the machine learning coding 89% of the time. Finally, news articles were matched back to the media outlets that had published the initial article. For each media outlet, the geographic region was coded based on the coverage area of that media outlet. For instance, if a newspaper listed its coverage area as covering five towns, the stories for that outlet were geocoded to those five towns. Geocoding was done using the Google Maps API.



Map 13.1 Cumulative media information ecosystem in New Jersey.

With those data in hand, and the processing complete, it was possible to generate a map of the information ecosystem within New Jersey. As noted in the prior review, it is necessary when mapping to make key decisions about what to map and how. In this case, the mapping was bounded by the state of New Jersey, and the content was categorized based on critical information needs. Map 13.1 shows the cumulative map of news content within the state.

The map shows the intensity of news coverage over the four-month period, with the darker red colors indicating more intense media coverage. Slices of the map were created for each of the critical information needs categories and additional analyses were performed to

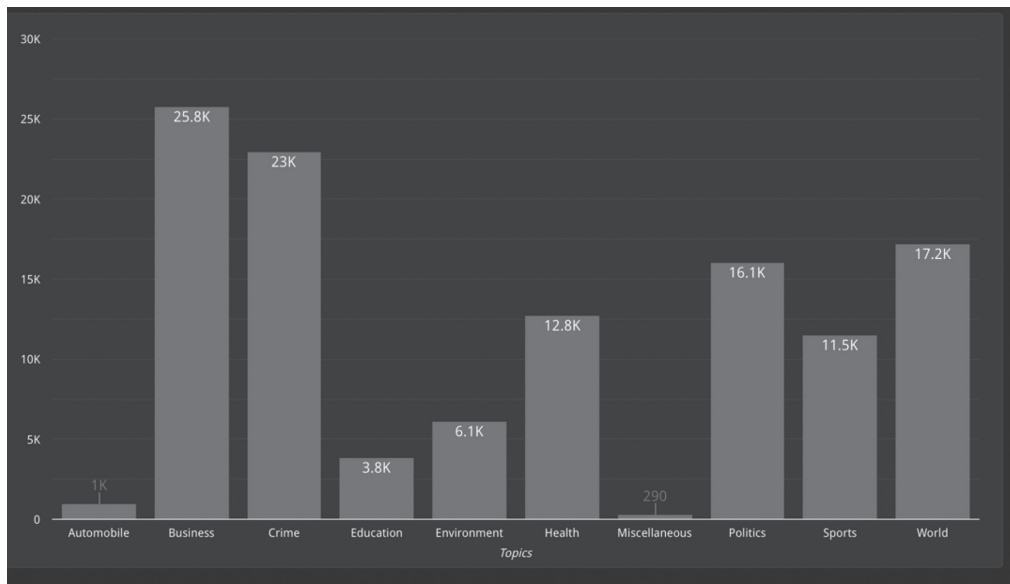


Figure 13.1 Breakdown of news articles by critical information need.

look at the breakdown of content. Figure 13.1 gives a count of news articles by critical information need category to show how news articles were categorized.

In part, the breakdown provides a validity check. The categorization based on critical information needs was validated against a national sample, which shows that the breakdown was roughly equivalent to the national distribution with some minor discrepancies (Napoli et al., 2018).

Limitations and persistent challenges

This chapter highlights the potential of research such as the WARCNet initiative to help fuel the growth of information ecosystem mapping initiatives. This work has the potential to help address critical questions of our time, such as better understanding the scope and scale of misinformation, as well as the spaces within our information ecosystems most susceptible to misinformation attacks. Simultaneously, this work points to several lingering challenges with regard to information ecosystem mapping. Specifically, ongoing challenges remain with regard to how information ecosystems are defined, in defining what is included and excluded from a given archive, and in gaining access to the right data in the right format. Each of these persistent challenges is outlined in the following.

Defining the information ecosystem

A central challenge outlined at the beginning of this chapter was the issue of how an information ecosystem is defined. In particular, the metaphor of rhizomatic versus environmental approaches was utilized. This case study took an environmental approach, capturing everything available in the news media space within a given state. The challenge of this

environmental approach is that the emphasis is placed on an aggregate analysis of what entities constitute a recourse space. On the other hand, a rhizomatic approach would emphasize the flow of information, and the mapping of the entities between which that information flows. For example, consider the context of the International Internet Preservation Consortium's efforts to map websites related to COVID-19. The creation of the archive, as discussed in this volume focused on the archiving of websites related to the creation of content pertinent to COVID-19 (Bingham et al., 2024). This more rhizomatic approach emphasized coverage of a given topic, and could, in the context of information ecosystem mapping, include websites that would be missed otherwise. This could include social media outlets and niche community websites that provide community news.

The transnational potential

National borders have often been demarcated by boundaries drawn on maps. But this work points to the analysis of information ecosystems as a means of understanding how information is made available within communities and regions, and how information provided by a given organization may extend beyond boundaries. For instance, the geolocation of stories published in an archive of the BBC would quickly reveal that information disseminated by the outlet has extensive reach across most of the United Kingdom, the European Union, and beyond. In an extension of the ecosystem mapping demonstrated in the preceding, researchers extended the approach to a snapshot of media in southern Australia. The findings illustrate how content produced by local Australian media organizations covered the local region, but also frequently mentioned locations across Europe and Asia. There are clearly methodological challenges that must be addressed, but there is also significant potential to better understand the transnational nature of Web archives.

Equity and inclusion of mapping approaches

The issue of the approach to information ecosystem mapping suggests that the choice of approach may result in the omission of certain critical information sources. Whether using a rhizomatic or environmental approach, a core challenge remains that certain types of websites may remain disconnected from the core of the Web. In taking an environmental approach, this means that a researcher is unlikely to find these websites listed in a traditional database. Taking a rhizomatic approach means that research is unlikely to find these websites by following the extant flow of information. For instance, ethnic minority websites can be challenging to locate and identify as they are often disconnected from the majority of web content (Gowricharn & Elahi, 2020). Certain subgroups can also fall to the periphery of the web, as noted in a recent study of female football fan groups on Tumblr (Kunert, 2021). One approach to counter this issue is the use of community informants in focus groups or interviews to elicit information about local sources that may otherwise be overlooked. Such approaches may also help to bring to light keywords that can be used to expand a search for relevant information.

Challenges of data and access

A final challenge brought to light in this case study is the issue of data and data access. The case study presented in this chapter focused on an example where existing archived data were

unavailable, and the research team had the opportunity to create a new data collection based on a customized list of information sources. In cases where a national archiving mandate exists, or national libraries have created large collections, data may be available, but access can still be an issue. For instance, the British Library and the Bibliothèque nationale de France have collections available, but in some cases those collections are restricted to local or on-site access. Further, even where access is readily available, the WARC file format can be difficult to analyze. The JSON format requires substantial additional coding to extract information that can be utilizable in research.

Implications and future questions

This chapter set out to examine the challenge of mapping information ecosystems as a means of understanding transnational media systems through an illustration from a regional case study. Across research domains, from health to misinformation, to social media, to news media, understanding and examining the breadth and depth of available information within a community remains a critical challenge for researchers across domains as well as national borders. This chapter outlined some of the key issues associated with mapping an ecosystem, from the scale of data to the challenges of scoping the analysis to capture the correct resources, to the final classification of data and subsequent analysis.

The closing sections of this chapter give a proof of concept with the mapping of news content in New Jersey, demonstrating the potential to move from web archiving through to content mapping and analysis. The analysis is not without limitations; for instance, the work illustrated above focused on online content online, meaning that if a story was broadcast on a radio station but not republished to the website then it was not possible to collect the data in this analysis. Further, the literature review outlined some of the challenges of collecting social media data; a potential solution for sampling was outlined, but that solution was not used in the analysis. Moreover, the perspective here is limited to a small region in the United States but provides a path for looking forward to the creation of transnational maps.

There are broader issues that must be addressed as this type of research becomes more critical, and as digital information ecosystems become more complicated. We need to better understand how we incorporate mobile and application-based domains, and how we track data from those domains. We need to develop better methods for integrating data from multiple archives; for example, television content and printed newspaper coverage often exist across different archives. Stitching these data together across these domains remains a lingering challenge.

Further, scale and accessibility remain an issue. For instance, while the Archive-It's ARCH research support services enable access to a broader array of data, and support ongoing scholarly inquiry, the scale of the data remains significant. The foundational work of the Archives Unleashed Project notes that during the first few years of the project, more than 200 researchers accessed the toolset analyzing more than 1000 datasets (amounting to 300 terabytes of web archive data). The sheer volume continues to be problematic, and while search tools are improving, many types of scholarly inquiry would still be lost in the sheer mass of a web archive.

Notes

- 1 Additional information about Internews can be found at the organization's main webpage <https://web.archive.org/web/20240328060412/https://internews.org/>. Internews partners with local communities to support the production of local news.
- 2 More information about the Institute for the Future (ITF) can be found at the organization's main webpage <https://web.archive.org/web/20240320223721/www.iftf.org/>. ITF is a think tank and policy group focused on research and education.
- 3 Information about SolrWaybackMachine can be found on the GitHub page for the project: <https://web.archive.org/web/20240305025759/https://github.com/netarchivesuite/solrwayback>.

References

- Abernathy, P. M. (2020). *News deserts and ghost newspapers: Will local news survive?* University of North Carolina Press.
- Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14), eaay3539. <https://doi.org/doi:10.1126/sciadv.aay3539>.
- Anderson, C. W. (2010, 2010/08/04). Journalistic networks and the diffusion of local news: The brief, happy news life of the "Francisville Four". *Political Communication*, 27(3), 289–309. <https://doi.org/10.1080/10584609.2010.496710>.
- Anderson, C. W. (2016). News ecosystems. In T. Witschge, C. W. Anderson, D. Domingo, & A. Hermida (Eds.), *The SAGE handbook of digital journalism* (pp. 410–423). SAGE.
- Armstrong, M. P., Wang, S., & Zhang, Z. (2019, 2019/01/02). The Internet of Things and fast data streams: Prospects for geospatial data science in emerging information ecosystems. *Cartography and Geographic Information Science*, 46(1), 39–56. <https://doi.org/10.1080/15230406.2018.1503973>.
- Bevendorff, J., Potthast, M., & Stein, B. (2021). FastWARC: Optimizing large-scale web archive analytics. *arXiv preprint arXiv:2112.03103*.
- Bimber, B. (2003). *Information and American democracy: Technology and the evolution of political power*. Cambridge University Press.
- Bingham, N., de Wild, K., Geeraert, F., & Nyvang, C. (2024). Surveying the landscape of COVID-19 web collections in European GLAM institutions: An explorative analysis. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge companion to transnational web archive studies* (pp. xxx–xxx). Routledge.
- Brena, G., Brambilla, M., Ceri, S., Di Giovanni, M., Pierri, F., & Ramponi, G. (2019, 07/06). News sharing user behaviour on Twitter: A comprehensive data collection of news articles and social interactions. *Proceedings of the international AAAI conference on web and social media*, 13(01), 592–597. <https://doi.org/10.1609/icwsm.v13i01.3256>.
- Brügger, N. (2011). Digital history and a register of websites: An old practice with new implications. In D. W. Park, N. W. Jankowski, & S. Jones (Eds.), *The long history of new media: Technology, historiography, and contextualizing newness* (pp. 283–298). Peter Lang.
- Brügger, N. (2018). *The archived web: Doing history in the digital age*. MIT Press.
- Castells, M. (1999). *Information technology, globalization, and social development*. UNRISD Discussion Paper. United Nations Research Institute for Social Development.
- Castells, M. (2004). *The network society: A cross-cultural perspective*. Edward Elgar.
- Chou, W.-Y. S., Oh, A., & Klein, W. M. P. (2018). Addressing health-related misinformation on social media. *JAMA*, 320(23), 2417–2418. <https://doi.org/10.1001/jama.2018.16865>.
- Cui, L., & Lee, D. (2020). Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.
- Delaunay, E. (2008). Newspaper digitization in France: Developments and strategies. *IFLA Publications*, 133, 223.
- Ferrier, M., Sinha, G., & Outrich, M. (2016). Media deserts: Monitoring the changing media ecosystem. In M. Lloyd & L. Friedland (Eds.), *The communication crisis in America, and how to fix it* (pp. 215–232). Springer.

- Firmstone, J. (2016, 2016/10/02). Mapping changes in local news. *Journalism Practice*, 10(7), 928–938. <https://doi.org/10.1080/17512786.2016.1165136>.
- Friedland, L., Napoli, P., Ognyanova, K., Weil, C., & Wilson III, E. J. (2012). *Review of the literature regarding critical information needs of the American public*. Communication Policy Research Network (CPRN).
- Gaff, D. H. (2017, 2017/06/01). Extra! Extra! Read All About It: Newspaper archives as archaeological site survey. *Journal of Archaeological Method and Theory*, 24(2), 451–465. <https://doi.org/10.1007/s10816-016-9273-3>.
- Gowricharn, R., & Elahi, J. (2020). Digital institutions: The case of ethnic websites in the Netherlands. *Identities*, 27(4), 442–461.
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Gulyas, A., O'Hara, S., & Eilenberg, J. (2019, 2019/10/03). Experiencing local news online: Audience practices and perceptions. *Journalism Studies*, 20(13), 1846–1863. <https://doi.org/10.1080/1461670X.2018.1539345>.
- Hockx-Yu, H. (2014, 2014/08/01). Access and scholarly use of web archives. *Alexandria*, 25(1–2), 113–127. <https://doi.org/10.7227/ALX.0023>.
- Holzmann, H., Ruest, N., Bailey, J., Dempsey, A., Fritz, S., Lee, P., & Milligan, I. (2022). ABCDEF: The 6 key features behind scalable, multi-tenant web archive processing with ARCH: Archive, big data, concurrent, distributed, efficient, flexible. *Proceedings of the 22nd ACM/IEEE joint conference on digital libraries*. IEEE. New York, NY.
- Jones, S. M., Nwala, A., Weigle, M. C., & Nelson, M. L. (2018). The many shapes of Archive-It. *arXiv preprint arXiv:1806.06878*.
- Khamis, M., Chen, J., Lee, J., Averyt, M., & Mitta, N. (2023). Crisis events webpages archiving. <http://hdl.handle.net/10919/114964>.
- Kramer, A. (2021). The internet is splitting apart: The Internet Archive wants to save it all forever. *Protocol*. www.protocol.com/internet-archive-preserving-future.
- Kuehn, E. F. (2023, 2023/04/01). The information ecosystem concept in information literacy: A theoretical approach and definition. *Journal of the Association for Information Science and Technology*, 74(4), 434–443. <https://doi.org/10.1002/asi.24733>.
- Kunert, J. (2021). The footy girls of Tumblr: How women found their niche in the online football fandom. *Communication & Sport*, 9(2), 243–263. <https://doi.org/10.1177/2167479519860075>.
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094–1096. <https://doi.org/doi:10.1126/science.aao2998>.
- Luraschi, J. (2022). *sparkwarc: Load WARC files into Apache Spark*. GitHub.
- Marciano, R., Lemieux, V., Hedges, M., Esteva, M., Underwood, W., Kurtz, M., & Conrad, M. (2018). Archival records and training in the age of big data. In J. Percell, L. C. Sarin, P. T. Jaeger, & J. C. Bertot (Eds.), *Re-envisioning the MLS: Perspectives on the future of library and information science education* (Vol. 44B, pp. 179–199). Emerald Publishing Limited. <https://doi.org/10.1108/S0065-28302018000044B010>.
- McKinney, W. (2011). pandas: A foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, 14(9), 1–9.
- Milligan, I. (2016, 2016/03/01). Lost in the Infinite Archive: The promise and pitfalls of web archives. *International Journal of Humanities and Arts Computing*, 10(1), 78–94. <https://doi.org/10.3366/ijhac.2016.0161>.
- Napoli, P., Weber, M. S., & McCollough, K. (2018). *Assessing local journalism: News deserts, journalism divides, and the determinants of the robustness of local news*. News Media Measures Project. https://dewitt.sanford.duke.edu/wp-content/uploads/2018/08/Assessing-Local-Journalism_100-Communities.pdf.
- Ogden, J. (2022, 2022/04/03). “Everything on the internet can be saved”: Archive Team, Tumblr, and the cultural significance of web archiving. *Internet Histories*, 6(1–2), 113–132. <https://doi.org/10.1080/24701475.2021.1985835>.

- Ou-Yang, L. (2021). Newspaper3k: Article scraping & curation — newspaper 0.0.2 documentation. Available on <https://github.com/codelucas/newspaper>. Accessed on 9th September.
- Pertwee, E., Simas, C., & Larson, H. J. (2022, 2022/03/01). An epidemic of uncertainty: Rumors, conspiracy theories and vaccine hesitancy. *Nature Medicine*, 28(3), 456–459. <https://doi.org/10.1038/s41591-022-01728-z>.
- Ruest, N., Fritz, S., Deschamps, R., Lin, J., & Milligan, I. (2021, 2021/11/01). From archive to analysis: Accessing web archives at scale through a cloud-based interface. *International Journal of Digital Humanities*, 2(1), 5–24. <https://doi.org/10.1007/s42803-020-00029-6>.
- Ruest, N., Fritz, S., & Milligan, I. (2022, 2022/09/02). Creating order from the mess: Web archive derivative datasets and notebooks. *Archives and Records*, 43(3), 316–331. <https://doi.org/10.1080/23257962.2022.2100336>.
- Ruest, N., Lin, J., Milligan, I., & Fritz, S. (2020). The archives unleashed project: Technology, process, and community to improve scholarly access to web archives. *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, Virtual Event, China. <https://doi.org/10.1145/3383583.3398513>.
- Schafer, V., Aasman, S., Clavert, F., de Wild, K., & Sirajzade, J. (2024). The challenges of searching for women in the COVID-19 web archive collections: Promises, achievements, and pitfalls. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge Companion to transnational web archive studies* (pp. xxx–xxx). Routledge.
- Shu, K., Mahudeswaran, D., & Liu, H. (2019, 2019/03/01). FakeNewsTracker: A tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*, 25(1), 60–71. <https://doi.org/10.1007/s10588-018-09280-3>.
- Stonbely, S. (2023). What makes for robust local news provision? Structural correlates of local news coverage for an entire U.S. state, and mapping local news using a new method. *Journalism and Media*, 4(2), 485–505. www.mdpi.com/2673-5172/4/2/31.
- Tanner, A., & Friedman, D. B. (2011). Health on the Web: An examination of health content and mobilising information on local television Websites. *Informatics for Health and Social Care*, 36(1), 50–61.
- Tranos, E. (2013). *The geography of the internet: Cities, regions, and internet infrastructure in Europe*. Edward Elgar Publishing.
- Tranos, E., Kitsos, T., & Ortega-Argilés, R. (2021, 2021/12/02). Digital economy in the UK: Regional productivity effects of early adoption. *Regional Studies*, 55(12), 1924–1938. <https://doi.org/10.1080/00343404.2020.1826420>.
- United Nations High Commission for Refugees. (2020). *Collaboration for connectivity*. www.unhcr.org/innovation/wp-content/uploads/2020/06/Collaborating-for-Connectivity_WEB062020.pdf.
- Usher, N. (2023). The real problems with the problem of news deserts: Toward rooting place, precision, and positionality in scholarship on local news and democracy. *Political Communication*, 40(2), 238–253. <https://doi.org/10.1080/10584609.2023.2175399>.
- Waldrop, M. M. (2017). The genuine problem of fake news. *Proceedings of the National Academy of Sciences*, 114(48), 12631–12634.
- Weber, M., Andringa, P., & Napoli, P. (2019). *Local news on Facebook: Assessing the critical information needs served through Facebook's TodayIn Feature*. University of Minnesota. <https://web.archive.org/web/20240404145123/https://hsjmc.umn.edu/sites/hsjmc.umn.edu/files/2019-09/Facebook%20Critical%20Information%20Needs%20Report.pdf>.
- Weber, M. S., & Napoli, P. M. (2018). Journalism history, Web archives, and new methods for understanding the evolution of digital journalism. *Digital Journalism*, 6(9), 1186–1205.
- Weber, M. S., & Yanovitzky, I. (2021). Knowledge brokers, networks, and the policymaking process. In M. S. Weber & I. Yanovitzky (Eds.), *Networks, knowledge brokers, and the public policymaking process* (pp. 1–25). Palgrave.
- Zannettou, S., Sirivianos, M., Blackburn, J., & Kourtellis, N. (2019). The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality*, 11(3), Article 10. <https://doi.org/10.1145/3309699>.

14

HISTORY OF VIRTUAL MUSEUMS AND WEB ARCHIVES

Opportunities for rescaling research

Nadezhda Povroznik

Introduction

Invented about 30 years ago, the World Wide Web brought museums a new environment to expand their activities and engage audiences. Museums underwent a significant transformation in the way they created and delivered digital content using their websites and communicating with communities online (Haitham, 2019). The coronavirus disease 2019 (COVID-19) pandemic outbreak and the series of subsequent national lockdowns announced brought an additional impetus to the development of digital initiatives by museums (Giannini & Bowen, 2022). To better understand the current state of the museums and future developments, it is important to trace the history of their online presence, and identify changes and dominant trends. Historical perspective is also important due to the fragile nature of born-digital heritage (Schafer & Winters, 2021) and the specifics of national programs of web preservation. Also, the transnational study of museums' websites has practical implications for web archives to refine preservation approaches to effectively safeguard born-digital cultural heritage (Ainsworth et al., 2015).

Several studies have looked into the history of museums on the web from different perspectives. The early stages of their development were considered in the framework of the transition from analog to digital presence (Povroznik, 2020b). By examining the patterns specific to museums' websites from the 1990s to the early 2000s, it was possible to show how web technology and innovative modes of public involvement were accepted by museums (Gaia et al., 2020) and how the emerging technologies were implemented in museums of different types (Bowen et al., 2023). Also, changes in audience engagement were a focus of research (Perera & Chandra, 2010), and the development of interactivity of museums' web resources (Fotakis & Economides, 2008; Arthur, 2018). However, bringing attention to the development of these websites in relation to a complex of museums' activities is needed for a better understanding of how the online and onsite activities can be aligned. Achieving this can be facilitated by employing mixed methods (Creswell & Creswell, 2018), integrating both qualitative and quantitative approaches. In this regard, analyzing web archives becomes

particularly valuable, as it allows researchers to leverage between close-reading techniques and more data-driven methodologies, thereby rescaling the scope and depth of research in the field. Additionally, historical analysis of the web-archived materials in a quantitative dimension has already been applied to other topics (Brügger, 2018), and research of the history of museums on the web has a great potential to benefit from these methods.

The chapter intends to investigate how web archives can be utilized for analyzing virtual museums' history. It aims to showcase the application of a close-reading approach in examining archived snapshots considering three case studies: The Metropolitan Museum of Art (MET) (New York, USA), the National Museum of Australia (Canberra, Australia), and the Deutsches Museum (Munich, Germany). These museums have been selected to offer a transnational perspective on the evolution of different types of museums, encompassing an art museum, a national museum, and a science and technology museum. Expanding the research framework, the chapter explores the various layers of data that aid in identifying crucial attributes and historical progressions of these museum websites.

Museum on the web: Theoretical implications of the phenomena of virtual museum

Analysis of museums' online presence contributes to understanding the phenomena of the virtual museum — which can be interpreted differently. The scholarly interdisciplinary project Virtual Multimodal Museum recognizes a virtual museum as a “digital footprint of a physical museum”, which is aimed to enhance “cultural presence” (Ioannides & Davies, 2018, 345). Another approach is based on relations between “physical” and “virtual” in museum practices (Bandelli, 1999; Nicholls et al., 2012). Also, the virtual museum is considered as a digital extension to the physical museum (Biedermann, 2021). As no general agreements have been established, Werner Schweibenz (2019) concludes that the concept of the virtual museum is still in progress.

To a certain extent, the virtual museum can be equated to the museum website (Gaia, 1999) or more precisely to “an information system ... which facilitates research, education, and discovery activities in the virtual space” (Povroznik, 2018, 1). Of course, not every website can be qualified as a virtual museum but only if it is designed for remote use by virtual audiences in a valuable way without a particular need to go to the physical museum. In this sense, the museum's website performs a function as a central resource of the virtual museum concept.

Researchers agree that virtual museums can be of different types and categories (Forte & Franzoni, 1998; Gaia, 1999; V-Must Deliverable Report, 2013). However, the classification according to the content of websites (Gaia, 1999) shows that not all museums' websites can be recognized in terms of what might be thought of as a “real” virtual museum. Therefore, the strategies that museums apply on the web require detailed analysis to understand the processes and trends of the evolution of these websites.

In this chapter, it is suggested that the virtual museum has to be considered through the prism of the multiplatform digital presence. Museums can create subwebsites or satellite websites for specific projects such as digital exhibitions published separately. Additionally, there are other types of resources intended for delivering museum functions via digital means. Social media is instrumental in enabling museums to engage with audiences online, promote their exhibits and events, share educational content, foster community interaction, and enhance their overall visibility and outreach (Zolo et al., 2021). Another element of a

museum's ecosystem is represented by the mobile applications, which are designed to be used not only onsite but can also be helpful to explore the museum remotely. They play a crucial role in providing an immersive and personalized museum experience, enhancing visitor engagement, accessibility, and enjoyment (Arvanitis, 2010). Mobile applications can be used to enhance learning opportunities, offering visitors the opportunity to probe and learn in a more interactive way and support the interpretations of museum content (Civantos et al., 2016). For that reason, mobile applications can be considered as a part of the virtual museum as well. In this regard, the virtual museum may be an integral part of the museum's ecosystem (see the Chapter "Information Ecosystems Through the Lens of Web Archives" by Matthew S. Weber in this book to get more detail about information ecosystems theory in practice related to web archives).

Museum websites as a multilayered historical resource

Museums on the live web engage audiences, inviting them to explore various content for research, education, entertainment, and different enlightening activities. Users may observe and interact with a kaleidoscope of resources collected, curated, and designed for them by museums. These resources can be represented by collections of digital objects, exhibitions, stories, educational and research materials, etc. Here websites perform as a specific display to orientate museums' ecosystems, comprising a set of interconnected elements. Shaping a sustainable ecosystem relates to the importance of considering the information architecture of museum websites, especially in the context of web preservation. The complexity of information architecture should correspond to a long-term preservation strategy to be accessed through web archives and to ensure the use, reuse, and exploration of these resources in the future.

Information architecture is aimed to organize and simplify information to make it user-friendly, or, in other words, to make it easier for people to find, explore, and interact with more effectively (Ding et al., 2017). However, the information architecture of a museum's website can undergo fragmentation during web preservation, and in the process lose some essential elements, meaning the subsequent reconstruction of the original state is impossible.

Some web archives such as the Internet Archive, for example, enable researchers to trace changes in the design, content, and functionality of museum websites over time, providing insights into social and cultural trends, shifting museum practices, and acceptance of technological advancements. In general, web archives tend to keep as much of the information architecture of the websites as possible in the framework of accepted strategies of web preservation, technical limitations, and other constraints (Brügger, 2018). Identification of layers of the information architecture on the museums' websites helps address the question of what can be discovered in web archives and what is missing. For researchers, consideration of these layers is necessary to reconstruct the history of virtual museums more comprehensively and also to provide adequate source criticism (Fickers, 2012).

The content layer of a website's information architecture reflects the museum's diverse activities and represents an external facet of their online presence. This layer constitutes the museum's engagement with the public and supports it in performing its cultural and societal roles. It displays the institution's initiatives, available for virtual and in-person attendees. Museums may publish valuable resources from the cultural, educational, and research perspectives, such as digital collections, virtual tours, multimedia content, lectures on various topics, guided tours in digital formats, blogs and vlogs, podcasts, interactive quizzes,

and games. Research shows that during recent decades, websites content has increased and progressed in terms of audience interactivity. Museums started by publishing short descriptions of their history and current activities, some iconic examples from the collections that they preserve, and exhibitions to welcome visitors online (Povroznik, 2020b). The process of the development of this content demonstrates how museums evolved in their understanding of the importance of the webspace and the creation of the digital environment to engage virtual visitors so as to create a “real virtual museum” (Gaia, 1999). This layer of information preserved in the web archives serves as a source base to reconstruct the history of museums’ online and onsite activities (Povroznik, 2020a). Some examples of such development are exposed in the case study section of this chapter.

The structural layer includes the hierarchy of web pages and the interdependencies of the elements (Li & Huang, 2022). Considering this layer of information helps to reconstruct the overall architecture of the website and how the content is organized logically. It involves understanding how information is organized, how users can interact with the website, and recognizing its logical flow including content, navigation, and functionality. The Wayback Machine,¹ particularly, provides researchers with statistics that introduce information on the development of the structure of the website, show the changes in the format of the published content, and support tracing the regularity of updates on the websites. In particular, radial-tree graphs generated by the Wayback Machine for each year of a website’s history show changes in the complexity of the structure. The example related to the MET helps to reveal the complex nature of the structure of the website during 1996–2024.² Studying the hierarchical arrangement of pages and sections within the website helps to grasp the changes in its structure and subsequently interpret these developments.

Also, the structural layer of information provides insight into the website’s navigation including menus and their evolution, and how it was designed to guide users in exploring information. Museums could follow different strategies or develop their vision on how to bring to the virtual audience better orientation and guidance on an exploration of the website.

In general, this layer of information helps to understand the evolution of human–computer interaction (HCI) and changes in user experience (UX) design implemented on museums’ websites. To make necessary adjustments to the websites to provide a better UX, museum practitioners analyze the behavior of online visitors who search and interact with digital content. Reconstruction of user behavior from a distance via the historical lens is rather difficult because this information is not usually captured or even published on the websites themselves. Snapshots of websites preserved in web archives facilitate the opportunity to discover and trace the changes on the menu, for example, and better grasp the trends that led to the particular state of HCI and UX design in museums, but the reasons for these changes are often uneven. Therefore, other sources should be used to analyze user behavior from the historical perspective. For instance, this can be achieved by aligning the changes on websites with surveys devoted to UX (if any). Primarily, these surveys are in use to clarify various aspects of visitors’ experience (Marty, 2008) and in this regard can serve as a valuable historical resource.

The history of websites is often approached by researchers in terms of screencast documentaries (Rogers, 2017) and, therefore, mainly the narrative layer of information of the websites is considered. However, a range of technologies has been involved in creating the website itself and to present the content. The historical development of websites should also be considered, addressing their features and limitations. This information corresponds to a technological layer, which includes the back-end architecture, such as the programming

languages, the database structure, and the server infrastructure. The technologies range from basic HTML and CSS to more advanced frameworks, JavaScript libraries, responsive design, and server-side scripting languages like PHP. All of these elements together shape functionality of the website, but evidence of only some of them can be captured through the web preservation processes. The Internet Archive is primarily focused on preserving the front-end content of websites, such as HTML, images, and other media files. It is not designed to provide detailed information about the back-end architecture of websites due to the technical limitations of the crawling process. To gain insights into the back-end architecture of a website, researchers need access to additional resources of information such as documentation, the website's source code if available, and interviews with the website developers.

At the same time, analysis of the implemented technologies helps to explain the specifics of the websites deployed in different time periods and is crucial for contextualization of the history of websites because they reflect the evolution of web development and design practices over time. Bringing a technological perspective allows researchers to explain the processes of how museums adapted to changing technologies and aligned their practices with innovations, and enhanced their online presence.

On the whole, consideration of the website as a part of a museum's ecosystem in a complex of layers opens up new perspectives to reconstruct the history of museums on the web and gain a more comprehensive understanding of museums' digital transformation. The transnational perspective enables consideration of best practices, achievements, and limitations of web preservation to see the gaps and to emphasize their significance.

A case study: Evolution of museums' web presence via close reading lenses

Case studies selection and methodology

The case study is aimed to highlight the outcomes of the close-reading approach to the history of museums' online presence and to argue for rescaling research and implementation of mixed methods to the analysis. The websites of the MET, the National Museum of Australia, and the Deutsches Museum were selected for the case study. These examples showcase transnational perspectives in museum development and compare "object"-oriented and more abstract "idea"-oriented museums (Velázquez Marroni, 2017). The MET and the National Museum of Australia are taken conventionally as "object"-oriented because they are designed to display the objects of art and history. Having a focus on science and technology the Deutsches Museum represents an "idea"-oriented museum. All the websites from the case study were established in the 1990s and have a relatively long web history to investigate.

The case study is based on the outcomes of a project on the *Digital History of Virtual Museums* conducted by the author. It refers to diagrams placed at the website of the project³ to provide visual aids for description and interpretation relevant processes. Particularly, the section "*infographics*" includes a subsection "*digital initiatives*"⁴ with the diagrams depicting models of the digital initiatives conducted by the museums. The models reflect not all of the content of the websites but exclusively the elements that aimed at a virtual visitor to explore remotely the museum itself, the collections, and other materials. To put it more precisely, digital initiatives refer to types of resources that are designed to be used by online audiences for research, education, entertainment, and other purposes and they are not dependent on the need to necessarily visit a "physical" museum in person for exploration. The complexity

of museums' websites is highlighted on diagrams by coloring the blocks according to time periods. It is worth mentioning that the diagrams show the introduced initiatives only and they do not indicate those ones that were removed at some point from the websites.

The content of the websites was traced accordingly to the classification given by Forte and Franzoni (1998) and summarized by Giuliano Gaia (1999). There are three main types of museum websites:

- 1) 'Simulated museum', trying to recreate the experience of visiting the actual [physical] museum (virtual tours, and so on) without adding any information; ...
- 2) '*Information*': the website is an instrument to use before or after the visit; it offers a lot of information not available at the actual [physical] museum; [...]
- 3) The real 'virtual museum', a website in many parts independent from the actual [physical] museum, with many sections and exhibitions residing only on the net.

(Gaia, 1999)

This classification treats websites from the point of design of a virtual museum on the web focusing on comparison of the opportunities of online and on-site exploration. The chapter is aimed to consider mainly the history of museums on the web; however, it is important to mention that the phenomenon of the virtual museum is broader than just the website of the museum, which has been already described in this chapter in the first section on theoretical implications of the phenomena of the virtual museum. The quoted classification is used as a theoretical background to this study to trace the evolution of the websites according to complexity of the content aimed at online exploration.

To avoid misinterpretations and mitigate potential confusion in the use of collocations of "real", "actual", "virtual", and "physical" museums, the third type of the websites is called subsequently in this chapter as an "actual" virtual museum.

The analysis of the websites in this case study is based only on the examination of snapshots from web archives and does not pretend to be exhaustive research. The Internet Archive and TROVE (Australian web archive) were the main resources to study the snapshots of the websites. The German National Library (Deutsche Nationalbibliothek) does not provide researchers with open access to the web-archived materials and because of that it is left for the next stages of implementation of this research.

The Metropolitan Museum of Art

Established in 1870, the MET, located in New York, USA, has grown to become one of the world's largest art museums. Another reason for selection is that the MET has been at the forefront of utilizing IT within the museum sector. A significant milestone occurred in 1968 when the museum hosted a conference titled "*Computers and their potential applications in museums*" (1968). This event showcased MET's early recognition of the potential benefits of IT in the museum field. The museum's extensive experience in application of IT, accumulated over the decades, served as a solid foundation for the development of a captivating and continuously evolving web resource. The repository of the Internet Archive preserved 10,102,070 captures of the MET website between November 11, 1996 and May 15, 2023, which includes 5,283,181 web pages and 189,019 images.⁵

The MET's website from the very first years could be used as an instrument to support the activities of the visitor before and after visiting the physical museum, which corresponds

to the “information” type of the website accordingly to Gaia (1999). The diagram devoted to the digital initiatives that were implemented on the website⁶ shows that already in the 1990s the website included:

- A database of the collection of digitized masterpieces; the individual pages of the objects included the picture of the object, short metadata, and concisely formulated historical narrative about the object.⁷
- The exhibitions accompanied with basic storytelling.⁸
- Educational resources, which included materials from the Hazen Center for Electronic Information resources,⁹ Thomas J. Watson Library Catalog,¹⁰ and various activities for online interaction (such as games and resources for teachers and families).¹¹

The subsequent development of the MET’s website led to the appearance in the 2000s of new ways to expose the objects in the digital collection (such as timelines¹²), implementation of personalized experience for users via the creation of the My MET Collection,¹³ and production of a specific education section MuseumKids¹⁴ aimed at children. These newly established features of the website brought it to the format of an “actual” virtual museum offering a wider range of activities for online visitors. These features were strengthened with other activities in the 2010s. For instance, the website received specific sections intended more precisely for “Teens”, “Adults”, and “Students”.

The snapshots of the MET website preserved on the Internet Archive provide some details from a quantitative dimension regarding the growth of the digital collection of the MET published online. The diagram “*Dynamics of Digital Objects Representation at the MET Collections Online*” illustrates the dynamics of publication of the digitized objects on the museum’s website.¹⁵ The diagram is built accordingly with information from snapshots from the Internet Archive. We can observe a significant surge in the number of digitized materials during the period of 2008–2012. This growth can be attributed to a large-scale digitization effort within the museum. To facilitate online exploration of the museum’s collections, visitors were provided with specialized guides directed them through collections and offered a range of related digital exhibits, combining visual aids with textual narratives.

Qualitative analysis shows the solid development of the museum on the web including evolving objects’ representation in qualitative and quantitative dimensions. The website changed significantly also in terms of design and combination of sections, testing various configurations and hierarchies of web pages. A more precise study, rescaling research, and applying data-driven methodologies, is needed to gain knowledge on the development of the website.

The MET offered a range of digital platforms to engage with audiences online. Already in 2008, it actively functionated on different social media platforms including Flickr, YouTube, I-Tunes U, Delicious, Facebook, My Space, and some others.¹⁶ In 2014 the museum released an application for mobile devices.¹⁷ This way, the museum moved to a multiplatform presence and creation of a diverse ecosystem. Therefore, expanding research from analysis of the museum’s presence on the web and studying of the MET functioning on other platforms is important for a comprehensive view of the evolution of digital activities of the museum, a deeper understanding of visitor engagement, and gaining insights into the development of the virtual museum as a multiplatform instance.

The National Museum of Australia

The National Museum of Australia, located in Canberra, Australia, stands as one of the country's largest and most recent museums, established in 1980 and made accessible to the public only in 2001. This institution provides extensive collections focusing on the cultural and social history of Australia. The museum was selected for the case study also because of its specific history on the web. The museum was opened for public admission in 2001, but it was already represented online in 1996, as evidenced from the snapshots preserved on TROVE and Internet Archive. The pages of the website of the National Museum of Australia have been captured 461,926 times between December 21, 1996 and June 18, 2023 by the Internet Archive, and its repository consists of 105,926 web pages and 65,048 images.¹⁸ The TROVE¹⁹ service of the National Library of Australia preserved 351,650 snapshots of this website crawled between December 21, 1996 and June 18, 2023. A thorough comparative analysis of the datasets captured by both institutions is needed to understand the gaps and differences and to inspect the opportunities of merging the collections for an in-detailed picture of the preserved materials.

Not yet open to the public, the national museum already in the 1990s organized tours in other places. The very early website was aimed at advertisement of these activities.²⁰ In the beginning, this website did not offer any particular activities online to make the visit to the website valuable for remote use for the museum's exploration and for this reason, it cannot be considered as a virtual museum of any type.

In 1998, the website was redesigned and it introduced a series of online initiatives including virtual exhibitions,²¹ educational materials, and also a library catalogue as it is exposed via the diagram.²² The digital initiatives were designed using multimedia means and included some interactive resources such as games.²³

In general, education was in the main focus of the museum's website supporting the teachers with preprepared tasks, "kits", and other materials to help the teachers to use the museum's resources in teaching.

The popular character of the website was mentioned in the Director's review of the museum's activities in 2001–2002. It was announced that 1 million people visited website in this period.²⁴ The Performance report proposed further development of the educational activities online and expanding of the "online programs to young people living in remote and outback communities".²⁵ Also, the report exposed strategic priorities in the implementation of ICT in the museum "to reach audiences beyond Canberra".²⁶ This announcement led to the subsequent development of the website into an "actual" virtual museum. It offered various activities including interactive projects and multimedia storytelling aimed at diverse audiences, such as "Kids", "Students" (by level), "Families", "Adults", and "Museum Professionals".²⁷

Social media accounts were introduced in 2010 and some information from them was captured and preserved by TROVE.²⁸ However, the social media presence of the museum requires more attention in terms of preservation due to the value of these resources and providing opportunities for their exploration and analysis. The National Museum of Australia is also an example of the multiplatform museum, and this should be under consideration while studying its historical development.

The Deutsches Museum

Deutsches Museum von Meisterwerken der Naturwissenschaft und Technik or, shortly, the Deutsches Museum is considered as one of the largest museums in the world among

History of virtual museums and web archives

museums dedicated to science and technology.²⁹ The museum was founded in 1903 and now stores and exposes exhibits distributed across dozens of thematic collections. For the case study, the museum is important with its “idea”-oriented perspective referred to processes and principles of science and technology and their historical development. The website of the Deutsches Museum was established in 1995,³⁰ but the first snapshot of the website on the web archive is dated to April 11, 1997. The repository of the Internet Archive includes 112,892 captures between April 11, 1997 and June 18, 2023, 11,837 web pages and 7,400 images.³¹ Also, the website of the Deutsches Museum is included in the program of web preservation and regularly crawled by the National Library of Germany since 2013.³²

Of particular interest is the first name of the website, which was indicated as the “Deutsches Museum Online”,³³ This self-identification specifically highlights the online presence of the museum offering digital initiatives, information, and resources openly accessible through the Internet.

The diagram with the model of digital initiatives shows a range of activities available for virtual visitors.³⁴ The central focus of the website was on to virtual exhibitions. During the mid-1990s, “exhibitions” were primarily presented in textual form, accompanied by photographs of individual exhibits. However, in the early 2000s, this format was enhanced with multimedia and hypertextuality, which offered nonlinear storytelling and interactive navigation through the exhibitions. The early version of the museum’s website included quite extensive multimedia content.³⁵ The multimedia resources demonstrated the concepts, processes, or technologies related to the particular topic, and also provided an interactive experience for the virtual visitor, simulated through simple animations. The website also had a panoramic view to demonstrate the museum’s interior and exterior views in 360 degrees and provided an option to download them.³⁶

The specificity of the Deutsches Museum website and the digital initiatives in the early years is that they duplicated previously developed content distributed on CDs. The museum adjusted this content for placing on the website. It is still possible to download the files from the Wayback Machine, but it is quite problematic to replay due to obsolescent technologies.

Another interesting feature of the website was implemented in 2000. Using the “Live Kamera”³⁷ option, it was possible to observe activities of the museum onsite via the camera and broadcasting technologies implemented on the website.

In 2006, the website underwent an update, resulting in a changed design while preserving the fundamental structure of its sections. The update introduced newly created databases of online publications, museum catalogues, and the museum’s archive. However, the multimedia sections were removed from the website. Over time, the website’s focus on digital initiatives diminished in favor of onsite activities. This transition aimed to encourage virtual users to visit the physical museum, emphasizing the value of the real-life museum experience.

In general, the website could potentially transit to an “actual” virtual museum but the approach to the webspace changed toward minimizing online initiatives valuable for remote exploration. This track requires subsequent investigation due to a variety of factors.

Results of comparison

The considered cases show different strategies for the exploitation of the web by museums and demonstrate the ways that museums followed to create diverse content, engage different audiences, and expand the activities of the museum beyond museum walls. “Object-oriented”

museums could develop their website into “actual” virtual museums even if they had different starting conditions.

Comparative analysis of the number of captures preserved on the Internet Archive shows the significant difference between the museum based in the USA and others. It demonstrates the significance of national programs for web preservation in sustaining research and securing born-digital heritage. However, it is important to raise questions about the factors influencing success on the web with diverse activities online and the conditions for failure in building an “actual” virtual museum, about acceptance of the technologies in the museum and their correlation with the development of digital initiatives, about the strategic decisions that were made to develop the website, and so on.

Rescaling research on virtual museums using web archives: Do we have a choice to study virtual museums comprehensively?

Changing the research optics and implementation of mixed methods enable a more comprehensive approach to better understand museum practices and their evolvement on the web. Carrying out holistic analyses requires capturing a complex picture (Creswell & Creswell, 2018) that includes identifying patterns, trends, and relationships across all elements of the system. Rescaling research enables moving beyond the study of individual cases or specific aspects of the museum activities using separate snapshots.

Implementation of a data-driven approach is considered as a necessary step in changing the scale of the research. It involves analysis of a larger quantity of data represented on a museum’s website. As has already been mentioned in section “Museum websites as a multi-layered historical resource” of this chapter, the website is a multilayered entity and its datasets can be built based on these layers to address particular research questions. For example, a collection of snapshots obtained from the web archive can be converted into text corpora for subsequent analysis. Expanding research methods and borrowing computer linguistics methods may help to identify key topics in a museum’s narrative and its changing through time and other patterns and specifics of the museum’s online presence. This approach has been experimentally implemented in regard to Danish museums’ websites. The diachronic analysis of the text corpus from the Danish web archive allowed the researchers to identify clusters and categorize websites within the national domain (Skov & Svarre, 2023).

Hyperlinks are another type of data for analysis to bring new knowledge about the historical development of museum websites. Based on live web scraping technologies, the study of external links to the British Museum’s website revealed the geographical locations and types of connected institutions that published references to the museum’s website (Kahn, 2019). Conducting similar research from the historical perspective requires gaining a massive amount of hyperlinks from the web archive. However, for doing that, the web archive should have a domain preservation focus, such as mentioned above — the Danish web archive (Netarkivet³⁸) has and provides this research option. At the same time, the Internet Archive does not provide the function to acquire the list of inbound links and that limits crucially the opportunities to analyze museums’ websites worldwide.

Another advantage of the national programs of web preservation relates to the opportunity for at-scale metadata analysis due to the systematic character of preservation, consistency in shaping metadata, and option for extraction of data according to specific characteristics of the websites, etc. For example, the research of the development of the Danish museums on the web in 2005–2020 showed dynamics in size, represented types of data, and some other

features of more than 100 museums in Denmark (Skov & Svarre, 2023). At the same time, building the relevant dataset based on the data from the Internet Archive would be more problematic due to technical and other constraints.

Dealing with other types of data related to the websites and their analysis requires more precise attention. The websites may include also images, audio and video files, and various multimedia formats at the content layer of information. The researcher needs to raise appropriate questions relating to how to use these data and to shape the dataset for answering them. However, it can be quite problematic not only due to the highly heterogeneous character of the websites but also due to disconnection of the objects from the contexts in the process of obtaining data from web archives. In this regard, critical reflection on the obtained data can become a very complicated process. Of course, providing a thorough source criticism is a necessary step for any type of historical research, but the reborn-digital resources (Brügger, 2018), such as web-archived websites, makes the source criticism process even more complex (Fickers, 2012).

Rescaling research also requires considering the multiplatform presence of museums and gaining relevant resources for such analysis. Social media accounts of museums have been recognized as born-digital heritage requires preservation (Zuanni, 2021). However, their preservation and subsequent representation via web archives is still challenging due to personal data protection regulations, copyright restrictions, and other reasons. At the same time, there are some projects relating to the preservation of museum accounts on social media that give a chance to rescale studying of museums' history on the web. For example, the platform Archive-It provides access to the collection the "Smithsonian Institution Social Media", regularly captured from 2012.³⁹

Despite the broad opportunities for rescaling research, there are strong limitations for comprehensive analysis of the history of museums on the web. Lack of preserved sources related to some elements of the museum's informational ecosystem is such a problem. In this regard, museums' mobile applications can be considered as heritage under threat due to lack of sufficient attention to their preservation.

Studying the history of technologies for building websites, creating content, multimedia creation and integration, and interactive features would be a valuable contribution to a deepened comprehension of technological advancements, to gain insights into the ways technologies influenced the presentation and delivery of digital content. Unfortunately, the evidence base for doing this is limited because many elements of such web pages have not been captured due to the specificity of formats and challenges for web preservation.

There are also other issues related to investigating the history of technologies. Researchers have already pointed out the rapid obsolescence of technologies (O'Dwyer, 2008) in the digital age. Regarding museums on the web, Flash is such an example. This animation technology was widely used on websites but was then abandoned by developers. Therefore, significant information placed on Flash was irretrievably lost and there is no way to reconstruct the full picture of a Flash-based website (Schafer, 2017). There are other gaps that still remain to be identified and addressed in the future.

Comprehensive analysis of museums' websites requires also analysis of the virtual audiences and their behavior. The case study discussed here helps to identify the main audiences for whom the museums developed their content. However, perception of this content and its use by the particular audiences is often unknown. Guest books on websites may serve as an interesting resource to gain the opinion of the users about their experience in the museum, onsite, and online (Villaespesa, 2019). On the live web, web analytics provide information

on user behavior including user interactions, such as page views, and session duration. This data helps museums to understand how visitors navigate through their websites, which pages or exhibits are most popular, and how users engage with the content. Web analytics also help to identify different segments of the online audience based on demographics, geographic location, interests, and other characteristics. Web analytics might be a very practical and important historical resource, but it can be accessed only on the live web. From the historian's point of view, this information might serve as a valuable resource to explain changes made to websites. However, web analytics are very rare to find on the archived web. Only small evidence from web analytics is published on these websites. For example, the Deutsches Museum in its first years on the web published these analytics⁴⁰ and this gives some insights on the virtual audience and traffic.

Conclusion

Studying the history of virtual museums opens up broad opportunities to gain new insights and benefit from the use of web archives. The research shows that analysis of the history of museums on the web can benefit from utilizing web archives and data-driven methodologies. In this regard, the domain-oriented approach of national systematic web preservation brings substantial opportunities for expanding research and raising new research questions in relation to the available data. However, national archives often limit research with restricted access to the archived materials, and in this regard the Internet Archive seems to be indispensable.

Application of mixed methods in combination with qualitative and quantitative methodologies enables a more efficient and systematic analysis of the phenomenon of a virtual museum. Rescaling research, expanding close-reading techniques, and implementing data-driven methodologies offers more complex analysis to address various questions in regard to the history of museums on the web and their multiplatform digital presence.

Conducting data-driven research requires building or using adequate datasets. The contemporary emphasis on preservation of born-digital heritage is focused on websites but attention is required to preserve other platforms, applications, and also the technologies themselves. In the current situation, the analysis of different aspects of the development of virtual museums is productive but can be fragmented depending on the available data and also limited from the perspective of comparative studies because it is not easy to find equally preserved cases represented in the web archives. In this regard, this chapter echoes another chapter of this book “An Inclusive approach to web archiving” by Sophie Gebeil. Therefore, refining and expanding preservation strategies of born-digital resources appears to be critical for ensuring their sustainability, accessibility, and discoverability.

Investigation of virtual museums at the moment is centered around the idea of the website as the key environment of museums' online presence. However, it should be expanded to analysis of other platforms (including social media and mobile applications) and consider the informational ecosystem in this regard as well.

Acknowledgements

I would like to express my gratitude to Prof. Andreas Fickers and the colleagues of the Luxembourg Centre for Contemporary and Digital History (C²DH) for the support of the project “Digital History of Virtual Museums” (2019–2022). Also, I am thankful to Prof.

Julianne Nyhan for advice and the opportunity to continue the research in the Humanities Data Science and Methodology (HDSM) at the Technical University of Darmstadt.

Notes

- 1 Wayback Machine. Retrieved March 15, 2024, from <https://web.archive.org/>.
- 2 The Metropolitan Museum of Art, Wayback Machine. Retrieved March 15, 2024, from <https://web.archive.org/web/sitemap/www.metmuseum.org/>.
- 3 Digital History of Virtual Museums. Retrieved March 21, 2024, from <https://web.archive.org/web/20240321235101/http://virtualmuseumhistory.com/infographics/>.
- 4 Infographics. Retrieved March 21, 2024, from <http://virtualmuseumhistory.com/infographics/>.
- 5 Metropolitan Museum of Art, Wayback Machine. Retrieved March 15, 2024, from https://web.archive.org/web/20040901000000*/metropolitan%20mucum%20of%20art.
- 6 Modelling of the Metropolitan Museum of Art Digital Initiatives Online, 1996–2019. Retrieved March 21, 2024, from <https://web.archive.org/web/20240321235101/http://virtualmuseumhistory.com/infographics/>.
- 7 For example, Mosque Lamp, Wayback Machine. Snapshot on February 17, 1997, <https://web.archive.org/web/19970217212424/www.metmuseum.org/htmlfile/gallery/second/islamicl.html>.
- 8 For example, The Glory of Byzantium, Wayback Machine. Snapshot on June 19, 1997, <https://web.archive.org/web/19970619043035/www.metmuseum.org/htmlfile/education/byzantium/title.html>.
- 9 Hazen Center, Wayback Machine. Snapshot on January 18, 1998, <http://web.archive.org/web/19980118222602/www.metmuseum.org:80/htmlfile/watson/hazen.htm>.
- 10 Thomas J. Watson Library Catalog, Wayback Machine. Snapshot on January 18, 1998, <http://web.archive.org/web/19980118222615/www.metmuseum.org:80/htmlfile/watson/index.htm>.
- 11 Education, Wayback Machine. Snapshot on November 11, 1996, <http://web.archive.org/web/19961111004958/www.metmuseum.org/htmlfile/Education/edu.html>.
- 12 Timeline, Wayback Machine. Snapshot on October 19, 2000, <http://web.archive.org/web/20001019060344/www.metmuseum.org/toah/splash.htm>.
- 13 My MET Museum, Wayback Machine. Snapshot on August 5, 2004, http://web.archive.org/web/20040805202951/www.metmuseum.org/mymetmuseum/index.asp?HomePageLink=mymetmuseum_r.
- 14 MuseumKids, Wayback Machine. Snapshot on February 19, 2003, http://web.archive.org/web/20030219075651/www.metmuseum.org/explore/museumkids.htm?HomePageLink=museumkids_c.
- 15 Dynamics of Digital Objects Representation at the MET Collections Online, 1996–2019. Retrieved March 21, 2024, from <https://web.archive.org/web/20240321235506/http://virtualmuseumhistory.com/met-museum/>.
- 16 MET Share, Wayback Machine. Snapshot on December 16, 2008, <https://web.archive.org/web/20081216031649/www.metmuseum.org/metshare/>.
- 17 Metropolitan Museum Launches Flagship App for iPhone, iPad, and iPod touch. Published online September 2, 2014. Retrieved April 18, 2024, from <https://web.archive.org/web/20240418030700/www.metmuseum.org/press/news/2014/metropolitan-museum-launches-flagship-mobile-app>.
- 18 The National Museum of Australia, Wayback Machine, https://web.archive.org/web/20040901000000*/National%20Museum%20of%20Australia.
- 19 The National Museum of Australia, Trove, <https://trove.nla.gov.au/search/category/webstes?keyword=http%3A%2F%2Fwww.nma.gov.au%2F&cl-site=nma.gov.au>.
- 20 For example, The Eyes on the Ball, TROVE. Snapshot on January 19, 1997, <https://webarchive.nla.gov.au/awa/19970119114645/www.nma.gov.au/NMA/3exheob.htm>.
- 21 For example, Gumnut Adventures - The World of May Gibbs, TROVE. Snapshot on July 13, 1998, <https://webarchive.nla.gov.au/awa/19980713113026/http://nma.gov.au/collections/ash/index.htm>.

- 22 Modelling of the National Museum of Australia Digital Initiatives Online, 1996–2019. Retrieved March 21, 2024, from <https://web.archive.org/web/20240321235101/http://virtualmuseumhistory.com/infographics/>.
- 23 For example, Ann Fibian, TROVE. Snapshot on December 1, 1998, <https://webarchive.nla.gov.au/awa/19981201085109/www.nma.gov.au/education/ann/index.htm>.
- 24 Director's review of operations, Wayback Machine. Snapshot on April 17, 2003, https://web.archive.org/web/20030417191917/www.nma.gov.au/annual_report_2002/__data/page/1645/nma_annrep2002_part01.pdf.
- 25 Performance reports, p. 25, Wayback Machine. Snapshot on April 17, 2003, https://web.archive.org/web/20030417235225/www.nma.gov.au/annual_report_2002/__data/page/1645/nma_annrep2002_part02.pdf.
- 26 *Ibid.* p. 41.
- 27 Modelling of the National Museum of Australia Digital Initiatives Online, 1996–2019. Retrieved March 21, 2024, from <https://web.archive.org/web/20240321235101/http://virtualmuseumhistory.com/infographics/>.
- 28 For example, Nat_Museum_Aust on Twitter, TROVE. Snapshot on April 6, 2011, https://web.archive.nla.gov.au/awa/20110406084034/http://twitter.com/Nat_Museum_Aust.
- 29 Deutsches Museum, Wikipedia. Retrieved January 11, 2024, from https://web.archive.org/web/20240111090301/https://en.wikipedia.org/wiki/Deutsches_Museum#cite_note-1.
- 30 Deutsches Museum, Wayback Machine. Snapshot on April 11, 1997, <http://web.archive.org/web/19970411131440/www.deutsches-museum.de/dme12.htm>.
- 31 Deutsches Museum, Wayback Machine, https://web.archive.org/web/20040901000000*/Deutsches%20Museum.
- 32 Deutsches Museum von Meisterwerken der Naturwissenschaft und Technik, Katalog der Deutschen Nationalbibliothek. Retrieved September 2, 2024, from <https://web.archive.org/web/20240902124613/https://portal.dnb.de/opac.htm?method=simpleSearch&cqlMode=true&query=idn%3D1048129039>.
- 33 Deutsches Museum, Wayback Machine. Snapshot on April 11, 1997, <http://web.archive.org/web/19970411130943/www.deutsches-museum.de/dme01.htm>.
- 34 Modelling of the Deutsches Museum Digital Initiatives Online, 1996–2019. Retrieved March 21, 2024, from <https://web.archive.org/web/20240321235101/http://virtualmuseumhistory.com/infographics/>.
- 35 Museum Multimedial, Wayback Machine. Snapshot on January 31, 1998, <http://web.archive.org/web/19980131152131/www.deutsches-museum.de/mum/index.htm>.
- 36 Video, Wayback Machine. Snapshot on January 31, 1998, http://web.archive.org/web/19980131174146fw/www.deutsches-museum.de/mum/video/e_video.htm.
- 37 Live-Kamera, Wayback Machine. Snapshot on March 3, 2000, <https://web.archive.org/web/20000303075527/www.deutsches-museum.de/mum/live/kamera.htm>.
- 38 Netarkivet. Retrieved March 15, 2024, from www.kb.dk/en/find-materials/collections/netarkivet.
- 39 Smithsonian Institution Social Media. Archive-It. Retrieved March 15, 2024, from www.archive-it.org/collections/3393.
- 40 Deutsches Museum, Wayback Machine. Snapshot on April 11, 1997, <http://web.archive.org/web/19970411131236/www.deutsches-museum.de:80/sys/stat/>.

References

- Ainsworth, S., Nelson, M., & Van de Sompel, H. (2015). Only one out of five archived web pages existed as presented. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. Association for Computing Machinery. [ISBN: 978-1-4503-3395-5] (pp. 257–266).
- Arthur, P. (2018). Engaging collections and communities: Technology and interactivity in museums. In *Proceedings of the Digital Humanities in the Nordic Countries, 3rd Conference*. CEUR-WS, 2084, 250–262.
- Arvanitis, K. (2010). Museums outside walls: Mobile phones and the museum of the everyday. In R. Parry (Ed.), *Museums in a Digital Age* (pp. 170–176). Routledge.

History of virtual museums and web archives

- Bandelli, A. (1999). Virtual spaces and museums. *Journal of Museum Education*, 24(1/2), 20–22.
- Biedermann, B. (2021). Virtual museums as an extended museum experience: Challenges and impacts for museology, digital humanities, museums, and visitors — in times of (coronavirus) crisis. *Digital Humanities Quarterly*, 15(3), n.p.
- Bowen, J., Borda, A., Gaia, G., & Boiano, S. (2023). Early virtual science museums: When the technology is not mature. *Internet Histories*, 8(1–2), 9–33. <https://doi.org/10.1080/24701475.2023.2298155>.
- Brügger, N. (2018). *The Archived Web: Doing History in the Digital Age*. MIT Press.
- Civantos, A., Brown, M., Coughlan, T., Ainsworth, Sh., & Lorenz, K. (2016). Using mobile media creation to structure museum interpretation with professional vision. *Personal and Ubiquitous Computing*, 20, 23–36.
- Computers and Their Potential Applications in Museums* (1968). A conference sponsored by the Metropolitan Museum of Art, supported by a grant from the IBM corporation, April 15–16–17, 1968. Arno Press. Retrieved March 12, 2024, from <https://web.archive.org/web/20221004063834/https://libmma.contentdm.oclc.org/digital/collection/p15324coll10/id/204737/>
- Creswell, J. W., & Creswell, D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE.
- Ding, W., Lin, X., & Zarro, M. (2017). *Information Architecture: The Design and Integration of Information Spaces* (2nd ed.). Springer Nature Switzerland AG.
- Fickers, A. (2012). Towards a new digital historicism? Doing history in the age of abundance. *VIEW Journal of European Television History and Culture*, 1(1), 19–26.
- Forté, M., & Franzoni, M. (1998). Il museo virtuale: Comunicazione e metafore. *Sistemi intelligenti*, 2, 193–240.
- Fotakis, Th., & Economides, A. (2008). Art, science/technology, and history museums on the web. *International Journal of Digital Culture and Electronic Tourism*, 1(1), 37–63.
- Gaia, G. (1999). The museum and its website. In *Proceedings of the Conference on the Museums and the Web. Archives & Museum Informatics*. Retrieved February 24, 2024, from <https://web.archive.org/web/20240224105120/www.museumstheweb.com/mw99/papers/gaia/gaia.html>.
- Gaia, G., Boiano, S., Bowen, J., & Borda, A. (2020). Museum websites of the first wave: The rise of the virtual museum. In *Proceedings of the Conference on Electronic Visualisation and the Arts EVA London 2020*. BCS, the Chartered Institute for IT, 2020 [ISBN 978-1-78017-538-6] (pp. 24–31).
- Giannini, T., & Bowen, J. (2022). Museums and digital culture: From reality to digitality in the age of COVID-19. *Heritage*, 5(1), 192–214.
- Haitham, E. (2019). *Museum Innovation and Social Entrepreneurship: A New Model for a Challenging Era*. Routledge.
- Ioannides, M., & Davies, R. (2018). ViMM — virtual multimodal museum: A manifesto and roadmap for Europe's digital cultural heritage. In *International Conference on Intelligent Systems*. IEEE (pp. 343–350).
- Kahn, R. (2019). The nation is in the network: Locating a national museum online. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities* (pp. 161–177). Routledge.
- Li, W., & Huang, X. (2022). The interactive design and user experience of virtual museums: Case study of the Virtual Palace Museum. In *14th International Conference on Cultural and Creative Design (CCD 2022)* (pp. 400–409).
- Marty, P. F. (2008). Museum websites and museum visitors: Digital museum resources and their use. *Museum Management and Curatorship*, 23(1), 81–99.
- Nicholls, A., Pereira, M., & Sani, M. (Eds.). (2012). *The Virtual Museum. Report 1*. Retrieved March 12, 2024, from https://web.archive.org/web/20180508124902/http://online.ibr.regione.emilia-romagna.it/1/libri/pdf/LEM_report1_theVirtualMuseum.pdf.
- O'Dwyer, A. (2008). European television archives and the search for audiovisual sources. In J. Bignell & A. Fickers (Eds.), *A European Television History* (pp. 257–263). Wiley-Blackwell.
- Perera, K., & Chandra, D. (2010). How museums in the digital age become more dynamic, visitor oriented? In *Proceedings of the International Conference on Qualitative and Quantitative Methods in Libraries (QQML 2010)*.
- Povroznik, N. (2018). Virtual museums and cultural heritage: Challenges and solutions. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*. CEUR-WS, 2084, 394–402.

The Routledge Companion to Transnational Web Archive Studies

- Povroznik, N. (2020a). Web archives in reconstructing history of virtual museums: Potential and limitations. *Perm University Herald. History*, 4, 95–105.
- Povroznik, N. (2020b). Digital history of virtual museums: The transition from analog to Internet environment. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*. CEUR-WS, 2612, 125–136.
- Rogers, R. (2017). Doing web history with the Internet Archive: Screencast documentaries. *Internet Histories*, 1(1–2), 160–172.
- Schafer, V. (2017). From far away to a click away: The French state and public services in the 1990s. In N. Brügger & R. Schroeder (Eds.), *The Web as History* (pp. 117–133). UCL Press.
- Schafer, V., & Winters, J. (2021). The values of web archives. *International Journal of Digital Humanities*, 2(1), 129–144.
- Schweibenz, W. (2019). The virtual museum: An overview of its origins, concepts, and terminology [Electronic version]. *The Museum Review*, 4(1). Retrieved February 13, 2024, from https://web.archive.org/web/20240213164047/https://themuseumreviewjournal.wordpress.com/2019/08/02/tmr_vol4no1_schweibenz/.
- Skov, M., & Svarre, T. (2023). A diachronic analysis of Danish museum websites. In *RESAW 2023*. Panel discussion: Museums on the Web: Exploring the past for the future.
- Velázquez Marroni, C. (2017). Beyond the ‘object-oriented vs. visitor/idea-oriented museum’ divide: The value of objects for museum experiences. *Museologica Brunensia*, 6(1), 12–20.
- Villaespesa, E. (2019). Museum collections and online users: Development of a segmentation model for the Metropolitan Museum of Art. *Visitor Studies*, 22(2), 233–252.
- V-Must Deliverable Report. (2013). Retrieved March 12, 2024, from www.academia.edu/6090456/Terminology_definitions_and_types_of_Virtual_Museums.
- Zollo L, Rialti R, Marrucci A, et al. (2021) How do museums foster loyalty in tech-savvy visitors? The role of social media and digital experience. *Current Issues in Tourism*, 25(18), 2991–3008.
- Zuanni, C. (2021). Theorizing born digital objects: Museums and contemporary materialities. *Museum and Society*, 19(2), 184–198.

15

EXPLORING SKILLS AND TRAINING REQUIREMENTS FOR THE WEB ARCHIVING COMMUNITY

Helena Byrne, Juan-José Boté-Vericad, and Sharon Healy

Introduction

This chapter is concerned with examining current trends in the challenges for library and archival roles in a digital context, in particular the skills and training requirements for information professionals who participate in web archiving as part of their role in a cultural heritage organization. There is much consensus that web archiving involves the selection and collection of web content, preserving it for the future and making the collected web content available for access and use (Niu, 2012). While a web archive may also come under the umbrella of a digital archive, it is nonetheless, “a specific type of digital archive” (Lomborg, 2019, p. 99). Costea (2018) notes that the term may not be apparent enough for some researchers, while other studies suggest that some individuals have a blurred understanding of the differences between a digital archive, a digital library, and a data repository (Healy & Byrne, 2023). Brügger (2018) adds that while the terms web archive and web archiving may be confusing, they were coined decades ago and are already part of the vernacular for this type of resource. On top of this, web archiving is a complicated process (Brügger, 2018; Bingham & Byrne, 2021) and is representative of the processes and activities described in the Archive-It web-archiving lifecycle model, which is inclusive of appraisal, selection, capture, storage, quality assurance, preservation and maintenance, replay/playback, access, use, and reuse (Boté Vericad, 2012; Bragg & Hanna, 2013; Boté & Termens, 2019).

Over the past decades, there have been continued advancements in the adoption of web archiving as a technique for the preservation of digital heritage on the web for research and other purposes. Web archiving also provides a means to combat the omnipresent problem for the citation of Universal Resource Locator (URL) references due to link rot. Link rot, also known as reference rot, broken links, or link decay, is used as a term to indicate that a URL no longer provides direct access to a file or web page as originally indicated. Moreover, even if a URL is stable, content drift can be an issue as the contents of a web page could change, so returning readers may not view the exact same cited content (Schneider et al., 2009).

Gomes et al. (2011) provide an overview of global development in web-archiving initiatives and observe that there was a significant growth in web-archiving initiatives from 2003, but mostly in developed countries. Such institutions include national and regional libraries and archives, university libraries and academic institutions, nonprofit organizations, and commercial organizations (Wikipedia, 2011). Thus, while it may not be “uncommon” for cultural heritage institutions such as national libraries and academic research libraries to employ staff to manage web-archiving activities, nonetheless, it is “still rather unusual that librarians or archivists who hold these positions had received any formal training in web archiving before they joined web archiving teams” (Holownia, 2020). This is why the International Internet Preservation Consortium (IIPC) (n.d.) and others (Byrne & Rarugal, 2019, 2020) advocate the need for beginners training in web archiving. Moreover, Truman (2016) suggests the need for continual training of existing staff, while Healy et al. (2022) suggest the need for more discussion and research to understand the necessary levels of training required across multiple levels of competencies, as the challenges “do not become less with increasing experience” (Healy et al., 2022, p. 103). Therefore, this chapter aims to build on these studies, and add to the discussion and literature on the topic.

Purpose of the research

This chapter aims to develop more discussion on the skills and training requirements for librarians and archivists who participate in web archiving as part of their role in a cultural heritage organization. In doing so, we examine the broader contexts of the challenges for library and archival roles within digital environments, and how this compares with the challenges for web-archiving communities. We also assess the landscape for training materials for library/archival roles, inclusive of web archiving. To end, we offer an overview of training opportunities that are available in higher education institutions across Britain, Ireland, and Spain. Through an analysis of online course descriptions of the curricula for a number of information studies courses across these three regions, we aim to evaluate whether there are any references to web archiving. We use this as a first step in measuring indicators as to the presence of web-archiving activities as part of course curricula. As part of this, the chapter will focus on two main research questions:

- What type of skills and training are required for library and archival roles for participating in web-archiving activities, and how does this compare with the skills and training requirements for library and archival roles in a digital environment, in general?
- How adequate are the current curricula for web archiving on courses in information studies across Britain, Ireland, and Spain?

We hope to offer some insights that will promote and foster further discussions on the need for continual development in skills and training for participation in web-archiving activities in cultural heritage organizations.

Approach and methodology

First, we use an exploratory approach to examine literature and case studies that provide observations on current trends in the challenges for library and archival roles in a digital

Exploring skills and training requirements

context. Then, we take a more focused look at the challenges, and skills and training requirements for librarians and archivists who participate in web archiving as part of their role in a cultural heritage organization. We also explore the landscape of web-archiving training materials, which may be useful for novices starting out.

Second, through an analysis of published website content, we assess the current status of training opportunities that are available through the curricula of courses on information studies in higher education institutions across the geographical regions of Britain, Ireland, and Spain to see if there was any reference to web archiving. These regions were chosen because they reflect the language of the authors involved, and also due to their understanding of different terminologies and scenarios in formal education across these regions.

Through desk research, the team compiled (for each region) a list of information studies courses in higher education institutions in the context of Library Studies, Archival Studies, Records Management, Digital Curation/Preservation, and Digital Humanities. The research team evaluated whether the course provided a web page with a course description for inclusion to the list and collated this data on a spreadsheet. The URL for the online curriculum overview was added to the spreadsheet and then captured using the “Save Page” functions provided by the Internet Archive and Arquivo.pt at the point of review. The links to the archival captures were also added to the spreadsheet. The spreadsheet is published as a dataset on the British Library Shared Research Repository (2023). All the courses reviewed in this study were advertised for a September 2023 start date. Some courses provided detailed course overviews while others only provided short descriptions. In Britain and Ireland, only postgraduate courses were reviewed while in Spain most information science courses are provided at undergraduate level, except for one postgraduate course.

Between February and May 2023, the research team assessed the information provided on the web page of each course description as to whether there was a presence of information on web-archiving activities as part of course curricula. Of course, that is not to say that web-archiving activities are not included at more granular levels of modules within a course, for example, web archiving could be included as part of a module on digital preservation or digital curation. This is why an additional column was included in the dataset to indicate if the course offered a digital-based module that covered digital preservation, digital curation, and/or digital information management. Due to the limited timeframe for this study only a close reading of the information on the website was included in the analysis. It would be interesting to conduct a follow-up study to see if a survey of higher education course providers would yield a different result. An analysis of this part of the research is included in the training and education section of this chapter.

Challenges for library and archival roles in the digital age

To understand the skills and training requirements necessary for library and archival roles in a digital environment, it is useful to start by examining the types of challenges experienced by these communities, and the skills and training requirements that are necessary to fulfil these roles. We then take a more focused look at the challenges for librarians and archivists who participate in web archiving as part of their role in a cultural heritage organization. In doing so, we lay the groundwork to open up some points for discussion further on.

Challenges for library and archival roles in digital environments

There are different challenges regarding library and archival roles in digital environments in general. The San José State University School of Information (SJSU-SI) annual report analyzed the latest career trends for information professionals. They collected data from a review of job duties from recent job postings for Library and Information Science (LIS) professionals and analyzed them. The results of the report considered different basic skills including computer technology (39%) and data management (10%) as skills expected from employers (39%) (SJSU-SI, 2023). The main challenges faced by librarians in the digital era include a gap between theoretical and practical aspects of librarianship in curricula, budgetary constraints, inadequate information technology (IT) infrastructure, poor information literacy skills and digital literacy, lack of accreditation, lack of incentives, and inactive role of LIS professional associations.

Davies (2022) and Olumide et al. (2022) also point to challenges regarding low information literacy rate among LIS professionals and users, lack of awareness about e-resources, and inadequate promotional activities are also major challenges. Librarians are faced with having to improve their information literacy and digital literacy skills, as well as their communicative skills to promote e-resources and engage in professional development opportunities. Olumide et al. (2022) recommend that libraries establish policies that prioritize quality over quantity and resolve issues related to librarian's scale, promotion, and salary on a priority basis so that librarians may work with devotion.

Archivists face a variety of challenges when working in digital environments including the need to ensure the long-term preservation of digital records, the management of large volumes of information, balancing access, and privacy concerns, navigating complex copyright and intellectual property laws, adapting to new technological tools and platforms, and ensuring the security of digital collections against cybersecurity threats. To address these challenges, archivists require expertise in digital technologies, information management, preservation, and legal and ethical considerations. They must also be able to collaborate effectively with stakeholders, including IT professionals, researchers, and donors. The complexity and dynamic nature of digital environments require archivists to stay current with technological advancements and be prepared to adapt to new challenges as they arise. By addressing these challenges, archivists can effectively preserve and provide access to digital records while ensuring compliance with applicable laws and ethical standards (Wright & Benoit Iii, 2019).

Another important aspect in relation to the skills for library/archival roles in a digital environment is big data, and the emergence of new roles in research such as “data scientist”. The definition of a data scientist is broad, including professions such as business intelligence analyst and data analyst. Data scientists can be categorized into vertical or horizontal types, and into developers, researchers, creators, business people, and mixed/generic types based on their skills. Data developers focus on technical analysis and data management, while data researchers are typically academics with a strong background in statistics or social sciences (Semeler et al., 2019). Probability models, data mining, data visualization, and machine learning are also crucial in understanding and utilizing large amounts of data. Domain expertise is also necessary as data science is applied in various fields such as economics, finance, and medicine. Critics argue that data science approaches often overlook the importance of domain knowledge (Virkus & Garoufallou, 2019). In this case, data science skills are the foundations of data librarians.

Exploring skills and training requirements

Data librarians are professionals who work with researchers to manage and preserve data collections. They share skills with managers in preserving and evaluating the economic values of data. They should possess knowledge of file formats, documentation and meta-data standards, and disciplinary research practices. Data librarians use the core values, ethical principles, skills, and professional knowledge of librarianship to acquire, hold, and manage data collections of files used for primary and secondary analysis; provide reference services and technical assistance; and provide education on access and use of datasets. The skills and competencies necessary for a data librarian include interpersonal skills, communication skills, and knowledge of the research environment, data usage, IT, and legal and regulatory commons. The profession of data librarianship involves data management and curation, data visualization and geospatial representation, and advanced data reference services (Semeler et al., 2019).

Archivists are also faced with challenges due to the transitions from traditional archival methods that were fundamentally based on paper, to a digital age whereby the “production and consumption of digital records are informed by social and industrial trends and by computer and data methods” that may display “little or no connection to archival methods” (Marciano et al., 2018). This necessitates a shift in archival training and education to incorporate transdisciplinary methods such as computational archival science, which would benefit both archival and computational science communities (Marciano et al., 2018).

Therefore, for librarians and archivists, collaborating with other professionals, such as data scientists, IT professionals, web developers, and researchers, is also crucial. By working together, librarians and archivists can ensure that the digital content is being preserved in a way that is useful and accessible (Sweester & Orchard, 2019; Perret, 2023). Using automation tools, optical character recognition (OCR) tools, metadata extraction tools, and quality control software can help librarians and archivists manage the increasing volume of digital content that needs to be preserved.

Finally, to remain relevant and effective in the digital era, librarians and archivists must adopt specific strategies that enable them to stay abreast of new developments in their field. These strategies include staying informed, being proactive, collaborating with other professionals, using automation tools, and emphasizing accessibility. Staying informed involves attending conferences, workshops, and seminars, as well as reading blogs, journals, and online forums. By doing so, librarians and archivists can keep up to date with the latest techniques, standards, and tools used in the field. Being proactive in identifying and collecting digital content that needs to be preserved is essential, especially for digital content that is important for future research and analysis (Vassilakaki & Moniarou-Papaconstantinou, 2017). In this way, librarians and archivists can maintain their relevance and effectiveness in the digital era and ensure that all types of digital content are protected and preserved for future generations.

In the next section, we examine the challenges for web archivists in cultural heritage organizations and thereafter assess whether there are any similarities or differences across these roles.

Challenges for web archivist roles in cultural heritage organizations

The chapter has several overlaps with other studies, which focus on the challenges for web-archiving initiatives, web-archiving practices and workflows, and the types of skills required

by staff for the successful development of a web-archiving program. In the next section, we take a look at several of these studies.

The National Digital Stewardship Alliance (NDSA) conducted web-archiving surveys in the United States in 2011, 2013, 2016, and 2017, which were aimed to get a better understanding of the types of web-archiving activities being conducted across the United States. Founded in 2010, the NDSA is a primarily US-based voluntary consortium of educational, governmental, nonprofit, and commercial organizations committed to the long-term preservation of digital information (Farrell et al., 2018). In broad terms, the NDSA web-archiving surveys targeted US web-archiving communities to examine the scope of their web-archiving activities including content selection, access and discovery, tools and services, staff levels, and the types of skills required by staff for the successful development of a web-archiving program. While there is not enough space here to explore each aim, or indeed each report, we will focus on the findings from the most recent report (Farrell et al., 2018).

In the 2017 NDSA web-archiving survey, participants ($N = 119$) were questioned on their top three considerations for the development of their web-archiving programs. Access and use were high on the list in terms of “researcher interactions, web analytics, use cases” (Farrell et al., 2018, p. 14). This was followed by considerations for the cost of web-archiving programs inclusive of staff level requirements and service provisions, followed by quality assurance in terms of “accuracy, completeness, replay fidelity” (Farrell et al., 2018, p. 14). Of interest here, Farrell et al. note how consideration for access and use saw an increase from previous studies. Participants were also asked for their top three staff skills essential for the successful development of a web-archiving program of which the configuration or operation of archival tools rated highest. This was followed by appraisal and selection, and the quality assurance and troubleshooting quality issues (Farrell et al., 2018). For Farrell et al. (2018), the responses from this survey “indicate that the importance of various staff skills” within web-archiving initiatives did not change demonstrably from the previous survey (Farrell et al., 2018, p. 15).

Truman (2016) conducted a study to document international web-archiving programs (with a focus on cultural memory institutions), and “to identify common concerns, needs, and expectations in the collection and provision of web archives to users; the provision and maintenance of web archiving infrastructure and services; and the use of web archives by researchers” (p. 6). Truman’s methodology includes independent research and participation in working groups at conferences. It also includes semi-structured interviews or email communications with individuals from institutions ($n = 23$) in the United States, Europe, and New Zealand with web-archiving program (or institutions intending to commence a program), two service providers ($n = 2$), and researchers who use web archives ($n = 4$). Truman analyzes institutional web-archiving services, addressing challenges, integration with library collections, and more. From the findings, Truman offers 22 opportunities for future research and development, organizing them into four main themes as follows: Increase communication and collaboration; focus on smart technical development; focus on training and skills development; and build local capacity. Truman suggests that the opportunities may fall under one or more themes, but the number one theme is to increase collaboration and communication in several areas (Truman, 2016). The next ranked theme focuses on smart technical development, in terms of “Gathering requirements for next generation tools” (Truman, 2016, pp. 42–43). Training and skills development ranks as the third most prevalent theme in the findings, with Truman (2016) suggesting the “need for training and skills development

Exploring skills and training requirements

for existing staff new to web archiving” but also an opportunity to consider the need to develop training for website developers with considerations for the future archivability of websites (p. 43). With regard to building local capacity, Truman (2016) suggests the need for “augmenting an institution’s resources and proficiency in the area of web archiving and related services”, such as “the dedication of full-time staff to work in web archiving so that they can stay abreast of latest developments, best practices, and consequently fully engage in the community” (p. 43).

Byrne and Rarugal (2019) conducted a conference workshop, titled “Reflecting on how we train new starters in web archiving” at the International Internet Preservation Consortium Web Archiving Conference (WAC) in June 2019. The purpose of the workshop was for participants to be more aware of what they do and why they do it to create a trainee-centered learning environment (Byrne & Rarugal, 2019). This was achieved through a number of activities including individual polling, group brainstorming, ranking as well as participants identifying what is achievable within their current work environment (Byrne & Rarugal, 2019). There were $N = 26$ participants at the workshop from the international web-archiving community. Byrne and Rarugal (2019, 2020) found that 65% of workshop participants ($n = 26$) responded “no” to the question if there was a structured training program on web archiving at their organization. Therefore, it was not surprising when these participants were asked “how were you trained in web archiving?”, that hands-on training was the most popular training method used. No formal definition of hands-on training was given, but it was considered to be inclusive of a mix of in-house formal and informal training sessions, webinars, talks, conferences, and other events. As the importance of web archiving grows, so too does the need for training in this field but these responsibilities are falling on web archivists. However, the demands on web archivists’ time is always high and it is challenging to find adequate time to develop materials for a structured training program (Byrne & Rarugal, 2020). While Byrne and Rarugal (2019) point to the availability of a suite of training materials for a beginner level looking to get started in web archive curation (IIPC, n.d., Training Working Group), they suggest that before taking on new strategies it is important to understand our own beliefs on training and what actions we currently take when training new staff. The aim of the workshop was to observe and reflect on how participants were trained in web archiving, what was effective/not effective and what participants could achieve to make training a more formalized process within their institution. Byrne and Rarugal (2019) note how reflecting on these points can help us to become more aware of any biases we may have in terms of preferred training delivery style which could be contradictory to what the trainee really needs (Byrne & Rarugal, 2020).

The Web Archives — Researcher Skills and Tools Survey (WARST) project conducted a study, which focused on individuals ($N = 44$) around the globe who participate in web archive research, and explored the skills, tools, and knowledge ecologies in the web archive research lifecycle (Healy et al., 2022). The researchers organized the findings around two communities of practice being individuals from a Library, Archive, or Web Archive environment and individuals who identified with being a Scholar, Academic, Lecturer, Student, or working in an IT/Web Design environment. Participants were asked a number of questions, including the types of data they collect, their research outputs, the type of tools they use for data collection and data analysis, and questions regarding their skills, knowledge, and challenges for participating in the activities of the web-archiving lifecycle, from creation to end use. In many instances, participants were given the option to enter free text. The free

text was coded through the number of times a particular topic/theme was mentioned in participants' answers ($R/r=$). Due to limited space, we focus on the findings for participants who identified themselves as working in a Library, Archive, or Web Archive environment ($n = 30$).

In terms of challenges encountered when working with web archives, Library, Archive, or Web Archive environment participants ($n = 25$) offered multiple representations including inconsistencies and incompleteness, legalities for acquisition/access, technical challenges, challenges with learning new skills (e.g., steep learning curve), financial challenges, producing documentation/metadata, volume of data, and institutional challenges (Healy et al., 2022). Also of interest are challenges regarding the “complexity” of WARC files and having limited “technical skills to analyse the WARC-files and the information within them” (Healy et al., 2022, p. 98). Therefore, similar to challenges for data librarians, mentioned earlier, web archivists would benefit from training in data analytics/science to assist in the processing of large web archive collections. The study also points to challenges for the provision of access to archived web materials due to legalities for archiving web content in the first instance, as well as legalities for providing access to the preserved content, and how such legalities vary from country to country. This also presents challenges for transnational web archive collections due to varying legal deposit laws across different countries, which affect how the data is preserved, stored, accessed, used, and by whom (Healy et al., 2022). Therefore, this necessitates the development of solutions and training materials, which consider the perspectives of stakeholders who engage in transnational web-archiving projects.

Summing up some similarities and differences

The literature offers an overview of both the challenges for web archivists and librarians/archivists working in a digital environment and reveals that there are many similarities across these roles. For example, web archivists have similar challenges as data librarians and computational archivists and would benefit from training in data analytics/science to assist in the processing of large web archive collections. Web archivists are also required to use automation tools, such as web crawlers, metadata extraction tools, and quality control software to help manage the increasing volume of born digital content that needs to be preserved. Moreover, to remain relevant and effective in the digital era, web archivists must adopt similar strategies as those working in the digital environment to enable them to stay abreast of new developments in their field. However, web archiving is further complicated by “ever-evolving” web and internet technologies (Truman, 2016) which requires evolving workflows that not only deal with the complexities of the characteristics of the web and social media as communications technologies, but also the broader changes in technology and how society and legal frameworks adapt to these changes (Gooding et al., 2021; Healy & Byrne, 2023). There is also the challenge for how such legalities affect transnational web archive collections and their preservation, storage, and access requirements.

Training and education

There is no doubt that web archiving adds a new dimension to the role of the librarian or archivist. It is for this reason that several organizations have taken on some responsibility to develop training materials. In this section, we examine some of the training materials, which are freely available online in addition to some of the options available through formal

Exploring skills and training requirements

higher education qualifications. Insights on the take-up of these courses (e.g., numbers of students/participants) and/or any critical reflections on the usefulness of the courses from either the students or employers in the archive/library sector could be explored in a follow-up study.

Training materials

Funded by the Institute of Museum and Library Services (IMLS) LB21 program in 2019, the Continuing Education to Advance Web Archiving (CEDWARC) was a collaborative project by research investigators at Virginia Tech, Los Alamos National Lab, Old Dominion University, George Washington University, University of Waterloo and the Internet Archive. CEDWARC had the purpose of developing a continuing education curriculum for teaching library and archive professionals in the USA advanced web archiving and analysis techniques (CEDWARC, 2019). They offered in-person training workshops and developed multiple online training workshops, which are freely available to access online through the project website.

The Society of American Archivists (SAA) provides fee-based training modules for archivists and records managers, through in-person workshops and training modules in the SAA “On-Demand Webcast Catalog” on topics such as web-archiving fundamentals and managing digital records. Elsewhere, the IIPC Training Working Group and the Digital Preservation Coalition (DPC) collaborated to develop training materials, which aim “to provide a complete beginner’s level training course in web archiving” (IIPC, n.d., Training Materials; IIPC, n.d., Training Working Group). There are plans to expand on these materials, but this will depend on resources available to the group. The IIPC Web Archiving Conference (2022) had a session dedicated to training. The session called “Teaching the Whys and Hows of Creating and Using Web Archives” had four sessions, two from North America and two from Europe (IIPC, n.d., Q&A: Session 3, WAC 2022 program; Personal communications with Anonymous member, 2023).

In the United States, there have been a number of stand-alone web-archiving modules included in postgraduate courses. The University of North Texas (UNT) launched a web-archiving module in Spring 2022. The Github page for INFO 5960.001 — web archiving at UNT signposts similar modules offered by other universities, but this list is not exhaustive (Phillips, 2022). The Archiving the Black Web (ATBW) project identified formalized training as a critical requirement to ensure the sustainability of their project. In March 2023, ATBW were awarded funding from the Mellon Foundation to “diversify and increase the number of web archiving practitioners and collections that focus on the Black experience, as well as bring together scholars and archivists who want to increase understanding, collections, and scholarship about how Black people use digital communication technologies” (ATBW News, 2023). The lessons learned from this project will be important for the wider web-archiving community.

Curriculum review for web-archiving provision

In this section, we offer an overview of training opportunities that are available through the curricula of information studies courses in higher education institutions across the regions of Britain, Ireland, and Spain. We use this as a first step in measuring indicators as to the presence of web-archiving activities as part of course curricula. As mentioned in

the methodology section, between February and May 2023 the research team reviewed a number of online descriptions of the curricula for information studies across these three regions to see if there was any reference to web archiving. The data was collated on a spreadsheet and is published as a dataset on the British Library Shared Research Repository (2023, DOI: 10.23636/ss6t-0857). In terms of courses reviewed, the total number of courses reviewed is (63), and the ratio of courses per country is as follows: Britain (41), Ireland (7), and Spain (15).

Overall, the review sample (63) reveals that only four courses explicitly mentioned web archiving in their module descriptions. Of these courses, there were zero courses with stand-alone courses in web archiving, rather web archiving was just one element of a module. Across Britain, there were four examples of where web archiving was provided as one element of an optional module; however, it must be noted that there were only two institutions that offered these modules, and the same module was available to two courses at both institutions. Across Ireland and Spain, we found zero references to web archiving in the courses reviewed (British Library Shared Repository, 2023). These findings suggest web archives and web archiving as a topic of study lurks on the fringes of information studies.

Discussion

There are many parallels between those working with web archive content and the challenges for archivists working with moving image content in the essay “Education, Training and Careers in Moving Image Preservation” published in 2000.

Aside from the occasional summer school for new archivists, organized by FIAF and held at the Staatliches Filmarchiv der DDR, and the equally infrequent single-session university classroom surveys of film preservation, there had been no attempt to offer courses of study in film preservation and film archives administration. The obvious reasons were that there are insufficient positions available for would-be archivists, and most of the work involved in the running of a film archive is far removed from the world of academia. The major figures in the American film archival field trained themselves ... [they] started at the bottom and slowly graduated to administration.

(Lukow, 2000, p. 2)

This is not surprising. Muldoon (1997) points out how revolutionary transitions in information communications technologies (ICTs) have all been marred with a similar phenomenon, and that is “the loss of information that occurs when a new technology is introduced and supersedes an older one” (p. 50). In *The Dark Side of Technology*, Townsend (2016) proffers that technological innovations, which speed up the process for the production of information come with the trade-off that the survival time of the information diminishes. Examples include transitions from stone engravings to inks on vellum, from typed print to word processing files, and from negative film to photographs on a mobile phone. For Townsend (2016) born-digital information such as social media, blogs, and web pages is the category, which epitomizes his argument “that the faster information is produced, the shorter is the survival time” (Townsend, 2016, p. 206). Nonetheless, when it comes to technological innovations in information production, the focus is on the creation of new information and not the preservation of the old (Lyman, 2002). Therefore, training and education is inevitably more directed toward the latest ICT technologies, which is certainly

Exploring skills and training requirements

needed. However, there is also a need for the training and education of archivists and librarians, which reflect the reality with regard to the long-term preservation of information, especially with the rapid advances of internet, web, and social media technologies (Truman, 2016; Healy et al., 2022).

Web archivists need training in all aspects of the web curation lifecycle but also an understanding of web publishing in general and how this relates to the technical limitations of current web crawler technology. In addition, practitioners working with web archives also need an additional set of training on understanding end user requirements and developing training resources to meet these needs (Truman, 2016; Gooding et al., 2021; Healy et al., 2022).

Relative to other information management specialties the web-archiving community is very small and there are not very many web-archiving jobs advertised, some roles include web archiving as part of a wider job profile. Therefore, it could be argued that the universities are being reactive to the job market, rather than proactive to the needs of the academic field, especially when it comes to the issues regarding link rot and the research integrity of URL citations.

Since the late 1990s commentators have discussed the problem of reference rot, and how it undermines the foundations of scholarly citation and research reproducibility (Harter & Kim, 1996; Spinellis, 2003; Zittrain et al., 2014; Zhou et al., 2015; Bansal & Parmar, 2020; Major, 2021; Craigle et al., 2022). Indeed, Major (2021) notes how reference rot “jeopardizes the scientific value of official publications because the experiments described can no longer be reproduced and the secondary literature referenced can no longer be reread” (p. 5). Nonetheless, there has been no widespread movement across academia to address this issue, or to advocate for the use of web archiving as a potential solution (Healy, 2022; Healy et al., 2022).

Furthermore, referencing style guides of organizations such as the American Psychological Association (APA), Modern Language Association (MLA), and the Institute of Electrical and Electronics Engineers (IEEE) have not been forthcoming with solutions for safeguarding URL citations. Nor are they discussing the potential of web archiving as a somewhat obvious solution, whereby researchers would become DIY web archivists to ensure that their research and bibliographies remain reproducible and accessible (Major, 2021; Healy, 2022; Byrne et al., 2023). However, as it stands the value of web archiving across academic disciplines as well as policymakers has not yet been fully realized. Web archiving is an important field, which needs to be understood in its own right for three broad reasons: preservation (collection building), research uses (research development) as well as preservation of URL citations, and online research (research data management).

The lack of web archive provision in higher education courses also highlights the disconnect between practitioners and academics within the broader information management field. Therefore, a more collaborative approach is needed, similar to the WARCnet.eu transdisciplinary research network, which incorporates the views and knowledge of cultural heritage information professionals and technical experts; data scientists/librarians; academic educators, researchers, and postgraduate students; and is representative of individuals, who participate in the activities of the web-archiving lifecycle from the point of creation through to end use and reuse. Furthermore, WARCnet also takes a transnational research approach toward understanding the history of (trans) national web domains and transnational events on the web and fosters discussion to address the challenges for developing transnational web archive collections.

A transdisciplinary/transnational approach would not only be beneficial for the development of educational materials and courses for web archive provisions, but it would also enable discussions to develop frameworks to provide course modules for students in the use of web archives for research, training courses for educators on how to incorporate web-archived content as part of their teaching materials and methods, and foster discussions on the creation, preservation, and storage and access requirements for transnational collections.

It is important that web archiving plays a more prominent role within higher education curricula in the information management sector. However, not all of the training needs can be addressed here, as some of the key skills required are gained through experience of working with web archive material over a number of years. Nevertheless, equipping the next generation of information professionals with a basic understanding of web-archiving technology, curation, and research potentials would also improve efficiencies in the broader field of curation, preservation, and access to born digital information. To improve web-archiving training provision a benchmarking system is required for the various participants within the full web-archiving lifecycle. However, as mentioned above, this will need to be developed through a collaborative transdisciplinary network of key stakeholders, which considers the collection and use of transnational web archive collections. For the purpose of this chapter, we were only able to focus on web archive curation due to space limitations. The DPC have already developed their Competency Audit Toolkit (CAT) for general digital preservation and this template could be modified to fit the needs of the web-archiving community (DPC, 2022).

Conclusion

In conclusion, this chapter examined the challenges faced by library and archival roles in a digital context, with a particular focus on the skills and training requirements for practitioners involved in web-archiving activities within cultural heritage organizations. We found that the challenges faced by librarians and archivists who participate in web archiving as part of their role, are in many ways similar to the challenges for library and archival roles in the digital environment, and in both cases, there is a need for ongoing professional development. In other ways, however, web archiving presents some unique challenges, which requires additional or alternative skills and training. Our exploration also highlighted the need for a collaborative transdisciplinary/transnational approach to designing new curricula for web-archiving education, which considers the broad range of skills and training required by practitioners in the web-archiving community, and the need to foster discussion for transnational web archiving. It is also essential to understand end-user requirements and develop training resources that meet these needs. The international nature of the web-archiving community is implicit throughout our analysis, reflecting a shared global challenge in this aspect of digital preservation. Emphasizing a transdisciplinary/transnational perspective enriches our understanding and highlights the importance of international collaboration in developing solutions for web archiving, and the provision of web-archiving education going forward.

Acknowledgements

Authors thank to Grup de Recerca Consolidat “Cultura, Biblioteques i Continguts Digitals (CBCD)” (grant num. SGR-2021-00282/Generalitat de Catalunya).

References

- ATBW News. (2023, March 8). *Creating a Professional Training Program and Research Initiative on the Study of the Black web, Archiving the Black Web Receives \$2.5 Million Grant from the Mellon Foundation*. Archiving the Black Web (ATBW). Retrieved from <https://web.archive.org/web/20240120042913/https://www.archivingtheblackweb.org/blog/>.
- Bansal, S., & Parmar, S. (2020). Decay of URLs citation: A case study of current science. *Library Philosophy and Practice*, 1–9. Retrieved October 30, 2020, from <https://web.archive.org/web/20230801212429/https://digitalcommons.unl.edu/libphilprac/3582/>.
- Bingham, N. J., & Byrne, H. (2021). Archival strategies for contemporary collecting in a world of big data: Challenges and opportunities with curating the UK Web Archive. *Big Data & Society*, 8(1), 1–6. <https://doi.org/10.1177/2053951721990409>.
- Boté, J.-J., & Termens, M. (2019). Reusing data technical and ethical challenges. *DESIDOC Journal of Library & Information Technology*, 39(6), 329–337. <https://doi.org/10.14429/djlit.39.06.14807>.
- Boté Vericad, J.-J. (2012). *Propuesta de un modelo de preservación digital para pequeñas y medianas instituciones sanitarias*. Ph.D. Thesis, Universitat de Barcelona. Retrieved from <https://web.archive.org/web/20240309111948/https://www.tdx.cat/handle/10803/96254#page=1>.
- Bragg, M., & Hanna, K. (2013). The Web Archiving Lifecycle Model. White Paper, The Archive-It Team, Internet Archive. Retrieved from https://web.archive.org/web/20220627214546/http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf.
- British Library Shared Research Repository Dataset (2023). Review of Information Studies Courses in Higher Education for Web Archive Provision. Retrieved from <https://web.archive.org/web/20240906020718/https://bl.iro.bl.uk/concern/datasets/df61a8d9-e8a1-442f-973e-e987c6b6361e>.
- Brügger, N. (2018). *The Archived Web: Doing History in the Digital Age*. The MIT Press.
- Byrne, H., Cannelli, B., Noguera, C., Kurzmeier, M., & de Wild, K. (2023). *Looking Ahead: After Web (Archives)?* WARCnet Papers, August 2023. Retrieved from https://web.archive.org/web/20230902214653/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Looking_ahead.pdf.
- Byrne, H., & Rarugal, C. (2019, June 6). *Workshop: Reflecting on How We Train New Starters in Web Archiving*. International Internet Preservation Coalition General Assembly and Web Archiving Conference, Zagreb, Croatia, June 6–7, 2019 (IIPC WAC 2019). Retrieved from https://web.archive.org/web/20231203013219/https://netpreserve.org/ga2019/programme/abstracts/#workshop-new_starters.
- Byrne, H., & Rarugal, C. (2020, May 24). *Reflecting on How We Train New Starters in Web Archiving*. IIPC Blog. Retrieved from <https://web.archive.org/web/20230922081028/https://netpreserveblog.wordpress.com/2020/05/24/reflecting-on-how-we-train-new-starters-in-web-archiving/>.
- CEDWARC. (2019). *Continuing Education to Advance Web Archiving (CEDWARC)*. Github.io. Retrieved from <https://web.archive.org/web/20230929005134/https://cedwarc.github.io/>.
- Costea, M.-D. (2018). *Report on the Scholarly Use of Web Archives*. NetLab. Retrieved from https://web.archive.org/web/20230918163907/http://netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf.
- Craigle, V., Retteen, A., & Keele, B. J. (2022). Ending law review link rot: A plea for adopting DOI. *Legal Reference Services Quarterly*, 41(2), 93–97. <http://doi.org/10.1080/0270319X.2022.2089810>.
- Davies, L. (2022). *7 Skills Every Library and Information Science Professional Needs*. Retrieved from <https://web.archive.org/web/20240206082930/https://resources.noodle.com/articles/master-of-library-and-information-science-skills/>.
- Digital Preservation Coalition, McMeekin, S., & Currie, A. (2022). *DPC Competency Audit Toolkit Guide* (1st ed.). Digital Preservation Coalition. <http://doi.org/10.7207/dpccat22-01>.
- Farrell, M., McCain, E., Praetzelis, M., Thomas, G., & Walker, P. (2018). *Web Archiving in the United States: A 2017 Survey*. National Digital Stewardship Alliance Report. <http://doi.org/10.17605/OSF.IO/3QH6N>.
- Gomes, D., Miranda, J., & Costa, M. (2011). A survey on web archiving initiatives. In S. Gradmann, F. Borri, C. Meghini, & H. Schuldt (Eds.), *Research and Advanced Technology for Digital Libraries* (Vol. 6966). Springer. https://doi.org/10.1007/978-3-642-24469-8_41.
- Gooding, P., Terras, M., & Berube, L. (2021). Identifying the future direction of legal deposit in the United Kingdom: The Digital Library Futures approach. *Journal of Documentation*, 77(5), 1154–1172. <http://doi.org/10.1108/JD-09-2020-0159>.

- Harter, S., & Kim, H. (1996). Electronic journals and scholarly communication: A citation and reference study. *Information Research*, 2(1). Retrieved from <https://web.archive.org/web/20231201063126/https://www.informationr.net/ir/2-1/paper9a.html>.
- Healy, S. (2022). *The Future(s) of Web Archive Research Across Ireland*. PhD Thesis, National University of Ireland Maynooth]. Retrieved from <https://web.archive.org/web/20231028175130/https://mural.maynoothuniversity.ie/17558/>.
- Healy, S., & Byrne, H. (2023). Scholarly Use of Web Archives Across Ireland: The Past, Present & Future(s). WARCnet Special Report. Retrieved from https://web.archive.org/web/20230902214739/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_Byrne_Scholarly_Use_01.pdf.
- Healy, S., Byrne, H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., & Jansma, R. (2022). Skills, Tools, and Knowledge Ecologies in Web Archive Research. WARCnet Special Report. Retrieved from https://web.archive.org/web/20230902214716/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Skills_Tools_and_Knowledge_Ecologies.pdf.
- Holownia, O. (2020, June 15). *Launching IIPC Training Programme*. IIPC Blog. Retrieved from <https://web.archive.org/web/20230922080839/https://netpreserveblog.wordpress.com/2020/06/15/launching-iipc-training-programme/>.
- International Internet Preservation Consortium. (n.d.). *Training Materials*. Retrieved from <https://web.archive.org/web/20240304032427/https://netpreserve.org/web-archiving/training-materials/>.
- International Internet Preservation Consortium. (n.d.). *Training Working Group*. Retrieved from <https://web.archive.org/web/20231127152801/https://netpreserve.org/about-us/working-groups/training-working-group/>.
- International Internet Preservation Consortium. (n.d.). *WAC 2022 Program*. Retrieved from <https://web.archive.org/web/20231003105634/https://netpreserve.org/GA2022/wac/>.
- Lomborg, S. (2019). Ethical considerations for web archives and web history research. In N. Brügger & I. Milligan (Eds.), *The SAGE Handbook of Web History* (pp. 99–111). Sage Publications.
- Lukow, G. (2000). Education, Training and Careers in Moving Image Preservation. AMIA the Association of Moving Image Archivists. Retrieved from https://web.archive.org/web/20150723182609/http://www.amianet.org/sites/all/files/ed_training_careers.pdf.
- Lyman, P. (2002). Archiving the World Wide Web. In Council on Library and Information Resources & Library of Congress (Eds.), *Building a National Strategy for Preservation: Issues in Digital Media Archiving*. Council on Library and Information Resources and the Library of Congress.
- Major, D. (2021). The problem of Web Ephemeria. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 5–10). Springer International Publishing.
- Marciano, R., Lemieux, V., Hedges, M., Esteva, M., Underwood, W., Kurtz, M., & Conrad, M. (2018). Archival records and training in the Age of Big Data. In J. Percell, L. C. Sarin, P. T. Jaeger, & J. Carlo Bertot (Eds.), *Re-envisioning the MLS: Perspectives on the Future of Library and Information Science Education* (Vol. 44B, pp. 179–199). Emerald Publishing Limited. <http://doi.org/10.1108/S0065-28302018000044B010>.
- Muldoon, J. (1997). Communications revolutions: Little history, much myth. In J. B. Gardner (Ed.), *Report on Documenting the Digital Age Conference*. Retrieved from <https://web.archive.org/web/20180524134523/https://www.historyassociates.com/wp-content/uploads/2015/09/Documenting-the-Digital-Age.pdf>.
- Niu, J. (2012). An overview of Web Archiving. *D-Lib Magazine*, 18(3/4). <http://doi.org/10.1045/march2012-niu1>.
- Olumide, O. H., Moyosore, S. A., Oluwaseyi, M. J., & Esther, E. (2022). Librarians' problems/challenges in the digital era: Focus on Southwest, Nigeria. *Engineering, Technology and Environment*, 10(1). Retrieved from <https://web.archive.org/web/20220630124948/https://seahipaj.org/journals-ci/mar-2022/IJIISTR/full/IJIISTR-M-4-2022.pdf>.
- Perret, R. (2023). Candles burning at both ends: Experiences of dual-role archivist/librarians. *Journal of Western Archives*, 14(1). Retrieved from <https://web.archive.org/web/20231010041834/https://digitalcommons.usu.edu/westernarchives/vol14/iss1/4/>.
- Phillips, M. (2022). *Web Archiving Course*. GitHub. Retrieved from <https://web.archive.org/web/20230918154000/https://github.com/vphill/web-archiving-course>.
- San José School of Information (SJSU-SI). (2023). LIS Careers Trends Report. Retrieved from <https://web.archive.org/web/20230921213540/https://ischool.sjsu.edu/lis-career-trends-report>.

Exploring skills and training requirements

- Schneider, S., Foot, K., & Wouters, P. (2009). Web archiving as e-research. In N. W. Jankowski (Ed.), *E-Research: Transformation in Scholarly Practice* (pp. 205–221). Routledge.
- Semeler, A. R., Pinto, A. L., & Rozados, H. B. F. (2019). Data science in data librarianship: Core competencies of a data librarian. *Journal of Librarianship and Information Science*, 51(3), 771–780. <http://doi.org/10.1177/0961000617742465>.
- Society of American Archivists (SAA). (n.d.). *On-Demand Webcast Catalog*. Retrieved from <https://web.archive.org/web/20240205204235/https://www.pathlms.com/saa/events/873>.
- Spinellis, D. (2003). The decay and failures of web references. *Communications of the ACM*, 46(1), 71–77. <http://doi.org/10.1145/602421.602422>.
- Sweetser, M., & Orchard, A. A. A. (2019). Are we coming together? The archival descriptive landscape and the roles of archivist and cataloger. *The American Archivist*, 82(2), 331–380. <http://doi.org/10.17723/aarc-82-02-18>.
- Townsend, P. (2016). *The Dark Side of Technology*. Oxford University Press.
- Truman, G. (2016). *Web Archiving Environmental Scan*. Harvard Library Report. Retrieved from <https://web.archive.org/web/20231021025944/https://dash.harvard.edu/handle/1/25658314>.
- Vassilakaki, E., & Moniarou-Papaconstantinou, V. (2017). Beyond preservation: Investigating the roles of archivist. *Library Review*, 66(3), 110–126. <http://doi.org/10.1108/LR-09-2016-0077>.
- Virkus, S., & Garoufallou, E. (2019). Data science from a library and information science perspective. *Data Technologies and Applications*, 53(4), 422–441. <http://doi.org/10.1108/DTA-05-2019-0076>.
- Wikipedia. (2011). List of Web archiving initiatives. Wikipedia. Retrieved September 18, 2021, from https://web.archive.org/web/20240203174212/https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.
- Wright, T., & Benoit Iii, E. (2019). Technology use in designing curriculum for archivists: Utilizing andragogical approaches in designing digital learning environments for archives professional development. *Preservation, Digital Technology & Culture*, 48(2), 85–94. <http://doi.org/10.1515/pdtc-2019-0005>.
- Zhou, K., Grover, C., Klein, M., & Tobin, R. (2015). No more 404s: Predicting referenced link rot in scholarly articles for pro-active archiving. *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 233–236. <http://doi.org/10.1145/2756406.2756940>.
- Zittrain, J., Albert, K., & Lessig, L. (2014). Perma: Scoping and addressing the problem of link and reference rot in legal citations. *Legal Information Management*, 14(2), 88–99. <http://doi.org/10.1017/S1472669614000255>.

16

TEACHING WEB ARCHIVING IN HIGHER EDUCATION

Best practices and future perspectives

Márton Németh

Introduction

Teaching the creation of web archives, using web archiving techniques in a research environment, and using archived web content as research materials in universities are essential to preserving our digital legacy, training future information professionals, and fostering interdisciplinary collaboration.

Web archiving education has various goals that focus on different aspects. A primary focus, primarily relevant for libraries and archives, is learning to preserve the various segments of the World Wide Web. Another objective is to learn how to apply archiving tools for specific purposes, for example in digital humanities research projects. The third focus is on using archived web materials for research, for example, in history. It is therefore worthwhile understanding how curricula in higher education can integrate various disciplinary perspectives about web archiving. Universities can contribute to a responsible and inclusive stewardship of our digital cultural heritage by equipping students with the necessary theoretical, and technical skills, competencies, and ethical and legal considerations. As web archiving becomes integral to academic programs, universities pave the way for a future where the digital realm and its historical significance are safeguarded for future generations.

This chapter introduces potential research and education areas, topics, and subjects related to these various contexts of web archiving. Accordingly, it is centered around the following research questions: What kind of disciplinary characteristics can be realized in current higher education programs related to the education of web archiving and the use of archived web materials? Is there any interdisciplinary collaboration that can be achieved in these contexts? How can we evaluate higher education programs with different disciplinary perspectives and collaboration forms related to web archiving and use archived web content mainly from North America and Europe? In which way newly emerging curricula frameworks in universities can lead to discussing disciplinary and interdisciplinary opportunities and challenges?

Methodology

The research design is entirely qualitative. At first, various disciplinary perspectives are described. Some best practices are introduced by analyzing course concepts, curricula, and module descriptions. In the first phase, general databases (EBSCO, ProQuest Central, ACM) were used to find related articles for teaching web archiving, and for teaching to use archived web content. Unfortunately, very few articles were found. By the next step, a wide range of educational programs was found by typing some basic questions used by search engines (Google and Bing) for the following keywords by various combinations: web archiving, long-term preservation, higher education, information science, and information technology (IT). In the second phase, I evaluated these hits and refined them by relevance and location. At this point, it has become clear that the major geographical focus should have points on Europe and North America. The final selection of programs is based on one hand disciplinary and geographical diversity. In addition, I will showcase relevant examples in various disciplinary contexts and describe how web archiving and using archived web materials can appear in different curricula environments. Some programs are selected because they offer interdisciplinary perspectives, combining different research contexts (for example gaining competencies about using Information Communication Technology (ICT) tools, studying theoretical aspects of archiving, and learning to manage small-scale archiving projects to study the web artifacts of complex social phenomena, then as the result building up complex narratives).

Disciplinary perspectives

Education of web archiving in universities can appear in different disciplinary contexts for various target groups of future professionals. The archived web can be defined as a significant research subject itself (Brügger et al., 2020; Milligan, 2020; Weber, 2020; Healy et al., 2022; Maemura, 2022). Librarians, archivists, information scientists, professionals in digital humanities, data scientists, and IT developers can work together on analyzing large, archived web corpora focusing on several structural and content-based features. Moreover, web archiving can be defined as a critical method for the future of digital research (Weber, 2020). A special report identifies and documents the skills, tools, and knowledge required to achieve a broad range of goals within the web archiving lifecycle and to explore the challenges of participation in web archive research (Healy et al., 2022). New scientific disciplines, such as web history, have emerged through these research activities in the past ten years (Brügger, 2009, 2013; Milligan, 2020). We can focus on creating, running, and managing web archives in public collections (Maemura, 2022, 2023).

A second disciplinary perspective treats archived materials as historical resources; web history is a particular sub-field.¹

Web archives can appear as resources for the history of communication. From this perspective, studying archived web content can be the focal point of communication science and media studies.

From a computer science perspective, the central focus points can be web archiving infrastructure, web archives as big data resources, and making web archiving more effective with artificial intelligence (Dubin et al., 2009; Fafalios et al., 2020). Digital Humanities scholars can focus on extracting links, and textual resources from web archives, analyze them, and create information visualizations about the results of analyses (Brügger & Finnemann, 2013; Brügger, 2021).

An overview of web archiving programs and courses

While the specific inclusion of web archiving in university curricula may vary, several universities worldwide have recognized this field's importance and incorporated it into their programs.

It is essential to note that including web archiving in university curricula is more expansive than the examples listed below. Many other universities and academic programs worldwide have recognized the significance of web archiving. Web archiving courses are integrated primarily in library science, information management, digital preservation, archival studies, and media studies. These courses' specific structure and focus may vary, but they aim to provide students with a comprehensive understanding of web archiving principles, techniques, and ethical considerations.

Information management, library, and information science

We start this inventory of teaching programs with two examples from the USA. Both are related to Library and Information Science master programs. The topic appears on the course level. These offer theoretical foundations; however, they are mainly based on learning by doing principles.

The University of California, Los Angeles (UCLA) offers a course in the Archival Studies specialization as part of its Master of Library and Information Science (MLIS) program about electronic records management that includes web archiving.² The course introduces students to web archiving principles, tools, and techniques and provides hands-on experience capturing and preserving web content. Web archiving appears in other specializations, such as library studies, media archiving, and informatics.

The University of Michigan's School of Information offers a course called "Web Archiving"³ as part of its Master of Science in Information (MSI) program. The course covers practical topics such as web crawling and metadata extraction. It offers a bit broader scope of the subject than the course at UCLA, as it focuses explicitly on building up web archiving strategies and managing web archiving projects and has a unique viewpoint on ethical considerations.

In Canada, web archiving appears in the digital data management and preservation context. The Faculty of Information at the University of Toronto offers courses called "Digital Preservation and Curation"⁴ and "Special Topics in Information: Digital Archives Workflows"⁵ as part of its Master of Information (MI) program.⁶ Web archiving appears in digital data management and preservation, as it is helpful for digital data experts to have specialized knowledge in this field. In this way, the general course includes specialized modules on web archiving, where students learn about the challenges and methodologies involved in capturing and preserving web content. The master's program offers several specializations. Web archiving is appearing in other courses of some specializations in different contexts. These are Archives and Records Management, Culture and Technology, and Human-Centered Data Science. Archived web data has become a significant resource for records management in archives; those competencies and skills can analyze datasets the data analytics course can acquire. The two-semester-long culture and technology courses can explore the theoretical foundations and social and ethical aspects of web archiving.

In Scotland, the University of Glasgow offers a master's program called "Information Management and Preservation."⁷ The course covers various aspects of digital preservation, including web archiving techniques and strategies. Information Studies at the University

Teaching web archiving in higher education

of Glasgow combines methods and research on the practices and collections of museums/galleries, archives/records, and libraries. They also focus on the overlap of skill sets required of cultural heritage practitioners working in a digital age with the theory and practice of digital humanities. Web archiving also appears in different courses in different contexts, like the previous case study from Toronto.

The Master of Digital Library Learning was an Erasmus Mundus program available from 2007 to 2020 in collaboration with Tallinn University in Estonia, Oslo University College in Norway, and the University of Parma in Italy.⁸ The author of this chapter graduated from this program in 2013 and possesses first-hand experience with the program structure and subjects' curricula. The program covered various technological, theoretical, and practical aspects of digital libraries. Web archiving and digital preservation were integrated into different main modules. Students generally learned about collecting, managing, preserving, and providing access to digital resources, including web content. In this way the use of archiving methods, tools, and the use of archived web materials for research also becomes relevant.

Archival Science and Record Management

Some programs specialized in teaching archivists offer courses about web archiving in the context of record management. These include modules on digital preservation and web archiving. Students explore the challenges of preserving digital records and web-based materials. Archived web material is appearing as a new document type besides born-digital, digitized, and traditional documents broadening the horizon of conventional record management with necessary new competencies and skills for future archivists. The University of London offers a one-year-long master's program called Archives and Records Management.⁹ Web archiving appears in the module called Creation and Capture. It introduces students to records management approaches and techniques (related to archived web materials, among other record types) that can be applied to information created, received, and used by organizations in a rapidly changing digital technology world.

Media Management, Media Studies, and Digital Humanities

Web archiving is increasingly incorporated into History, Museology, or Media Studies education programs. Some universities have started integrating web archiving into their humanities and digital humanities programs, recognizing the importance of preserving and applying digital sources for historical research. In most cases, this subject appears in a different explicit way than in the agenda for information professionals. The focus is mainly not on the archiving theories and practice but on using web archives for research purposes. In the digital humanities context, the archiving methods and tools are also relevant to managing research projects based on archiving and then analyzing archived web materials. More and more departments and institutes are building web corpora to make new perspectives in their research. Mainly potential appliance, integrated ways of archived web materials to research are described in the following in different program and course contexts.

Web archiving is more multi-disciplinary and is a subject for future information professionals. It can still occur as a specific area of digital preservation, as in the previous examples but in a different disciplinary context. Master of Arts in Digital Asset and Media Management (DAMM)¹⁰ program offered by the University of London includes modules

on digital preservation and on web archiving. Students gain knowledge and practical skills in managing digital assets and preserving digital content in a digital scholarship context.

A similar approach is evident in the Netherlands. Master of Media Studies by the Preservation and Presentation of the Moving Image¹¹ at the University of Amsterdam (UVA) offers a one-and-a-half-year long master's program covering various aspects of digital preservation, including web archiving. Students learn about preserving and presenting digital audio-visual materials, including web-based content. The program seems to be unique in Europe due to its combination of the in-depth study of critical issues and practice-based learning, its international orientation, and its broad outlook. You can have the chance to study different types of moving images and sound, like cinema, broadcasting, and media art (all in their analog and digital manifestations) and their relations to each other. Digital media preservation issues such as web archiving are appearing in a broad context with several other aspects of media studies.

The UVA offers another one-year master's program called Media Studies: New Media and Digital Culture.¹² It provides a comprehensive and critical approach to new media research, practices, and theory. An ongoing research project, *The Web that Was*, focuses on the web's history from social and historical perspectives. History and social aspects of various digital platforms and ecosystems appear together with archiving challenges to use archived materials from these platforms and ecosystems for future researchers. Students can focus on building up different narratives based on archived research materials about various perspectives on the history of the internet. The program offers a unique perspective to mix analysis of historical context with current social theories and analysis of new media and platform capitalism.

The two-year research master's in media studies program of UVA offers an even broader perspective to combine archiving and preservation issues with historical research, media research, and social analysis. Each student can build their project to reflect media as complex social and cultural phenomena in their historical and contemporary contexts with the intersection of information, technology, and people.¹³

Furthermore, another elective course for all the master students at the Faculty of Humanities called Museums, Heritage and Digital Curation¹⁴ focuses on the use of various born-digital materials (like archived web content) or digitized content to search new narratives in the practice of museums by stimulating participation and dialogue: internally, locally, nationally or internationally to illustrate society's search for greater inclusivity and multivocality.

The Digital Humanities Institute at the University of Sheffield offers a master's program called Digital Culture and Communication.¹⁵ The focus is on how technologies transform the human experience and the cultural contexts of digital media developments through historical, critical, and theoretical perspectives. Students can investigate how digital content and theories are used in culture and heritage, how digital projects are managed, and how digital content is used to represent human culture and society in the past and present, such as ancient documents, artifacts, and environments, as well as present-day social media, digital arts, and virtual worlds. The appliance of archived web content as a significant source of digital cultural heritage may appear related to archiving theories and, management and historical representations, social and cultural context.

The Cultural Data Management and Communication master program¹⁶ focuses on the concepts and methods for how cultural data is created, managed, analyzed, and communicated. As live and archived web data represent and influence culture, it is essential to understand how the modern world works and how to create more value from it. A module focuses on linguistic analysis of cultural data; archived web resources can be used effectively

in this context. In another modular context called Designing Cultural Data Products, a significant issue may be realizing how current cultural products can be made using archived web materials. In this framework, the only challenge can be the relatively high level of technical expertise to do research with archived web data by the current research tools. However, the development in this field shows us that the entry barriers will be significantly lower shortly (for example by the appearance of introductory Jupyter Notebooks for historians, supported by the International Internet Preservation Consortium¹⁷). It makes it possible to use archived web content in research projects in non-technical and not information science-oriented environments such as these programs from Sheffield offer. The Digital Humanities Institute puts a great emphasis on publishing their results on specific websites that are archivable, and they are archiving their websites by themselves.

A concrete research product is also available.¹⁸ The Digital Humanities Institute developed a Dewdrop tool to find digital orphans. Those online assets (research resources) are deemed undiscoverable, unused, unknown, or forgotten by the broader research community because they need to be made more visible to standard discovery mechanisms, such as search engines, subject catalogs, aggregation sites, and other subject-specific websites. The tool can develop a discovery-friendly version of a resource's textual content at the record level. As a set of optimized data records, it can mediate between the resource and discovery services. The tool achieves this in two ways: a crawler retrieves a copy of the resource's textual content, including data in databases, and an analyzer generates discovery-friendly records from the content using natural language processing techniques.¹⁹

The Centre for Internet Studies (CFI) at Aarhus University, Denmark, was established on September 18, 2000, to promote research into the Internet's social and cultural implications and functions. It is at the School of Communication and Culture, Faculty of Arts.²⁰ Niels Brügger, professor, and his team have pioneered fostering new research directions in the history of the internet as a means of communication, analyzing the online artifacts of current leading social and political events, and analyzing the history of the Danish domain of the web. The Centre has hosted Ph.D. projects in these fields; it has become a meeting place and forum for local and international researchers working with web-archiving related research within the humanities and social sciences and serving as a center of knowledge in this field. The NetLab and DigHumLab project that finished at the end of 2022 made a considerable contribution to familiarizing graduate students, Ph.D. students, researchers, and public collection professionals with web archiving and research on archived web materials in Denmark (through various courses and publications) but also in Europe by organizing online courses in English. The author of this chapter participated in an introductory online course on web archiving together with colleagues from Belgium. This practice-based course offered a major help in forming the foundations of a permanent web archiving service at the National Széchényi Library, in Budapest, Hungary.

The Digital Humanities Department at the Faculty of Arts, Eötvös Loránd University (ELTE), Hungary, offers undergraduate and graduate students an introductory course in Digital Humanities. It is grounded in the fundamental theoretical and practical aspects of digital humanities. The course provides a solid base of academic background, using various tools focusing on their history and their literary and artistic depictions. Digitally generated content and web archiving modules examine the essential differences between digitally generated and analog counterparts. It also looks at the benefits of preserving the past web through web archiving. It offers some basic skills and competencies to do small-scale archiving projects in a digital humanities research environment.²¹ The department provides

free training programs for doctoral students from ELTE about managing digital cultural heritage, handling, archiving, publishing, and describing cultural objects such as web content.

Computer Science, Informatics

We can find only a few examples in Computer Science, Informatics, and Data Science to incorporate web archiving or using web collections as resources into curricula. Different web ecosystems, platforms, and web archives of public collections can offer a large base of unstructured data or structured datasets for data mining, visualization, and analysis. The potential of working with data from the web is enormous to familiarize students with robust data science workflows and teach them statistical data analysis.

At first, a case study shows the potential appliance of web scraping in IT education. As Dogucu and Çetinkaya-Rundel point out, web scraping allows more significant amounts of data to be collected in a shorter period and in an automated fashion that minimizes errors. They focus on teaching web scraping in undergraduate data science and statistical data analysis courses. Web scraping techniques allow students to gather unstructured information from the internet and convert it into a structured format. “There are two types of web scraping. The first is screen scraping, where you extract data from the source code of a website with an HTML parser or regular expression matching. The second is using application programming interfaces, commonly referred to as APIs. This is where a website offers a set of structured HTTP requests that return JSON or XML files” (Dogucu & Cetinkaya-Rundel, 2021, p. 112). The authors offer tools to teach web scraping and clean, analyze, and visualize structured datasets. They point out that the web offers current and interesting data sources. Furthermore, not just the data itself is interesting but the story behind why it is relevant to scrape these data. It can enforce motivation for students to work with web-based resources. Web scraping can bring data science to the classroom by solving real problems, for example, extracting data from a web page given by its URL (Dogucu & Cetinkaya-Rundel, 2021, pp. 119–120). In some situations, the dataset contains exactly what the students want and fits exactly to their research aims. In some other cases, the structured format of data limits the options of students. It means that their data analysis projects are limited by the availability of structured data instead of what they want to investigate (Dogucu & Cetinkaya-Rundel, 2021, p. 120). The authors cannot envision a data science curriculum without web scraping. The newly acquired skills by harvesting web data can be integrated into individual small-scale research projects, even for undergraduate students. “Teaching the concept and mechanics of iteration early on within the context of web scraping paves the way nicely for teaching statistical topics that require iteration on a later phase of a course.” “These are already difficult topics for students, so not having to teach the programming mechanics simultaneously as the statistical concepts is a nice benefit of introducing iteration early” (Dogucu & Cetinkaya-Rundel, 2021, p. 120). Another important aspect is to describe the ethical element of scraping data. An open question is whether it is ethical to make data harvests from a designated resource and whether the owner allows this activity from a website.

The MSc Data Science program at the University of Southampton offers a course called Social Media and Network Science.²² In this course, students will look at social media to analyze its use, see how conceptualizing social media as networks allows us to use techniques from network science as part of that analysis, and explore some of the personal and social issues that have arisen and the technologies that might address them. The course offers an overview of the history of social networking technologies and the web, highlighting connections

Teaching web archiving in higher education

among linked data and the social semantic web, discussing issues of graph theory, game theory, and social networks, and methods to analyze data from social media. They highlight the issues of digital literacy, network dynamics, privacy, identity, power, and trust in social networks. As these networks are mainly closed platforms, retrieving and analyzing content with the help of data science and network science is critical to access a significant part of our cultural heritage in the future.

The University of Sheffield offers a Data Science MSc program accredited by CILIP, the UK's Library, and Information Science Association.²³ The program provides significant competencies and skills in data visualization, machine learning data analytics, database design, and a core module focusing on data and society. One can find modules on Human–Computer Interaction and researching social media among electives. This module examines the key theoretical frameworks and methods used in social media studies by describing research methods and tools to study social media and focusing on the social and ethical context. This program can be ideal for future web archivists who want to be information and data specialists. For those who want solid technical skills and competencies, one hand can analyze and visualize archived web content to run data mining projects on large corpora. On the other hand, students can learn to manage information resources and use social narratives to build up complex research projects by working with professionals with information science, social science, and humanities backgrounds.

In Hungary, a new three-semester specific course related to data stewardship started in autumn 2023 with the collaboration of the Digital Humanities Department, Faculty of Arts, and Faculty of Informatics, Loránd Eötvös University (ELTE) in Budapest. The focus is on data management, data curation, and data security. In various modules, the methods of web archiving and managing archived web content as digital data appear.²⁴

Discussion

Web archiving can be an essential part of educating future information specialists. By integrating web archiving into their curricula, universities empower students to become proficient in this specialized field. In this context, the focus is on creating and managing web archives.

Teaching web archiving equips students with the necessary skills to capture and preserve this digital cultural heritage, safeguarding it for posterity. Web archiving courses can be incorporated into library science, information science, archival science, museum studies, information management, or digital preservation programs, providing students with a comprehensive understanding of archival principles, tools, and techniques specific to the web.

Web archiving requires technical expertise in web crawling, scraping, metadata extraction, and digital preservation. Teaching web archiving for future information professionals enables students to gain hands-on experience with various tools and technologies employed in the archiving process. They learn to select, capture, curate, and store web content, ensuring its authenticity, integrity, and long-term accessibility. Moreover, students can acquire data analysis, content curation, and information retrieval skills, making them valuable assets in the evolving field of digital preservation.

“The absence of efficient and meaningful exploration methods of the archived content is a significant hurdle in turning web archives into usable and valuable information resources” (Fafalios et al., 2020, p. 1). A major challenge in information science is adapting semantic web tools and methods to web archive environments. The web archives must be a part of the

linked data universe with advanced query and integration capabilities and be directly exploitable by other systems and tools.

In this context, archived web content appears as a historical resource. The object of web history research can be a history of the web as a technical infrastructure; the history of the web as a communication and publication platform; history of a specific topic, event, institution, person, etc., as it was reflected on the web; archived textual and visual web content or webserver logs as subjects of extensive data analysis (e.g., for machine learning, for analyzing user characteristics). The level of research according to Niels Brügger's model can focus on individual web pages, individual websites, a specific range of domains, or the whole web sphere (Brügger, 2009).

In informatics, web archiving can appear as a challenging subject related to developing new archiving models, workflows, and software products and discussing managing archived web data from an ICT perspective.

Web archives, as large corpora, can be an education and research subject of several projects in the data science field. The concept of linked and open data has led to the need to quickly process large amounts of semi-structured data in web archives and retrieve valuable information. Data sources can be web and transaction data (e.g., log data, geolocations); structural data (e.g., link graphs); and content data (e.g., textual or visual information). In this context, a new way of collaboration can be formed among public collections, web archivists, and data scientists. The focus on data mining can be on web usage mining, web structure mining, and web content mining (Fafalios et al., 2020).

Web archiving can appear in the context of using artificial intelligence. A plan can be essential to create new workflows that make web archiving more effective. Other AI-related aspects of education and research in web archiving context can be automatic language detection of web content; entity detection; connecting entities with Wikidata; speech-to-text transformation; adding automated keywords and other metadata elements to archived web content; image recognition (helping to retrieve archived web content more comprehensively).

A significant benefit of web archiving as an education field is that it can transcend disciplinary boundaries and requires collaboration (Brügger, 2021; Healy et al., 2022) among diverse areas such as computer science, information science, history, social sciences, and humanities. By offering web archiving courses, universities encourage interdisciplinary collaboration and foster partnerships between researchers, librarians, archivists, and technologists. Such collaborations facilitate innovative research, promote cross-pollination of ideas, and contribute to a holistic understanding of the web as a cultural artifact.

Web archiving offers a wealth of research opportunities across various academic disciplines. In this way, the answer to the research question directing to the existence of these collaboration forms is entirely positive. By incorporating web archiving and the methods of using archived web data into university programs at the master's and Ph.D. levels, students gain access to vast digital collections for research and analysis. They can explore the evolution of ideas, track social and cultural phenomena, examine historical events, and analyze changes in online discourse (Ogden & Maemura, 2021; Ogden et al., 2021; Schafer & Winters, 2021; Maemura, 2023). Web archives become invaluable resources for graduate and postgraduate research, enriching the academic landscape and contributing to knowledge creation.

In education and research, web archiving raises ethical and legal considerations, such as copyright issues, privacy concerns (Gambell et al., 2021; Schafer & Winters, 2021; Corrado, 2022), and the contextual capture of dynamic web content. Teaching web archiving in universities allows for discussions on these ethical dimensions, enabling students to navigate

Teaching web archiving in higher education

the complexities and dilemmas inherent in web archiving. By understanding and adhering to ethical guidelines, students can ensure responsible and inclusive archiving practices that respect the rights of content creators and users.

In the final part of the discussions, I am reflecting on the research questions. The first research question focused on the characteristics of current higher education programs related to the education of web archiving and the use of archived web materials. During the evaluation of the different programs, certain connections could be found mainly related to information science, archival science, media management, media studies, and digital humanities. Discussions about web archiving and archived web content have been less integrated into the History and IT programs. Web archiving is solidly integrated into multiple programs focusing on information professionals in library and archiving fields. Media management and media studies programs also can integrate web archiving and working with archived web materials in many ways into curricula. The integration of the use of archived web resources could be more concrete in the history programs of study at the master's level. However, at the Ph.D. level, many projects can be found focusing on web history issues. Web history as a discipline has emerged with its peer-reviewed journal, and there is a good hope in the author's view, that it will be integrated into graduate history curricula shortly. In ICT, data science education shows the most significant potential benefits of integrating web archiving into curricula. Web archive collections can be ideal objects for data science projects. However, we still cannot find as many examples of this curricula integration as the opportunities emerge in this field.

A research question focused on the evaluation ways of different higher education programs related to web archiving and archived web content. It has become clear that in each disciplinary framework that I could describe working with web archiving and archived web content can open new theoretical and practical perspectives, creating new ways to study the development of the web, archiving methodologies and practices, offering various perspectives to discover archived web artifacts, and presenting the ways of describing various historical interpretations based on archived web resources.

Related to the last research question, about describing the various ways of interdisciplinary collaborations to web archiving and archived web content, by analyzing different curriculum conceptions, it has become clear that interdisciplinary collaboration is essential on the one hand to get archiving competencies and skills to work with archived web data and run various research programs related to any of the described disciplines, on the other hand, to describe the human experience and the cultural contexts of web content through historical, critical, and theoretical perspectives. We should notice that the emerging ways of archiving practice and the wider cultural, and social conceptions of working with archived web content are reflecting each other in more and more ways, offering various possibilities of colorful interdisciplinary collaboration forms among education programs in different disciplinary fields. Creating new tools just as Jupyter Notebooks for Historians and integrating them into the collections of various web archives can offer easier access to archived web content for research and education. Behind these developments joint work of web archivists, IT developers, and researchers is essential.

Conclusion

This chapter introduced the potential benefits of integrating on one hand web archiving, on the other hand, competencies, and skills to analyze archived web content into higher

education curricula at various levels and in different contexts. The described case studies show the relevance of web archiving in different environments and the depth of integration to Media Studies, ICT, Library and Information Science, and Digital Humanities. Media in general can appear in a complex social and cultural phenomenon in their historical and contemporary contexts with the intersection of information, technology, and people. Technologies can convert the human experience and the cultural contexts of digital media developments through historical, critical, and theoretical aspects. In another context, we can focus on the education of web archiving in a framework that describes how cultural data is created, managed, analyzed, and communicated. Cultural data in web archive collections can be used as a basis for linguistic analysis. Datasets from web archives can be useful resources for data mining projects.

The above-mentioned frameworks also show that interdisciplinary collaboration appears in the design of several programs, modules, and courses among media studies, information studies, digital humanities, and informatics. We are using ICT tools to manage datasets from web archive collections. With these datasets, we can analyze the digital artifacts of historical movements, make linguistic analyses on archived web content, and so on. Then we can put the results of different kinds of research activities into a common narrative. A major research question of this study focuses on whether it is possible to find best practices in the USA and Europe for integrating web archiving and studying archived web content into curricula in different environments. The evaluation of selected higher education programs confirms that web archiving has already been integrated into different programs and modules in multiple contexts in various countries.

Researchers and students can explore the evolution of ideas, follow social and cultural phenomena, examine historical events, try to archive major artifacts of online artists, and analyze changes in online discourse. A major task of education programs is to make it realize that web archiving is not just a merely technical issue, as the public often thinks at first glance. Furthermore, incorporating the ethical and legal context (copyright, privacy) of web archiving into various programs and modules is also essential to realizing the social responsibility of developers, social scientists, information scientists, and historians.

This chapter is an initial attempt to describe the appearance of web archiving education in different curricular contexts. The horizon of investigation can be broadened later focusing on Asia, Latin America, and Africa. More resources can be found to offer a more complex and comprehensive comparative analysis. A major limitation of the study is that it focuses mainly on master's and Ph.D. levels. Using the methodology and practical tools related to web archiving and studying archived web content in my view needs some prior basic general experience at the bachelor's level from one or more related disciplines. (However, it is important to highlight the importance of preserving our digital legacy at the bachelor's level as well.) The other main limit of the research described in this chapter is geographical. The focus is mainly on North America and Europe, mostly because curricula resources can be obtained more easily from this region and there was a limited time to obtain this research, in this way, the scope could not be global. A future study perhaps later can widen the horizon.

Hopefully, web archiving will appear in more and more contexts in separate curricula frameworks and offer prospects of interdisciplinary collaboration in education and research.

Teaching web archiving in higher education

Program, module, and course descriptions

- UCLA School of Education and Information Studies, Description of the Master Program of Library and Information Science* (n.d.). Retrieved May 27, 2023 from <https://web.archive.org/web/20240313171014/https://seis.ucla.edu/departments-and-degrees/departments-of-information-studies/master-of-library-and-information-science>.
- University of Michigan, School of Information, Web Archiving Course Description* (n.d.). Retrieved May 27, 2023 from <https://web.archive.org/web/20240313171150/www.si.umich.edu/programs/courses/639>.
- University of Toronto Faculty of Information, Digital Preservation and Curation Course Syllabus* (n.d.). Retrieved May 27, 2023 from <https://web.archive.org/web/20240313171413/https://ischool.utoronto.ca/course/digital-preservation-and-curation/>.
- University of Toronto Faculty of Information, Special Topics in Information: Digital Archives and Workflows Course Syllabus* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313171628/https://ischool.utoronto.ca/course/special-topics-in-information-digital-archives-workflows/>.
- University of Toronto Faculty of Information, General Information about Master of Information Program* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313171802/https://ischool.utoronto.ca/current-students/programs-courses/programs-of-study/master-of-information/>.
- University of Glasgow, Information Management and Preservation Master Program Description* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313172203/www.gla.ac.uk/postgraduate/taught/informationmanagementpreservation/>.
- University College London Archives and Records Management Module Description* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313172336/www.ucl.ac.uk/module-catalogue/modules/creation-and-capture-INST0028>.
- King's College London, Master of Arts in Digital Asset and Media Management Program Description* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313172456/www.kcl.ac.uk/study/postgraduate-taught/courses/digital-asset-and-media-management-ma>.
- University of Amsterdam, Master of Information Studies by the Track of Preservation and Presentation of the Moving Image and Sound Program Description* (n.d.). Retrieved January 11, 2024 from: <https://web.archive.org/web/20240313172634/www.uva.nl/shared-content/programmas/en/dual-masters/preservation-and-presentation-of-the-moving-image-heritage-studies/preservation-presentation-of-the-moving-image.html>.
- University of Amsterdam, Master of Information Studies by the Track of Preservation and Presentation of the Moving Image and Sound Description of Specializations* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313172756/www.uva.nl/shared-content/programmas/en/dual-masters/preservation-and-presentation-of-the-moving-image-heritage-studies/study-programme/study-programme.html#Electives>.
- Charles Sturt University, Digital Curation and Preservation Course Description* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313172902/https://handbook.csu.edu.au/subject/2023/INF443>.
- The University of Amsterdam, Media Studies: New Media and Digital Culture Program Description* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313173121/www.uva.nl/shared-content/programmas/en/masters/new-media-and-digital-culture-media-studies/study-programme/study-programme.html>.
- The University of Amsterdam, Research Master's in Media Studies Program Description* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313173239/www.uva.nl/shared-content/programmas/en/research-masters/media-studies-research/study-programme/study-programme.html>.
- Sheffield University, Digital Culture and Communication Master Program Description* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313173350/www.sheffield.ac.uk/postgraduate/taught/courses/2024/digital-culture-and-communication-ma>.
- Description of Dewdrop Software Tool* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313173501/www.dhi.ac.uk/projects/dewdrop/>.

The Routledge Companion to Transnational Web Archive Studies

- Dewdrop Software Tool Source Code and Documentation* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313173623/www.dhi.ac.uk/dewdrop/>.
- Centre for Internet Studies, Faculty of Arts, Aarhus University Profile Description* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313173735/https://cfi.au.dk/about/profile>.
- ELTE University Competencies in Digital Humanities Course Description* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313173845/https://elte-dh.hu/en/competences-in-digital-humanities/>.
- University of Southampton, Social Media and Network Science Course Description* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240314084653/www.southampton.ac.uk/courses/modules/comp6250>.
- Sheffield University, Master of Science in Data Science Program Description* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313174242/www.sheffield.ac.uk/postgraduate/taught/courses/2024/data-science-msc>.
- ELTE University, Data Steward Special Program Description (in Hungarian)* (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313174412/https://elte-dh.hu/adatgazdasz/>.
- What is Digital Library Learning or DILL* (n.d.). Retrieved January 10, 2024 from: <https://web.archive.org/web/20240313174558/www.tlu.ce/meediarav/blogid/what-digital-library-learning-or-dill>.

Notes

- 1 The discipline even has a peer-reviewed scientific journal called *Internet Histories* by Taylor and Francis focusing on the Internet's cultural, social, political, and technological histories of the Internet and associated digital cultures.
- 2 Description of the Master program retrieved May 27, 2023 from <https://web.archive.org/web/20240313171014/https://seis.ucla.edu/departments-and-degrees/departments-of-information-studies/master-of-library-and-information-science>.
- 3 Course description retrieved May 27, 2023 from <https://web.archive.org/web/20240313171150/www.si.umich.edu/programs/courses/639>.
- 4 Course syllabus retrieved May 27, 2023 from <https://web.archive.org/web/20240313171413/https://ischool.utoronto.ca/course/digital-preservation-and-curation>.
- 5 Course syllabus retrieved May 27, 2023 from: <https://web.archive.org/web/20240313171628/https://ischool.utoronto.ca/course/special-topics-in-information-digital-archives-workflows/>.
- 6 Information about the programme retrieved May 27, 2023 from: <https://web.archive.org/web/20240313171802/https://ischool.utoronto.ca/current-students/programs-courses/programs-of-study/master-of-information/>.
- 7 University of Glasgow, Information Management and Preservation Master Program description (n.d.). Retrieved May 27, 2023 from: <https://web.archive.org/web/20240313172203/www.gla.ac.uk/postgraduate/taught/informationmanagementpreservation/>.
- 8 What is Digital Library Learning or DILL? Retrieved January 10, 2024 from <https://web.archive.org/web/20240313174558/www.tlu.ce/meediarav/blogid/what-digital-library-learning-or-dill>.
- 9 Syllabus of the module retrieved May 27, 2023 from <https://web.archive.org/web/20240313172336/www.ucl.ac.uk/module-catalogue/modules/creation-and-capture-INST0028>.
- 10 Syllabus of the programme retrieved May 27, 2023 from <https://web.archive.org/web/20240313172456/www.kcl.ac.uk/study/postgraduate/taught/courses/digital-asset-and-media-management-ma>.
- 11 Program description retrieved in January 11, 2024 from <https://web.archive.org/web/20240313172634/www.uva.nl/shared-content/programmas/en/dual-masters/preservation-and-presentation-of-the-moving-image-heritage-studies/preservation-presentation-of-the-moving-image.html>.
- 12 Program description retrieved in May 27, 2023 from <https://web.archive.org/web/20240313173121/www.uva.nl/shared-content/programmas/en/masters/new-media-and-digital-culture-media-studies/study-programme/study-programme.html>.

Teaching web archiving in higher education

- 13 Program description retrieved in May 27, 2023 from <https://web.archive.org/web/20240313173239/www.uva.nl/shared-content/programmas/en/research-masters/media-studies-research/study-programme/study-programme.html>.
- 14 Course description retrieved in May 27, 2023 from <https://web.archive.org/web/20240314083118/https://studiegids.uva.nl/xmlpages/page/2022-2023-en/search-course/course/100682>.
- 15 Program description retrieved in May 27, 2023 from <https://web.archive.org/web/20240313173350/www.sheffield.ac.uk/postgraduate/taught/courses/2024/digital-culture-and-communication-ma>.
- 16 Program description retrieved in May 27, 2023 from <https://web.archive.org/web/20240314083714/www.sheffield.ac.uk/postgraduate/taught/courses/2024/cultural-data-management-and-communication-ma>.
- 17 Asking Questions for Web Archives — Introductory Notebooks for Historians retrieved in January 10, 2024 from <https://web.archive.org/web/20240314083918/https://netpreserve.org/projects/jupyter-notebooks-for-historians/>.
- 18 Description of Dewdrop retrieved in May 27, 2023 from <https://web.archive.org/web/20240313173501/www.dhi.ac.uk/projects/dewdrop/>.
- 19 Dewdrop code and documentation retrieved in May 27, 2023 from <https://web.archive.org/web/20240313173623/www.dhi.ac.uk/dewdrop/>.
- 20 CFI profile retrieved in May 27, 2023 from <https://web.archive.org/web/20240313173735/https://cfi.au.dk/about/profile>.
- 21 Description of the course retrieved in May 27, 2023 from <https://web.archive.org/web/20240313173845/https://elte-dh.hu/en/competences-in-digital-humanities/>.
- 22 Course description retrieved in May 27, 2023 from <https://web.archive.org/web/20240314084653/www.southampton.ac.uk/courses/modules/comp6250>.
- 23 Program description retrieved in May 27, 2023 from <https://web.archive.org/web/20240313174242/www.sheffield.ac.uk/postgraduate/taught/courses/2024/data-science-msc>.
- 24 Program concept (in Hungarian) retrieved in May 27, 2023 from <https://web.archive.org/web/20240313174412/https://elte-dh.hu/adatgazdasz/>.

References

- Brügger, N. (2009). Website history and the website as an object of study. *New Media & Society*, 11(1–2), 115–132. <https://doi.org/10.1177/1461444808099574>.
- Brügger, N. (2013). Web historiography and Internet Studies: Challenges and perspectives. *New Media & Society*, 15(5), 752–764. <https://doi.org/10.1177/1461444812462852>.
- Brügger, N. (2021). Digital humanities and web archives: Possible new paths for combining datasets. *International Journal of Digital Humanities*, 2, 145–168. <https://web.archive.org/web/20240313181406/https://link.springer.com/article/10.1007/s42803-021-00038-z>.
- Brügger, N., & Finnemann, N. (2013). The web and digital humanities: Theoretical and methodological concerns. *Journal of Broadcasting & Electronic Media*, 57(1), 66–80. <https://web.archive.org/web/20240313175702/www.tandfonline.com/doi/abs/10.1080/08838151.2012.761699>.
- Brügger, N., Schafer, V., & Winters, J. (Eds.). (2020). *Perspectives on Web Archive Studies: Taking Stock, New Ideas, Next Steps*. https://web.archive.org/web/20230924000852/https://cc.au.dk/fileadmin/user_upload/WARCnet/Bru_gger_et_al_Round_doc.pdf.
- Corrado, E. M. (2022). Digital preservation is not just a technology problem. *Technical Services Quarterly*, 39(2), 143–151. <https://web.archive.org/web/20240313181636/www.tandfonline.com/doi/full/10.1080/07317131.2022.2045432>.
- Dogucu, M., & Çetinkaya-Rundel, M., (2021). Web scraping in the statistics and data science curriculum: Challenges and opportunities. *Journal of Statistics and Data Science Education*, 29(1), 112–122.
- Dubin, D., Futrelle, J., Plutchak, J., & Eke, J. (2009). Preserving meaning, not just objects: Semantics and digital preservation. *Library Trends*, 57(3), 595–610. <https://web.archive.org/web/20240313181912/https://muse.jhu.edu/article/265536>.
- Fafalios, P., Holzmann, H., Kasturia, V., & Nejd, W. (2020). Building and querying semantic layers for web archives (extended version). *International Journal on Digital Libraries*, 21(2), 149–167.

- <https://web.archive.org/web/20240313182018/https://link.springer.com/article/10.1007/s00799-018-0251-0>.
- Gambell, S., Gooding, P., Hughes, L., Doran, M., Murphy, O., Tupman, C., & Smithies, J. (2021, July 23). *Communicating the Value and Impact of Digital Humanities in Teaching, Research, and Infrastructure Development*. <https://web.archive.org/web/20240313182245/https://osf.io/z8v9c/>.
- Healy, S., Byrne, H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., & Jansma, R. (2022). *Skills, Tools, and Knowledge Ecologies in Web Archive Research*. WARCnet Papers. https://web.archive.org/web/20230924000852/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Skills_Tools_and_Knowledge_Ecologies.pdf.
- Maemura, E. (2022). *Towards an Infrastructural Description of Archived Web Data*. WARCnet Papers. https://web.archive.org/web/20230924000852/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Maemura_Towards_an_Infrastructural_Description.pdf.
- Maemura, E. (2023). Sorting URLs out: Seeing the web through infrastructural inversion of archival crawling. *Internet Histories*, 7(4), 386–401. <https://web.archive.org/web/20240313182542/www.tandfonline.com/doi/full/10.1080/24701475.2023.2258697>.
- Milligan, M. (2020). *You Shouldn't Need to be a Web Historian to Use Web Archives*. WARCnet Papers. https://web.archive.org/web/20230924000852/https://cc.au.dk/fileadmin/user_upload/WARCnet/Milligan_You_shouldn_t_Need_to_be__2_.pdf.
- Ogden, J., & Maemura, E. (2021). 'Go fish': Conceptualising the challenges of engaging national web archives for digital research. *International Journal of Digital Humanities*, 2, 43–63. <https://web.archive.org/web/20240313182840/https://link.springer.com/article/10.1007/s42803-021-00032-5>.
- Ogden, J., Summers, E., & Walker, S. (2021). Patterns of Use: Conceptualising the role of web archives in online discourse. *Abstract from 4th RESAW (Research Infrastructure for the Study of Archived Web Materials) Conference*, Luxembourg.
- Schafer, V., & Winters, J. (2021). The values of web archives. *International Journal of Digital Humanities*, 2, 129–144. <https://web.archive.org/web/20240313183055/https://link.springer.com/article/10.1007/s42803-021-00037-0>.
- Weber, M. S. (2020). *Web Archives: A Critical Method for the Future of Digital Research*. WARCnet Papers. https://web.archive.org/web/20230924000852/https://cc.au.dk/fileadmin/user_upload/WARCnet/Weber_Web_Archives_A_Critical_Method.pdf.

17

CONVERSATION 3

Communities

*Friedel Geeraert, Márton Németh, Saskia Huc-Hepher,
Matthew Weber, Jane Winters, Valérie Schafer, Friedel Geeraert,
Nadezhda Povroznik, Karin de Wild, Moisés Rockembach,
Susan Aasman, Xiao Ma, and Sophie Gebeil*

Terms such as ‘nation’ and ‘community’ are important for many web archives and researchers studying web archives — should we strive to define these terms in a rigorous way, how could this be done, and who are the ‘we’ to take on this task (if needed)?

- Friedel Geeraert:* It is always useful to have a clear definition of terms, including terms as ‘nation’ and ‘community’. A web archiving Glossary is being developed within the IIPC. A potential collaboration could be set up and these terms could be added to their Glossary.
- Márton Németh:* Defining nations is a complex issue, my favorite conception is from Benedict Anderson (1983). He is talking about modern nations as imagined communities that had evolved from the 18th century. Furthermore, nowadays in Europe within a country, you can find different communities with various sociocultural backgrounds, and it is hard to define common points of national cohesion. I think ‘communities’ can be used better in a web archiving context as it is useful to describe any kind of human community including nations. If someone would like to write about a specific country or nation it just needs to be briefly described in what context s/he is talking about a nation or communities of a nation.
- Saskia Huc-Hepher:* There are, however, troubles with the concept of community too. I agree with Friedel Geeraert that when it comes to defining terms such as ‘nation’ and ‘community’, the ‘we’ taking on the task must be a collaborative, cross-institutional, transnational, and potentially a cross-sector one. The suggestion to add these fundamental terms to the IIPC glossary is excellent, but the collective defining process cannot end there. It is essential that members of the ‘communities’ themselves are consulted and take an active role in the shaping of the terminology.

- Matthew Weber:* Community is a particularly challenging term, and one that I have struggled with in my own research. In defining terms such as community, or even more nuanced, local, one has to think about the multitude of perspectives and possible definitions. For instance, someone living in a rural portion of a given country likely has a very different perception of community than someone living in a dense urban area. In that regard, I greatly appreciate Márton Németh drawing on imagined communities; the notion truly captures the fluidity of communities today. From a practical point of view, there is also a distinction to be drawn between community as a sense of belonging, and community perceived as a proximal notion.
- Jane Winters:* I very much agree with Saskia Huc-Hepher that there are challenges when communities are defined without the involvement of members of those communities (who may, of course, not wish to be included in some kinds of research or defined in particular ways). I am not too sure about the value of closely defining terms like ‘community’ or ‘nation’ in a central glossary, as researchers will most likely define them differently depending on their own positionality and research aims. Perhaps it is more useful to encourage all of us to think carefully about how we use such terms and not assume a shared understanding.

**How important is the role of a mixed community of scholars,
archivists, and data scientists, and how to keep such a community
alive and relevant?**

- Valérie Schafer:* Mixed communities of scholars, web archivists, and data scientists are absolutely essential for conducting ambitious studies on web archives. Over the past decade, thanks to initiatives like the RESAW conferences or those of the IIPC, projects such as WARCnet, and other efforts at national levels, it is clear that a solid community of interest, exchange, and practice has been established. However, this largely depends on the commitment and goodwill of all these stakeholders because project fundings are not well suited to these kinds of hybrid constellations, especially when they are transnational. While the academic and archival fields are inventive in finding collaborative solutions, they are constantly forced to innovate, with the added challenge of sustainability.
- Friedel Geeraert:* I concur with Valérie Schafer; it is primordial that there are mixed communities of scholars, web archivists, and data scientists. The cross-fertilisation that has happened within RESAW and WARCnet has had a direct and positive impact on the establishment of web archiving procedures and policies. It has allowed our institution to take the needs of the research community into account when thinking about the web archive we wish to establish at the national level.
- Nadezhda Povroznik:* Disregarding the scale of a research project, this collaboration is essential and all the parts benefit from these joint activities learning from each other and contributing with their expertise and knowledge. The cycle of the research process requires maintaining these connections and

Conversation 3: Communities

moving back and forth from the scholar with the research question to the archivist (librarian) for the resources, and then to the data analyst to proceed with the data and back to the scholar to take the further decisions and collectively reflect on the processes. In this regard, to foster research of the web archived materials and maintain these collaborations the developed research workflow is essential. It would transparently organise the research process, identify phases and the necessary steps, outline the tasks for the involved parties, suggest the tools and instruments useful at the particular stages, and guide the researcher to advance effectively being a part of the collaboration network. At the same time, this cooperation is hardly possible in those academic cultures where the national programs of web preservation are fragmented or do not exist. In this case, researchers rely mainly on the Internet Archive. However, the research workflow is needed as well but it should be developed in such a way that it is able to cover the expertise and responsibilities of archivists in the research process. It means that the main actors, in such a case, are the researchers and data scientists. Still, their expertise, exchange of ideas, problem-solving, and collaboration are indispensable for the efficient research process. As a result, the unity (or collaboration) of three can be transformed in the interaction of two actors but it means that expertise should be taken by one of the parties or consulted at external resources. Anyway, it is hardly possible to conduct high-quality at-scale research performing all the tasks by a single scholar.

Márton Németh:

I mostly agree, with all of the opinions above. In some cases, it is difficult to find new ways of collaborating with people from different academic backgrounds. However, to realise a successful research project based on archived web data, at least some people have to work together, and find some common ideas, and workflows. The real difficulty is that we do not have rich traditions of how people with completely different academic mindsets can work together in the same framework, for example historians, IT people, librarians, archivists, and communication scientists. Joint efforts, best practices of collaboration, and some evolving traditions can be needed to put this kind of collaboration to a higher level.

Matthew Weber:

Interdisciplinarity is complicated, but the scope and scale of research questions that many of us in this volume aim to address extend well beyond the expertise of any individual domain. I have worked at the intersection of disciplines for the majority of my career, and I believe that this type of work will continue to be an important part of the ongoing work of scholars, archivists and data scientists. Volumes such as this help to foster these conversations and continue the development of this type of community.

**How to involve a critical community of users
(researchers/IT/ data scientists) in curatorial practices?**

Friedel Geeraert:

In countries with young or new web archives at the national level, it is important to do outreach to universities and other research institutions early on as web archiving as a practice will not yet be well-known

and, as a consequence, web archives will be underused as sources for research. GLAM institutions could set up long-term collaborations with researchers and subject experts external to the web archiving institution to work on thematic collections, for example.

Márton Németh: I agree with Friedel Geeraert. In my view, GLAM institutions working in a university environment or a part of a university based on the existing personal connections among IT people, researchers, and GLAM people, can form interesting projects, with challenges that can be inspiring for data scientists and other IT people. They can build up not just thematic collections but run data-oriented projects that can put historians, data scientists, social scientists, librarians, and archivists on the same platform. In the longer-term new competence centres can be created, in this way these kinds of collaborations can get an institutionalised form. The benefits of these projects can be obvious for universities, and also to the wider public.

Karin de Wild: In the Netherlands, the KB, National Library of the Netherlands, does not operate within the university environment. However, there are University Libraries and Digital Humanities Centers that could play a pivotal role in making web archives accessible for researchers and students from various disciplines.

Moisés Rockembach: I believe that the two main challenges we face are outreach and community building around web archives. It is essential to make sure these archives do not turn into forgotten artefacts lost in the digital world, but are recognized by users, including researchers, IT professionals, and data scientists, as rich sources of information. By also engaging the community involved in open science and open data, we can tap into a group that is likely to be receptive to using these research data. By actively promoting these archives and fostering communities engaged in their use and curation, we can elevate their status from mere storage to dynamic resources integral to our understanding of digital memory and culture.

How do you see the relationship between web archives and citizens?

Friedel Geeraert: One of the most important contributions of web archives is that they further citizens' right to information as defined by Article 10 of the European Convention on Human Rights: 'Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.'¹

Nadezhda Povroznik: It is a very nice thing to find these connections and establish them for wider outreach and for enriching the institutions responsible for web archiving with new insights and revealing potential benefits for both parties. I am in doubt of 'citizen science' because it is a more crafting type of activity which requires a lot of effort, skills, and commitment from all involved and without instant fuelling some of these essential components, the relevant projects will not be sustainable. However, curiosity-driven projects can be a potential strength of

Conversation 3: Communities

these relations. In my experience as a teacher, I had an idea to implement using the web archives to inspire non-research audiences and conduct creative projects. It is obvious that web archives keep treasures from the past (from the recent past to the middle of the 1990s). Many of their secrets remain to be discovered and opened up to society. The creative industries require new sources of inspiration, and the web archives are able to bring up the new flow. I have developed an educational course ‘Digital Cultural Heritage: Theory and Practice’ for the project-oriented master’s program ‘Digital Technologies for Creative Industries’ at the Faculty of History and Politics in Perm State University. The course included an extensive section on the use of materials from web archives for the creative industries. As part of the course, students learn the basics of working with web archives, and searching for data. They design the projects on the creative use of the web archives for the development of different projects in the field of digital cultural heritage. For example, some of these projects were devoted to the exploitation of early web images for the development of modern website design, inspiration from the past for the future in company branding, creation of virtual exhibitions of digital art, and some others. Of course, this example is not purely the involvement of ‘citizens’ but already goes beyond ‘traditional’ use of the web archives by researchers.

Márton Németh:

As the previous examples show further efforts must be taken to talk about web archiving with various groups in society. Web archiving can be defined as a special part of personal information management (PIM). In Hungary, the National Széchényi Library started a project some years ago from state funds to make this integration and offer practical guidance for high school students on how to archive their materials and how they can work together on school projects by digital archiving topics. Unfortunately, the termination of the projects by cultural and educational decision-makers made it impossible to get these training materials into schools. However, more and more forms of crowdsourcing and practical involvement have to be realised in my view.

Oral histories and expert interviews are excellent ways of ‘putting people back in the data’. What other ways are there to promote web archive research that sees people instead of data?

Valérie Schafer:

Big data and data-driven analysis indeed tend to obscure singularities, identifying trends rather than considering the individuals behind the data. This is undoubtedly a fruitful approach to handling the vast amount of data, but combining this data with more qualitative studies and close reading can also introduce nuances and reintegrate individuals into the storytelling.

Susan Aasman:

In his blog ‘Big Data for Dead People’²² historian and digital humanities scholar Tim Hitchcock addressed the challenge of the macroscope

that combines distant reading with close reading. No matter how much data or innovative tools are explored, one should never forget the real object of study, which is ‘an emotional, imaginative, and empathetic engagement’ with human subjects. The use of digital hermeneutics as a multi-layered methodology might be an approach that guides such a goal in a more systematic manner.

Saskia Huc-Hepher: To oral histories and expert interviews as ways of ‘putting people back into the data’, I would add ethnography. Ethnography is a slow and intimate approach to research which involves observing and interacting with people in their own environment(s) and their own (trans)language(s). It can take place exclusively online, in the form of ‘netnography’ (Kozinets, 2010), or the web data can be part of a wider ethnographic project, bringing an extra layer of material and, as such, thickness to the offline data. The latter brings living, breathing people back into the digital, bridging the material-digital divide and instead favoring the dynamics of the space in between. Another way to put people back into the archived online data is to take a *multimodal social semiotic* approach to its analysis (Kress, 2010). Multimodal social semiotics not only acknowledges the materiality of data by exploring the multiplicity of modes through which meaning is made, but it necessarily recognises the relationship between the digital material and the social, cultural, interpersonal, extratextual and intertextual context in which it was produced and — crucially — received. This emphasises the infinite meaning-making potential of online data and its reliance on the people interpreting it to (re)construct meaning.

Xiao Ma: Saskia Huc-Hepher’s point about ethnography resonates with me, particularly when the web archive research focuses on a specific group identity. Incorporating archived online data into a broader online and offline ethnographic study would illuminate the complex process of identity-category-making and the contextual contingencies of community-building. Community is built, not found; it is an ongoing process, not a fixed entity. People may participate in community-building to address various contemporary needs and concerns, and this process can manifest in different forms, such as in-person or online events. Conducting ethnographic research would aid in contextualising the archived web data and formulating research questions, and vice versa.

**What are the best ways to address questions of representation/presence/
absence of communities in web archives?**

Saskia Huc-Hepher: Language has to be central. National web archives need to ensure that web artefacts presented in languages not considered representative of the nation are included in their collections. This is particularly — but not exclusively — important in the case of nations whose official primary language is English. Given the contemporary role of English as a global superpower among languages, or what Zeng

Conversation 3: Communities

and Yang (2024, n.p.) refer to as the ‘hegemony of English’, it is essential that web archives overcome technical, practical and competency hurdles in order to provide a lasting platform for, and record of, diverse multilingual and translingual voices. Failure to do so will mean that minority-language groups find themselves excluded from the archive and, in turn, their presence ‘white-washed’ from history. With over 300 languages spoken in London alone, English is by no means the exclusive language of the British population, but there is arguably a cross-sector institutional blind spot when it comes to fully embracing the UK’s multilingualism. If national memory institutions are genuinely to meet their Equality, Diversity and Inclusion (EDI) objectives, they need to open their eyes — and ears — to the multiple languages around them and to the blending of languages in the transnational, translingual online space. This means rethinking English as a unique mouthpiece for Englishness, French for Frenchness, Spanish for Spanishness, and so on. It also means facilitating transnational web archiving collaborations that are plurilingual, interlingual, and potentially translingual. Granted, accommodating multilingualism in technologically mediated settings such as web archives, which are often structured around English can be problematic, but if we are to ensure against cultural absences of the future, it is vital we include the full range of linguistic presences today.

Sophie Gebeil:

I endorse the views expressed by Saskia Huc-Hepher. There is presently a discussion surrounding the digital archiving of non-Latin languages. Concurrently, for the Arabic language, digital mediation has led to a new form of writing using Latin characters: Arabizi, a practice that is widely transnational and visible on the web. The lack of representation of linguistic diversity is further exacerbated by several phenomena that hinder the archiving of minority webs. For example, when I worked on web pages dedicated to the places of Maghrebian immigration history in France, these were low-visibility contents, and I observed that sites with low search engine visibility were often not archived during automatic collection. Thus, low visibility on the living web is coupled with a low probability of collection. One perspective to address these challenges would be to collaborate with web actors from marginalised groups to propose targeted collections that are documented with the involved groups. This approach works in regions where web archiving is accessible. Indeed, the question also arises in states that lack institutional or private collection systems. In such cases, archiving may be established from outside, with the risk of depriving populations of their digital heritage and perpetuating North/South power dynamics. Therefore, this is a transnational endeavour that can only be pursued through cooperation between actors proficient in web archiving and those who are experts in the relevant web spheres, even if they do not have access to archiving facilities.

Friedel Geeraert:

It is clear that language can be an important factor to scope of web archive collections. It is important to keep in mind, however, that many

countries are multilingual and that languages can be shared between countries. Countries like Belgium and Switzerland share languages with their neighboring countries (Belgium: Dutch, French and German, and Switzerland: German, Italian and French). It is important to be able to capture this linguistic diversity in a representative manner, but it is equally important to devise strategies that allow excluding content in the same language coming from another country when working on a web archive with a national scope.

Speaking from a purely organisational point of view, it is also important that the cultural diversity that can be found in society at large is also reflected in the human resources that are working for web archiving institutions. The more culturally diverse the people involved in selecting content to be scoped in, the more representative the corpus will be.

Nadezhda Povroznik: I would expand this question from the communities to the teams and individuals that may be indicated on the websites as contributors or developers but often missing. My research interests are around the history of the museums on the web and the communities, teams and individuals play an essential role in establishing, developing, designing, curating, and maintaining the online environment. What do we know about these communities being at historical distance even within these 25+ years of the web? So often, they remain hidden. How can we identify these teams? It is a really puzzling question but necessary to address. In the case of the museums, we can search for additional resources such as reports on the museum's activities, documents on the development of digital strategies and similar initiatives. At the same time, it is critical to identify these communities and individuals for many reasons to make the research more comprehensive. In the history of the museums on the web there are so many hidden processes under the interface of the website that users explore, there are so many various decisions on layout, technology, content, curation, distribution and other processes made by these people and communities. Recognition of people and their contribution to the development of the museums' web presence would be a right step to be done in shaping a responsible environment and also a solid contribution to further research. Implementing methods from Oral History and conducting interviews with past and present museum staff members can uncover personal insights, challenges and contributions that are not visible in the final digital product. Also, the interviews can be aligned with the practice to discover how the strategies of digital transformation were implemented and more particularly, how the certain initiatives were implemented in the museum's practice.

Conversation 3: Communities

Notes

- 1 European Court of Human Rights & Council of Europe (2021). European Convention on Human Rights. Archived at Wayback Machine April 4, 2024, https://web.archive.org/web/20240404130952/www.echr.coe.int/documents/d/echr/Convention_ENG.
- 2 Tim Hitchcock, “Big Data for Dead People: Digital Readings and the Conundrums of Positivism”, *Historyonics*, 09-12-2013. <https://web.archive.org/web/20240404133109/https://historyonics.blogspot.com/2013/12/big-data-for-dead-people-digital.html>.

References

- Anderson, B. (1983). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso.
- Kozinets, R. V. (2010). *Netnography: Doing Ethnographic Research Online*. Sage.
- Kress, G. (2010). *Multimodality: A Social Semiotic Approach to Contemporary Communication*. Routledge.
- Zeng, J., & Yang, J. (2024). English language hegemony: Retrospect and prospect. *Humanities & Social Sciences Communications*, 11, 317. <https://doi.org/10.1057/s41599-024-02821-z>.

PART IV

Politics of web archives as collections

18

THE TROUBLE WITH COMMUNITY

Constructing, deconstructing and reconstructing transnational “community” micro-archives

Saskia Huc-Hepher and Xiao Ma

Introduction

Despite the ubiquity of the term “community” in British institutional, political and media spheres, the concept lacks critical engagement in the context of transnational web-archiving. In the mid-1990s, social anthropologist Gerd Baumann (1996) observed how “community” was regularly used in reference to “ethnic minorities”, and since, its use has only grown. Designations such as the “Chinese community” or “French community”, “LGBTQ community” or “online communities” are now generally regarded as “culturally sensitive” without further interrogation. In this chapter, we draw on French philosophical Nancian and Derridian thought, alongside contemporary xenofeminist theory, to explore the often overlooked practical and conceptual complexities of transnational “community” web archiving.

The chapter stems from our lived experience of constructing transnational micro-archives (Brügger, 2013), within the overarching UK Web Archive (UKWA), for two ostensibly different diasporic groups: Chinese and French. Currently, diaspora-based special collections, including ours, are categorised under “Society & Communities” in the UKWA. This classification reflects a popularised perception in the UK heritage sector that “ethnic minorities” are “ethnic communities”. The idea of “community heritage” is, however, contested due to the artificial homogenisation and divisions it establishes in society, which underscore “presumed differences between the white, middle class and the ‘rest’” and between “heritage experts and ‘everybody else’” (Waterton & Smith, 2010, p. 5).

It is in this context that we define “transnational ‘community’ micro-archives” as selective collections dedicated to perceived ethnic, cultural and/or linguistic “communities”, who themselves have transnational migration backgrounds and whose online presence challenges national digital borders. But defining and delimiting transnational “community” micro-archives is troublesome. On one hand, their inclusion in the UKWA (hosted by the British Library) pushes the boundaries of the national archive and transforms it into a fundamentally transnational space. That is, by bringing culturally, ethnically and

linguistically diverse voices into the national framework, the diasporic micro-archives serve to destabilise the borders of the predominantly (white) Anglophone .uk digital territory typically represented in the national archive and, in so doing, contribute to its decolonisation (Huc-Hepher, 2021c). On the other hand, labelling the micro-archives as “community” collections erects new borders, thereby creating a sense of a mainstream “us” and a marginal(ised) “them”. These rigid “community” parameters are inconsistent with the ambivalence and fluidity that characterise transnationalism, as transnational identities — and archives — are “not rooted in a singular place, but are forged in and through movement between places” (Ang, 2011, p. 84). The digital traces of diasporic lived experience presented in our two web collections also travel between and across borders, in what we conceptualise as a transnational “in-between space”, envisioned through a metaphorically trans-inclusive lens. It is in this blurry online context, where the Chinese web artefacts rub shoulders with the French and together diversify the national UK archive, that our definition of “transnational ‘community’ web archives” comes into its own and where we explore the trouble with “community”.

The idea of a Chinese or French “community” micro-archive necessarily emanates from an idea of “Chineseness” or “Frenchness”, but how does Chineseness or Frenchness function as a category in the post-imperial, post-print-media, transnational web-archiving context? Is a stable “community” construct compatible with the inherent fluidity of web-curation practice in the transnational in-between space? How does “community” as a term and concept travel across linguistic, cultural and ideological borders? And how can a deconstructive application of xenofeminist principles help us reimagine the web archive along (trans-)inclusive lines that acknowledge the “trans-ness” (Ehrig, Jung, & Schaffer, 2022, p. 14) of the diasporic and the digital? These are some of the questions we grapple with in our discussion of constructing, deconstructing and reconstructing the transnational micro-archive.

Given the paradoxical scarcity of transnational web-archiving scholarship on “community”¹ and its omnipresence as a term in the UK web-preservation context, there is a pressing need to problematise “community” in both its practical and conceptual dimensions.² This chapter will attempt to make inroads into precisely that interrogative space. First, based on our experience of *constructing* transnational micro-archives, we discuss the practical trouble with categorisation along “community” lines. Here, we highlight the dangers of ascribing a homogeneous “ethnic” and/or culturo-linguistic identity to diverse individuals and the postcolonial implications of linguistic choices. In the second section, we *deconstruct* the concept of “community”, drawing on Anderson’s (1983) idea of the *imagined community* and Nancy’s (1986) philosophical exploration of *being-in-common*.³ Here, we focus on the boundaries of “community” and its epistemological reach in a transnational archiving environment. In section three, we turn to contemporary xenofeminist theory (Hester, 2018; Laboria Cuboniks, 2015, 2022) as a means of *reconstructing* diasporic web archiving. Here, using a metaphorically trans-inclusive lens, we propose to inoculate against the binary symptoms of the *mal d’archive* or archive fever (Derrida, 1995), by optimistically foregrounding in-betweenness and constructively deconstructing the archive. Although our collections and discussion of “community” are situated in the UK context, our fundamental arguments on the trouble with oversimplistic identity-based categorisation transcend national borders and, as such, are transferable to other transnational web-archiving settings.

Constructing transnational “community” archives: Process and principles

Notwithstanding the obvious demographic differences between the London French Special Collection (LFSC) and the Chinese in the UK Special Collection (CUKSC) — which later became part of the East and Southeast Asians in the UK Special Collection (ESEAUKSC) — constructing them presented common *troubles*, particularly regarding institutional/curatorial language set against individual/collective identity. Similarly, while our curatorial methods diverged, we had a shared objective to preserve the fragile digital heritage of migrant populations in collections that are both *representative* and *inclusive*.

To meet the objective of *representativeness* when constructing the LFSC, an ethnosemiotic theoretical model was applied. The approach, as previously established (Huc-Hepher, 2015), draws on Bourdieu’s three-stage field analysis paradigm (Bourdieu & Wacquant, 1992) and entails: (1) contextualising the field of study (i.e., the French diaspora in London) within the field of power (e.g., governing bodies, national institutions); (2) ascertaining the relationships and capital flows (economic/cultural/social capital) between different individual/collective players in the field(s) (e.g., commercial, media, associative groups); and (3) scrutinising habitus in relation to field (i.e., individual characteristics and practices). In practical terms, this involved collecting web artifacts corresponding to each of the three diasporic-field scales: macro (e.g., the French Embassy, French Institute and Franco-British Council), meso (e.g., Poule au Pot restaurant, *Ici Londres* magazine, Tamise En Scène theatre group, London Basque Society and London French Rugby Football Club), and micro (e.g., individual artists’ websites, like singer-songwriter Anne B’s and the visual artist Eléonore Pironneau’s, alongside personal blogs, such as Food for Thoughts, Lost in London or Teatime in Wonderland).

Curating the CUKSC and ESEAUKSC was a more organic process. The initial aim was to create a “Chinese in London” micro-archive to record the diverse experiences of Londoners who acknowledge “Chinese” as part of their multiple identities and to form a bridge between their offline and online lives. The strategy was to collect UK websites created by, for or related to “Chinese people”, indicated by having “Chinese” (in English or Chinese) in the website name or “About” page (or equivalent). Three months into the process, however, on a backdrop of COVID-19-related racism, multiple challenges cast doubt over this research and curation rationale. One of the main troubles concerned the creation of a seemingly natural category of “Chinese” in the UK Web Archive and its incompatibility with the complexities of “Chineseness” as an ethnic and/or racialised category in Britain. Yeh argues that “the Chinese” are perceived “not only as an ethnic other, but as a racial other” (2020, n.p.). For obvious reasons, and as violently played out in totalitarian conceptions of community in the 20th century (Nancy, 1986), it is hence highly problematic to apprehend “Chineseness” purely in terms of phenotype, just as it would Frenchness.⁴ Indeed, the racialisation of COVID-19 as a “Chinese” virus (Yeh, 2020) led to a dramatic increase in racial violence against “Chinese” people. Significantly, though, the perpetrators of COVID-19-related hate crime and microaggressions targeted anyone displaying a perceived “Chinese look”, irrespective of the person’s national or ethnic background (Yeh, 2020). As a result, many light-skinned East and Southeast Asians (for example, people of Korean ethnicity) found themselves victims of “anti-Chinese” racism (Yeh, 2020).⁵

Similar, albeit less extreme troubles presented themselves when constructing the LFSC. The field-based principles outlined above produced a corpus whose breadth only partially

met the objective of *representativeness*. That is, the LFSC blogs, though predominantly created by French women, appear to belong to a middle-classed, ostensibly white, subgroup of the wider “community” (thus perpetuating the French diaspora stereotype; Huc-Hepher, 2021a). Equivalent observations apply to many LFSC websites, despite the triadic field-based modus operandi. And while our *inclusion* ambitions were partly addressed through a participatory curation model, it was difficult to involve diaspora members beyond a self-selected subgroup who play an active role in the “community”. Granted, inviting diaspora members to suggest websites for the collections avoided a troublesome top-down curation process. Nevertheless, as these crowdsourced curatorial opportunities were often shared at public outreach events in conventional and/or official “community” spaces (e.g. the French Institute), or online via platforms associated with established organisations and authorities (e.g. China Exchange,⁶ the French Embassy or British Library), the genuine inclusiveness of the co-creative practices is questionable.

Indeed, the act of web *curation* effectively becomes one of *creation*, in the sense of fabrication, when there is a disconnect between the “community” collection and those it claims to represent (Huc-Hepher, 2015). Fieldwork with French Londoners confirmed such a discrepancy, as most did not self-identify as “members” of the “French community”, perceived as an elite based in the affluent South Kensington diplomatic quarter of London (Huc-Hepher & Wells, 2021). Some even expressed a strong sentiment of *exclusion* from the group (Huc-Hepher, 2021a, 2021b).⁷ Validating Smith’s assertion that migrants “are still classed, raced, and gendered” in the diasporic field (2005, p. 238), a French research participant from a working-class background confided how he has “always felt awkward in French establishments [in South Kensington] [...] where people look down on [you] a bit or speak in a slightly rude way; where you sometimes feel inferior”, to the extent that he now systematically seeks to “avoid [...] these cultural and administrative spheres”.⁸ This transfer of pre-existing social structures and class-related insecurities underlines the potential conflict of “community” archiving with individual lived experience.

Likewise, the term “Chinese community” is often ascribed to the entire UK-based Chinese population from diverse backgrounds, either imposed as a hegemonic homogenising label, implying imagined commonality, or strategically self-appropriated to increase visibility and deconstruct dominant discourses from within, recalling Spivak’s “strategic essentialism” (1988, 1996). “Chinese community leaders”, for example, sometimes adopt self-racialised constructions of Chineseness and promote a “shared community” for socioeconomic and political gain (Christiansen, 2003). Although the term “community” is a useful one for ethnographers and memory institutions, empirically it is at odds with many diasporic residents’ perceptions of selfhood, as evidenced by the contested nature of the social construction of a “Chinese community” and the class-based power dynamics experienced in London’s official French institutions. The term “community” was consequently avoided in the naming of both collections to challenge the assumption that “ethnic minorities” are naturally “ethnic communities” and address “the fault lines of ‘community’ thinking along ethnic lines alone” (Meissner & Vertovec, 2015, p. 547).

The notion of “class” brings us to that of “classification” and the trouble with naming/metadata. From its roots in biological compartmentalising and hierarchising, classifying information has been a process of *differentiation* and, thus, implicitly of discriminating and othering. Institutionally, however, it is deemed essential due to the link between descriptive metadata and “value”, for “[o]nly rarely is unlabelled or insufficiently labelled material of any use” (Cain, 2023, n.p.). According to this rationale, our web artifacts need labelling to

The trouble with community

be of “use”, but such naming is not trouble-free (Huc-Hepher, 2015). Assigning titles and organising web objects, and by extension individuals, into ordered categories, reproduces the (b)orders, classes and socially constructed taxonomies (and racialisations) that exist off-line, or worse, introduces ones that were not there at the outset. To avoid relying on abstract systems of class(ification), which reflect the researcher-researched power relationship, Bourdieu foregrounds *practices* as meaningful entry points into understanding social and cultural distinctions (Bourdieu, 1994; Reed-Danahay, 2020). Similarly, Anderson (1983) refers to *style*, materialised through a common *language*, as fundamental to the building of “imagined communities” as nations. This praxis-, style- and language-centred construct is effective theoretically and useful from a curatorial perspective, as it contributes to the three-stage approach outlined above and ensures migrants’ everyday online lives are recorded in the archive (Huc-Hepher, 2015). However, it is more problematic when organising and hence “making sense” and “use” of the archived diasporic web material. For how can troublesome divisions be avoided when classifying information, even according to a logic of practice or (generic, discursive, etc.) style?

In our desire to bring order and discoverability to the large-scale messiness of reborn-digital data (Brügger, 2016; Kitchin, 2014; Ogden & Maemura, 2021), we inevitably cause a tension between *inclusivity* and *navigability*. By *including* them in a community-based category, we *exclude* them from others. How can an inclusive curatorial strategy which rejects the othering associated with migration through negative UK media discourse and a politically “hostile environment” (Mazzara, 2019) be reconciled with an institutional framework grounded in utilitarian classificatory and navigational division? Take, for example, the founding principle of creating and naming a corpus of “French” or “Chinese” community web artifacts. While the preservation intentions are laudable, the implications are manifold. Should common language be the key defining feature of “community” identity, as initially suggested by Anderson (1983)? In such a framework, a London-French corpus would include web artifacts belonging to North and Sub-Saharan African “communities”, as well as those from North American, West Indian or Polynesian webospheres, connected through a shared language and imperial legacy alone (Fanon, 1952).⁹ Should, then, the corpus instead acknowledge the “fluidity of community boundaries and [...] the challenge, and arguably futility, of attempting to assign specific habits and practices to a single heritage or community”, including language (Wells, 2022, p. 57)?

In practical and conceptual terms, the community micro-archive and its naming and cataloguing processes suffer from the tension at the heart of the archive itself. Derived from the Greek word *arkhē*, which simultaneously designates a hospitable home for authored (hi)stories *and* the authority of the memory institution bringing order and sense to material, the archive is, at its very core, troubled by its fundamental ambiguity (Derrida, 1995). Its hospitable function performs an openness to the Other, the *xenos*, as prescribed by xenofeminist principles, particularly in the case of multilingual diasporic collections. Yet, its naming and classifying conventions inevitably instate unwelcome/unwelcoming perimeters around categories, and around and between “communities”. By ordering material into *types*, the risk of introducing stereotypes is never too distant, and so the discriminatory spiral of othering is perpetuated. Likewise, when constructing “community” collections and boxing material into “user-friendly”, navigable compartments, the archive’s authority takes precedence over its hospitality. And this is only accentuated in the case of multilingual collections within the predominantly Anglophone context of the UK Web Archive.

In the case of the CUKSC, we note several terminological/linguistic troubles connected to bringing classified order. According to the 2021 Census, only 0.7% of the overall population self-identified as “Chinese” in England and Wales (Office for National Statistics (ONS), 2022). However, this single umbrella term encompasses a (growing) population who are ethnically, linguistically and culturally diverse, with complex migration trajectories and sociocultural influences from a range of geographical locales, including Hong Kong, Malaysia, Singapore, Vietnam, mainland China, India, and the Caribbean. The complexities of “Chineseness” (Wu, 1991) begin with the English word “Chinese” itself, which not only fails to communicate certain meanings but gives rise to misinterpretation (*ibid.*). A variety of terms are used in Chinese languages to reflect racial, ethnic, cultural and national nuances. In Mandarin, for example, these include *Zhonghua minzu* 中华民族, *Hanren* 汉人, *Tangren* 唐人, *Huaren* 华人, *Huaqiao* 华侨, and *Zhongguoren* 中国人 (Wu, 1991; Lee, 2022). The meaning and use of each term are socially, historically, and politically specific.¹⁰ In the UK, however, all these terms are typically translated as “Chinese”, which flattens the subtleties of people’s identities and complex migration trajectories.

As a direct response to COVID-19-related “anti-Chinese” racial violence (Yeh, 2021), the years following the start of the pandemic witnessed a stark increase in the number of UK websites using the term “East and Southeast Asian (ESEA)” in their names (e.g. Britain’s East and South East Asian Network). This form of self-identification, although not entirely new, was borne of political mobilisation (*ibid.*), online and offline, and as a means of individual and collective healing. Online workshops and events were organised by new ESEA anti-racism and advocacy groups (in which the authors participated). Rather than privileging the “Chinese” experience, the term “East and Southeast Asian” was deployed as a tool to foster solidarity and combat racial inequalities. In a constructively destructive move, therefore, which, as xenofeminist ideals dictate, “seizes alienation as an impetus to generate new worlds” (Laboria Cuboniks, 2015, p. 0x01), a fast-growing ESEA political identity (Yeh, 2021) was challenging the depoliticised “Chinese community” label. Actively engaging in this *collective* category-making in our web-archiving practice therefore became essential to our commitment to anti-racism, particularly in a post-BLM (Black Lives Matter) context. Documenting the rise of this “East and Southeast Asian” identity online also became a new imperative. From August 2021, the curation process therefore switched from being one focused on a self-professed “Chinese” population only, to one geared toward the collection and preservation of websites reflecting the bottom-up, pan-ethnic development of an ESEA (political) entity.

Bringing this section to a close, it is important to stress how these contemporary web-classificatory barriers to inclusion and language-based metadata ramifications echo colonial power imbalances of the past. In the 17th century, indigenous American languages were shoe-horned into the rigid and limited European alphabet of the colonisers for the sake of documentation and preservation (Dodds Pennock, 2023). Today, for the same reasons, overlapping, translingual, transnational and disorderly web data produced by diasporas finds itself ordered into the similarly confining categories and labels of the archiving authority. To agentively avoid reproducing imperial hegemonies of the past, it is therefore necessary to address these practical troubles when constructing “community” collections today. In the next section, we delve deeper into the complexities of “community” curation by deconstructing the concept of “community” itself.

Deconstructing “community” in the transnational web-archiving space

Neal and Walters argue that a community is “an incomplete process through which people construct and create identities, and bond themselves to others, whether geographically, virtually or imaginatively” (2008, p. 237). In other words, a community is built not found, and identities within these communities are shaped by a range of factors, from age, gender, sexuality, and profession, to culture, ethnicity, race, immigration status, socioeconomic status, etc. It is thus highly problematic to assume that “ethnic, cultural or national minorities” are naturally “ethnic, cultural or national communities”. “Minorities” and “communities” are not synonymous. People who recognise migration as part of their personal or family history might participate in community-building, but it is not necessarily related to cultural, national, ethnic, or racial identity. There is no such entity as “the white community in the UK”; neither is there a “Chinese community in the UK” nor a homogenous “French community”. Community is an ongoing process, not a stable entity. It is also, as Anderson (1983) argues, a creative work of the *imagination*, rather than an experiential fact. Just as it is impossible for every member of a nation to know every other member (*ibid.*), so it is impossible for every member of the Chinese or French diaspora in London or the UK to have a tangible connection with each other. Consequently, a Chinese or French “community” is an “imagined community” (Anderson, 1983, p. 65), which is not to say “there is no such thing as community” in a neoliberal, Thatcherian sense.¹¹ Rather, a collective sense of belonging is materialised textually through *communication*, through a common vernacular shared previously via print media (Anderson, 1983) and today chiefly on digital/social media.

This conceptualisation of community as a process constructed through communication, as opposed to an essentialist fact, resonates with the notion of community that French philosopher Jean-Luc Nancy spent a lifetime developing. For Nancy (1986), community does not involve an essential state of being, nor does it imply a split or contradiction between the individual and the collective, reminiscent of neoliberal political agendas, because community is only made possible through the *relationship* between the self and others, referred to as a “relational ontology” (James, 2020, p. 16). One’s existence, rather than being confirmed by Descartes’ dualistic (mind/body) thought-based model (“I think, therefore I am”), is, according to Nancy, realised through *communication* at a more fundamental level: I express myself therefore I am. It is impossible for an individual to state “I am born” or “I am dead”, as both extremes of existence are reliant on other people’s articulation of one’s presence in, and absence from, the world (Nancy, 1986, p. 68). Existence is, therefore, necessarily a common state, what Nancy terms “l’être-en-commun” (1986, p. 99), or being-in-common. The “in” is crucial here, as it places the emphasis on the in-between space, rather than on the idea of commonality in the sense of sameness. So, combining Anderson’s imagined community through materialised communication and Nancy’s notion of expression as the essence of existence, we see how our diaspora-based web collections realise a community presence through the expressive materiality of the web artifacts and how their shared (re)presentation in the transnational web archive constructs a sense of common cohesion. However, this common (re)presentation is distinct from the acknowledgement of a commonality of experience or identity.

Indeed, accepting that existence is conditional on a dynamic between oneself and others does not signify that the self is a common being, for “there is no common being, no common

substance, essence or identity” (Nancy, 1986, p. 224): it is the existence of self *with* others that is the essence of our being, rather than a likeness *to* others that is our essence. In this sense, we form communion through our communication with others and not through a common set of (biological or cultural) characteristics, which, as Nancy is acutely aware, can lead to dangerous totalitarian conceptions of community at both ends of the political spectrum. If communication is essential to “being-in-common”, then sharing a common language, as Anderson asserts but Nancy curiously overlooks in *La communauté désœuvrée* (1986), seems crucial. Nonetheless, as discussed above, regarding both the French and Chinese diasporas, language muddies the waters of the being-in-common concept due to its postcolonial and transnational undercurrents, not to mention many second and third generations not speaking the language of the “homeland”. Does this, in turn, prevent them from “community” belonging?

One trouble with “community”, then, is its insistence on a shared language, together with the tension between its simultaneous conceptual liquidity (as process and multiplicity) and fixity (as stable and universal). Moreover, this liquidity seeps into the conceptualisation and production of the archive itself. As Brügger asserts,

in contrast to the digitisation of a stable newspaper [the medium considered crucial to the construction of the “imagined community” in Anderson’s paradigm], the Web that is to be archived is (potentially) in a state of flux and constant updates and changes.
(2021, p. 219)

Unlike the “print-capitalism [of newspapers, which] gave a new fixity to language [and] created language-of-power” (Anderson, 1983, pp. 44–45), our diasporic web collections provide a counternarrative to the official media representations of migrant minorities and do so in an intrinsically heterogenous, messy, and fragmented fashion. Missing links, multiple versions and intertextual contradictions are steadily becoming the norm rather than the exception. In this sense, our collections in the UK Web Archive provide an archontic home (Derrida, 1995, p. 12), a protective refuge for the diasporic web artifacts, but in a way that reflects the liquidity of the transnational experience and distinguishes itself from the concept of “community” as a hermetic whole. Rather than “community collections”, we are effectively dealing with “[l]iquid homes” (Cohen 2018, p. 241, citing Bauman) and recognising the potential for multiple and changing identities within the singular conception of “community”, as well as the dis-placed nature of the migrant experience and the fragmented reality of web archives.

If we accept, therefore, that a “community collection” transcends a single homogeneous classification of Chineseness or Frenchness and comprises manifold identities, themselves “always plural and in process” (Brah, 2018, p. 238), its fluid in-betweenness is patent. In the transnational context of the London–French experience, the concept of “community” is also trapped between two rigid culturo-ideological frames, leading to a fundamental tension between the opposing understandings of “community” and “*communauté*”, despite their lexical proximity. In Britain, the term “community” typically has positive, inclusive connotations in the public space. Now regularly used in the plural form by contemporary politicians as a substitute for *society*, “community” is defined by the NHS/ONS as “a very general term referring to the people living in a locality or to the locality itself” (Datadictionary, 2023, n.p.). This location-based definition fails to acknowledge the additional semantic reach of the term in English, namely, “people who are considered as a unit because of their common

The trouble with community

interests, social group, or nationality” (Cambridge Dictionary, 2023). In France, while the first dictionary definition of “*communauté*” aligns with the second aspect described above, i.e. a “social group whose members live together, or have common goods or interests” (Le Robert, 2023), there is a key difference between the English and French definitions. In addition to there being an absence of geographical (and national) meaning in the latter, there is a religious aspect not included in the former: “Group of religious people living together” (Le Robert, 2023).

This addition is not innocuous, as it points to the underlying trouble with the idea of community in the French public consciousness. Being a vehemently secular State since 1905, France apprehends communities as being contrary to the Republic’s universal ideals. As a religiously rooted construct, intertwined with sectarian notions of identity, “*communauté*” is consequently seen as a threat to the ultimate overarching “imagined community”, namely the nation-state, or what Laborde refers to as “le communautarisme national” (2016, cited in Kiwan, 2020, p. 4). Further, having been weaponised in French political and media discourses as a critique of both “racialised individuals [...] unwilling to integrate” (Wolfreys, 2018, p. 131) and of the so-called “Islamogauchiste”, “woke” academic elite (Kiwan & Wolfreys, 2023), the notion of “*communauté*” has become highly charged and inextricably linked to “communautarisme” (communitarianism), itself regularly serving as “a metaphor used by ‘respectable racism’” to oppose US/UK pro-diversity ideologies (Wolfreys, 2018, p. 100).

Beyond the religious/secular dissonance in the concept of *communauté*, the difference in the terms’ semantic reach in English and French therefore results from the nations’ diametrically opposed immigration paradigms: a multicultural model versus an assimilationist one. Thus, while ostensibly equivalent terms, the respective positive and negative connotations of “community” and “*communauté*” embody the transcultural and translingual tensions within the Anglo-French transnational space to which the LFSC corresponds. The in-betweenness of the “community” concept is also applicable to the British–Chinese context, with multiple equivalents existing in Chinese languages. Originally referred to as “*jiben shehui* 基本社会” (basic/fundamental society) and “*defang shehui* 地方社会” (local society) in the early 20th century based on translations of the English “community”, by the 1930s, “*shequ* 社区” was coined to blend the earlier terms (Ding, 2020). *Shequ* is still commonly used in China, but its meaning has evolved. While *shequ* tends to focus on locality, other terms describe “community” in different contexts and discourses, including *wangluo shequn* 网络社群 (online community), *ouzhou gongtongti* 欧洲共同体 (European Community) or *guoji shehui* 国际社会 (international community). These examples again highlight the fluidity of “community” across the transnational space-time, but only scratch the surface of its conceptual complexities and elasticity in the Sinitic context. Furthermore, being based solely on Mandarin within China’s borders, they do not reflect variations brought by differing uses and perceptions of “community” in Sinitic-language societies and cultures outside China (e.g., Singapore). In the UK, “Chinese community” translations include *shequ*, *shequn* and *shehui*, substantiating the multifarious interpretations of the “Chinese community” construct and loss of nuance in the Anglophone web preservation setting.

The conceptual tensions and inequivalences discussed in this section illustrate the ideological trouble with “community” in relation to the LFSC and CUKSC, and emphasise the significance of the in-between space in transnational web archiving, a point to which we return in the final section. With these complex community conceptualisations in mind, we need to ask what our so-called “community” web collections are for? Are they principally

repositories for documenting “communities”, or sites for documenting and preserving the multiple and intrinsically fluid processes and interpretations of community-building in the transnational space? As described in the previous section, grassroots groups are actively contributing to Chinese and/or ESEA identity category-making in the digital space. Likewise, although most French field-research participants unequivocally oppose the “community” label (Huc-Hepher, 2021a), having been conditioned by France’s *communitarian* political and media discourses, the “common-unity” of their digital communicative practices (Huc-Hepher & Wells, 2021) testifies to the existence of a collective identity which belies their words. Moreover, through the materiality of the web artifacts and their common semiotic affordances, this collective identity, though plural and liquid, is given lasting solidity within the overarching national archive, rendering them digital fossils of the future, connected through their co-location in the collections. This confirms Nancy’s contention that history-making and community-making are interdependent, with the “history of a single person [...] never becoming historic until it belongs to the community” (1986, p. 253). As researchers and curators of these diverse diasporas, archiving their web artifacts thus serves to cement their collective presence and allows us to map contested processes of category- and identity-making, and better understand negotiations of borders and belongings. We are nevertheless aware that mapping, being built on the colonial “basis of a totalizing classification” (Anderson, 1983, p. 173), brings complex and controversial troubles of its own.

Deconstructing the term “community” in this section has raised important conceptual questions for the broader field of transnational web archiving. In the next and final section, we turn to the future. In the spirit of optimism highlighted by Ogden & Maemura (2021, p. 19), this chapter constitutes “a prompt to explore the space between the ‘future promise’ of web archives and the contemporary challenges they invoke for researcher use”. Having looked back on the diasporic-web-collection construction process and looked inwards to deconstruct the very concept of community, it is now time to look forward and explore potential ways to relieve the “community web archive” of its chronic troubles through active, xenofeminist reconstruction work.

Reconstructing transnational “community” archives: A xenofeminist cure for the trouble?

From Derrida’s original deconstructionist questioning of established oppositions and Nancy’s “singular plural existence [...] [as] problem, ambivalence” (James, 2020, p. 21), through Judith Butler’s *Gender Trouble* (1990), Donna Haraway’s *Staying with Trouble* (2016) and Alison Phipps’s *Me, Not You: The Trouble with Mainstream Feminism* (2020), to this chapter, challenging established binaries and embracing in-betweenness through the prism of *trouble* has constituted a leitmotif in these different currents of philosophical and feminist thought. The focus of this chapter has been on the *trouble* with “community”, as practice, term and concept. Such trouble is only intensified in the transnational web-archiving space, where terms and their underlying (political) ideologies come adrift in a disconcerting in-between space, floating amidst pre- and post-migration cultural frameworks. Equally disorientating is the in-betweenness of the web archive itself, embodying single–collective, online–offline, past–future, public–private binaries in one entity (Huc-Hepher, 2021c). Indeed, playing on the ambivalence of *trouble* in French (the noun meaning *problem/disorder* and the adjective, *blurry*), Derrida, in a Freudian challenge, posits that the archive’s innate ambiguity is its “trouble” (1995, p. 141). And this blurry, generically trans identity (*genre* covering

The trouble with community

discursive/textual genre and gender in French) is seen to be at the root of the archive's conceptual and material malaise (Huc-Hepher, 2021c). In this section, we explore how a xenofeminist (XF) framework, which welcomes deliquescent ambiguity, can be applied to the dualistic troubles of *trans*-national community web archiving.

Although XF writing does not explicitly refer to “trouble”, there are implicit allusions, such as the intertextual nod to Derrida's *Mal d'archive* (1995) in the contention that it is against political “maladies that XF inoculates” (Laboria Cuboniks 2015, p. 0x09). XF is also a self-proclaimed trans-inclusive, technomaterialist “politics for alienation” (*ibid.*), which aligns with the gender-que(e)rying theory of Butler (1990), the post-gender cyborg feminism of Haraway (1990) and the post-#MeToo feminism of Phipps (2020). Phipps challenges “borders of class and nationality [because they] are at one with the borders of gender” and contends that “[b]inary gender is a colonial and capitalist project” (*op. cit.*, p. 156). Being closely linked to cyberfeminism from the outset, the Laboria Cuboniks xenofeminist collective reassert their openness to liquid genders and identities in their 2022 publication (Laboria Cuboniks, 2022), confirming XF's “commitment to *transfeminisim*” and rejection of “*categorical identities* in favour of a feminism based on always *malleable states* and *processes* — *transits* and *transformations*” (n.p.; our italics).

The above quotation shows how XF addresses several of the troubles discussed in this chapter: classificatory trouble when constructing “community” collections; the transits and transformations affecting understandings of “community” in transnational web-archiving spaces; deconstructionist recognition of the relevance and malleability of in-betweenness; and the processual nature of identity and archival construction. In view of XF's pledge to dismantle binaries and celebrate fluid identities, its embracing of contemporary technologies to achieve change and its core tenet to be welcoming of difference (Laboria Cuboniks, 2015, 2022), XF is a theoretical toolkit apt for the reconstruction of diasporic “community” web archiving. We contend that applying XF to transnational web archiving allows for the productive problematisation of “community” and transcends the dichotomies outlined above. XF provides a means to reconstruct web-archiving discourse/practice, avoiding post-colonial and “patri-archaic” notions of categorisation and differentiation, through its “construction of non-oppressive norms and hegemonies [...] so important now as technologies develop at an ever-accelerating rate” (Laboria Cuboniks, 2022, n.p.). To ensure that our transnational micro-archives counter persistent hegemonies, beyond their mere presence in the Anglophone national patri-archive (Derrida, 1995), which alone contributes to a positive translingual and transcultural queering (Huc-Hepher, 2021c), it is essential that web curators actively work against them and are attentive to the “inequalities of gender, race, and class” coded into our digital worlds, without which “these blind spots will continue to determine our future” (Laboria Cuboniks, 2022, n.p.).

When reconstructing “community” web-archiving practice, we therefore need a form of “xeno-hospitality” (Hester, 2018, p. 64) that simultaneously recognises and resists the reproduction of a colonial order through the naming and framing of digital diasporic culture, or what Tony Bennett (2009, p. 102) proposes as the “Museum-Field-Colony” triad. One way to mitigate an implicitly neo-colonial paradigm is to embed diasporic web-archiving in a wider ethnographic project and embrace co-creative, rather than appropriative, methods (Phipps, 2019). This transforms the process from a patronising “armchair” collecting model (Bennett, 2009) to one component of an ethnographic whole, grounded in reflexivity. Furthermore, “insider” researchers-cum-web-curators are equipped with the culturo-linguistic knowhow to cut through semiotic barriers to understanding, which guards against

the warped (re)presentation of collective digital identity and future blind spots highlighted in xenofeminist thought (for what is considered of historic value for the collecting institution may not align with the priorities or lived experience of the diaspora). Importantly, therefore, this reconstructed framework defends “the immanent real of open community as the real itself” (James, 2020, p. 26) — a realness that recognises the messiness of everyday practice and blurriness of “community”, while rejecting imperial logic founded on principles of ownership and division.

Indeed, geared towards the future as well as the past, “archiving produces as much as it records” (Derrida, 1995, p. 34; Appadurai, 2019). Archival inclusions, omissions and classifications are all building blocks for future memory, determining the discourse, filling or introducing gaps in knowledge and understanding. As such, to pave the way for the fullest record and most faithful narrative possible, a reconstructed transnational web-archiving model must explore the breadth and depth of the diasporic webosphere. To minimise present curatorial bias and future skewing of the past, it must actively seek out transnational spaces hidden beneath unexpected top-level domains and ensure that institutional logics of class(ification) do not serve to divide and exclude, but are purposefully fluid, overlapping, potentially unintelligible (for the linguistically lacking), and consistently inclusive of the entities (re)presented. Brah contends that diasporas are “sites of hope and new beginnings. They are contested cultural and political terrains where individual and collective memories collide, reassemble and reconfigure” (2018, p. 237). The same can — or should — be said of diasporic micro-archives. With a moral obligation to the future as well as the past, “community” curators must resist the “[b]iological racialization” carved into collecting institutions’ traditional “territorially and racially defined galleries [and classification systems] (the Black African Hall, the White African Gallery, the European Gallery [...])” (Bennett, 2009, pp. 106–107), or what Haraway terms the “Teddy Bear Patriarchy” (Haraway, 1984, p. 21). We need an archiving practice that transcends these neo-colonial divisions and agentively incorporates the heterogeneity of “the textual traces” of culture (Bennett, 2009, p. 104). A reconstructed transnational archival praxis must revel in the spaces *between* the contours of biology and in the fluidity of identity. It should produce — albeit a cliché — a unity borne of plurality, or in Baker’s (2020, p. 225) construal of Nancy, one that acknowledges “that we are all singular human beings, with a matrix of differences in common”.

In its non-binary attitude to difference, which recalls Derrida’s deconstructive aspirations and “openness to alterity” (Luszczynska, 2012, p. 3), xenofeminism provides a pathway from Bennett’s historic Museum-Field-Colony model to a dynamic Community-Archive-Community one. It is a model where the diasporas co-construct the micro-archive on the basis of their own insights and knowledge, and the archive lays the foundations for a complete picture of the diasporic presence, in all its slimy (VNS Matrix, 1991, cited in Laboria Cuboniks, 2022) and contradictory heterogeneity, for the benefit of future generations.

Let us espouse the xenofeminist force that recognises how we “are thrown into this world of flesh, carved up by gender [and race] and skewed by power” (Laboria Cuboniks, 2022, n.p.) and resist these predestined inequities by reconstructing “community” micro-archiving according to XF’s inclusive and gender/genre-abolitionist principles. The effectiveness of applying such an XF theoretical framework to the LFSC has been previously demonstrated (Huc-Hepher, 2021c), but, as yet, other transnational micro-archives have not been exposed to an XF lens. So, regarding the CUKSC and ESEAUKSC, xenofeminist reconstruction work could involve renaming the collections as “Becoming (British) Chinese in the UK”

The trouble with community

and “ESEA political mobilisation and community-building” respectively. This would highlight the dynamism of group-making and contextual contingencies of community-building. Web artifacts in both collections could be tagged in several collections to acknowledge their multiple belongings and dismantle classificatory borders. When analysing ESEAUKSC web artifacts, researchers could use “issue-mapping”, i.e. identifying and analysing the issues (e.g., hate crime) triggering the term “ESEA” online, to critically examine the “ESEA” construct (Wark et al., 2022). This approach would enable researchers to switch their focus from the identity category to the complex process of identity-category-making.

Embracing the “questioning paradigm” of archival thought, which acknowledges themes of diversity and representation and understands the communicative function of archives (Ridener, 2009, p. 151), we therefore propose a reconstruction of the “community” micro-archive along the following metaphorically trans-inclusive xenofeminist lines. National archiving institutions should support efforts to involve researchers or subject-specialists in transnational selective web archiving. Their culturo-linguistic expertise and sensibilities will help to: (1) empathetically decode the translingual web objects identified and avoid cultural *faux pas*; (2) support inclusive classification methods; and (3) actively engage grassroots (social) networks in the curation process, including groups/individuals not explicitly geared toward “community” work (e.g. religious, sports). Moreover, involving humanities’ researchers in web-curation activities and incorporating web archiving in an overarching (ethnographic) research project has significant symbiotic value. Knowledge of the offline diasporic spaces provides leads for the identification of collectable web artifacts, and searching for web artifacts in the transnational webosphere introduces new offline spaces worthy of attention in the field. True to xenofeminist principles, this dynamic archiving-research process is welcoming of the “entanglements of locations (both on and off-line) that mutually co-constitute one another” (Laboria Cuboniks, 2022, n.p.), and is concurrently immersive and productive. It deepens the researcher’s experiential knowledge of the population and provides a lasting digital record of their physical presence. Beyond the exploratory benefits of the archiving process and consultative stability and framing affordances of the micro-archive as product (Ogden & Maemura, 2021), memory institutions also benefit from the symbiosis. Researcher dissemination, academic repurposing (e.g., in the classroom) and public-engagement activities bring new audiences to, and uses of, web archive(s), as well as bringing new, transdisciplinary theorisation and methodologies.

To challenge the borders that “community” categories erect, collecting institutions should encourage the cross-collection tagging described above. This would perform both Nancean being-*in*-common and inclusive XF objectives. They must endeavour to provide curator tools in multiple languages and facilitate naming/metadata protocols that recognise diasporic fluidity and plurality. Memory institutions should also consider referring to diaspora/ic, as opposed to “community” or even “transnational”, as the notion of transnationalism is based on one of nationalism. Identification with an imagined national space, an “imagined homeland” (Rushdie, 2018) or an “imagined community” is often at odds with individual belongings and powerful regional or translocal selfhoods (Gielis, 2009; Huc-Hepher, 2021a; Huc-Hepher & Lyczba, 2024). In view of this blurry, unstable transness, web curators should move towards a “deterritorialized diaspora” (Cohen, 2018, p. 242; Taylor & Pitman, 2020) that transcends geographical, phenotypical, cultural, and linguistic divides. According to Laboria Cuboniks (2022, n.p.), “[b]iology ceases to be destiny only insofar as we are free to control our biology and refuse to let it merely be a means by which we, ourselves, are controlled by others”. Web curators and archives should, then, defy the

binary hegemonies embedded in conventional preservation practices and embrace a post-biological, post-taxonomic, translanguaging and transcultural order.

Concluding thoughts: Towards trans-unity

Drawing on Nancy's and Derrida's philosophy, alongside Laboria Cuboniks' contemporary xenofeminist thought, in this chapter we have examined the trouble with "community" in three dimensions. First, we addressed constructing community, as web curation/archiving process and principles. Second, we deconstructed community as a concept in the transnational archiving space. Third, we explored how community could be reconstructed as a xenofeminist antidote to the archive's endemic trouble. The chapter has shown that the transnational web archive is a liquid space, where the fluid seeping of identities and blurring of meanings across linguistic, cultural and national borders are only troublesome when attempting to contain them in hermetically bordered "community" categories. If we accept Nancy's contention that community is necessarily diverse and that, like (online) communication, it is "without limits and the being in common communicates itself to the infinity of singularities" (1986, p. 167), we can contribute to a positive rethinking of the diasporic web collection along permeable "open community" lines (James, 2020).

In keeping with both the conception of community developed by Nancy and that of space developed by Massey, the transnational *community* web-archiving *space*, is, therefore, in "the sphere of possibility [...] of coexisting heterogeneity" (Massey, 2005, p. 9). It is an in-between space (hyper)linked through a common-unity of online semiotic and linguistic practices, yet fundamentally plural. This liminal space where the self ends and the other begins, and where nations merge in diasporic lived experience, should not be apprehended as a divisive borderland. Rather, it is a deterritorialised "contact zone" (Pratt, 1991), blurring the boundaries between pre- and post-migration places and between the physical places that migrants inhabit and the digital worlds they create. It is a post-taxonomical space that permeates beyond transnationalism and beyond community to a moist and fertile "transunity".

Notes

- 1 See, however, Ogden and Maemura (2021) for critical analysis of the boundaries of web-based collections and Huc-Hepher and Wells (2021) for work on diasporic web-archives; Taylor and Pitman (2020) on the (in)adequacy of the "classic" nation-state/region in digital contexts; and Diminescu (2012) for a conceptualisation of "e-diaspora" as an online post-migration community (which differs from our blended offline-online understanding).
- 2 Other digital projects bearing the "community" label include the Latin America UK web-archiving project (Cross-Language Dynamics: Reshaping Community) and the 2023 launch of the UK-Ireland Digital Humanities Association's "Community Interest Groups".
- 3 See Luszczynska (2012) and Baker (2020) respectively for rare, combined interrogations of Derrida and Nancy, and Anderson and Nancy.
- 4 There is a popular misconception in the UK that London's French inhabitants are an exclusively white, wealthy elite, meaning the experiences of Black French Londoners and other minorities are overlooked.
- 5 Lighter skin tones are stereotypically associated with essentialist notions of "Chineseness".
- 6 China Exchange is a UK-registered charity raising awareness about China, Chinese culture and London's Chinatown.
- 7 For information on methodologies and participants relating to the authors' broader ethnographic research, see Huc-Hepher (2021a) and Ma (2024).
- 8 All quotations from primary research data and secondary literature recorded or published in French or Chinese have been translated by the authors.

The trouble with community

- 9 At the time of France's colonial empire, significantly known as the *Communauté française*, the French language was imposed to erase cultural/ethnic/racial differences in the name of a collective "French" identity. Such linguistic "white-washing" is evidently at odds with the inclusive aspirations of the LFSC.
- 10 This brief discussion serves only as a starting point to unpack the complex meanings of the term "Chinese".
- 11 Conservative prime minister, Margaret Thatcher, (in)famously declared: "[t]here is no such thing as society" (Thatcher, 1987, p. 30), underlining *individual* rather than *State* economic responsibility.

References

- Anderson, B. (1983). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso.
- Ang, I. (2011). Unsettling the national: Heritage and diaspora. In H. Anheier & Y. R. Isar (Eds.), *Heritage, Memory and Identity* (pp. 82–94). Sage.
- Appadurai, A. (2019). Traumatic exit, identity narratives, and the ethics of hospitality [Electronic version]. *Television & New Media*, 20(6), 558–565.
- Baker, E. (2020). From 'imagined' to 'inoperative' communities: The un-working of national and Latin American identities in contemporary fiction. In C. Davies & R. O'Bryen (Eds.), *Transnational Spanish Studies* (pp. 213–226). Liverpool University Press.
- Baumann, G. (1996). *Contesting Culture: Discourses of Identity in Multi-Ethnic London*. Cambridge University Press.
- Bennett, T. (2009). Museum, field, colony [Electronic version]. *Journal of Cultural Economy*, 2(1–2), 99–116.
- Bernal, V. (2018). Nations, migration, and the world wide web of politics. In K. Stierstorfer & J. Wilson (Eds.), *The Routledge Diaspora Studies Reader* (pp. 257–264). Routledge.
- Bourdieu, P. (1994). *Raisons pratiques: Sur la théorie de l'action*. Editions du Seuil.
- Bourdieu, P. & Wacquant, L. (1992). *Réponses: Pour une anthropologie réflexive*. Editions du Seuil.
- Brah, A. (2018). Cartographies of diaspora. In K. Stierstorfer & J. Wilson (Eds.), *The Routledge Diaspora Studies Reader* (pp. 235–238). Routledge.
- Brügger, N. (2013). Web historiography and Internet Studies: Challenges and perspectives [Electronic version]. *New Media & Society*, 15(5), 752–764.
- Brügger, N. (2016). Digital humanities in the 21st century: Digital materials as a driving force. *Digital Humanities Quarterly*, 10(2), n.p.
- Brügger, N. (2021). The need for research infrastructures for the study of web archives. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 217–224). Springer.
- Butler, J. (1990). *Gender Trouble: Feminism and the Subversion of Identity*. Routledge.
- Cain, A. J. (2023). Taxonomy. *Encyclopaedia Britannica*. Retrieved February 27, 2024, from <https://web.archive.org/web/20240307212609/www.britannica.com/science/taxonomy/Nomenclature>
- Cambridge Dictionary. (2023). Community. Retrieved February 27, 2024, from <https://web.archive.org/web/20240307214020/https://dictionary.cambridge.org/english/community>
- Christiansen, F. (2003). *Chinatown, Europe: An Exploration of Overseas Chinese Identity in the 1990s*. Routledge Curzon.
- Cohen, R. (2018). Solid, ductile and liquid: Changing notions of homeland and home in diaspora studies. In K. Stierstorfer & J. Wilson (Eds.), *The Routledge Diaspora Studies Reader* (pp. 239–243). Routledge.
- Datadictionary. (2023). *NHS Data Model and Dictionary — Community*. Retrieved February 27, 2024, from https://web.archive.org/web/20240307214126/www.datadictionary.nhs.uk/nhs_business_definitions/community.html
- Derrida, J. (1995). *Mal d'archive: Une impression freudienne*. Editions Galilée.
- Diminescu, D. (2012). Introduction: Digital methods for the exploration, analysis and mapping of e-diasporas [Electronic version]. *Social Science Information*, 51(4), 451–458.
- Ding, Y. Z. 丁元竹 (2020). *Zhongwen shequ de youlai yu fazhan jiqi qishi* 中文“社区”的由来与发展及其启示 (The Origin, Development and Inspiration of the Chinese Concept of “Community”). *Minzu yanjiu* 民族研究 (*Ethno-National Studies*), 4, 20–29.

- Dodds Pennock, C. (2023). *On Savage Shores: How Indigenous Americans Discovered Europe*. Alfred A. Knopf.
- Ehrig, S., Jung, B. C., & Schaffer, G. (2022). *Exploring the Transnational Neighbourhood: Perspectives on Community-Building, Identity and Belonging*. Leuven University Press.
- Fanon, F. (1952). *Peau noire, masques blancs*. Editions du Seuil.
- Gielis, R. (2009). A global sense of migrant places: Towards a place perspective in the study of migrant transnationalism. *Global Networks*, 9(2), 271–287.
- Haraway, D. (1984). Teddy Bear Patriarchy: Taxidermy in the Garden of Eden, New York City, 1908–1936 [Electronic version]. *Social Text*, 11, 20–64.
- Haraway, D. (1990). *Simians, Cyborgs and Women: The Reinvention of Nature*. Routledge.
- Haraway, D. (2016). *Staying with the Trouble: Making Kin in the Chthulucene*. Duke University Press.
- Hester, H. (2018). *Xenofeminism*. Polity.
- Huc-Hepher, S. (2015). Big Web data, small focus: An ethnosemiotic approach to culturally themed selective Web archiving [Electronic version]. *Big Data & Society*, 2(2), 1–15.
- Huc-Hepher, S. (2021a). *French London: A Blended Ethnography of a Migrant City*. Manchester University Press.
- Huc-Hepher, S. (2021b). Navigating the London-French transnational space: The losses and gains of language as embodied and embedded symbolic capital [Electronic version]. *Languages*, 6(1), 57.
- Huc-Hepher, S. (2021c). Queering the web archive: A xenofeminist approach to gender, function, language and culture in the London French Special Collection [Electronic version]. *Humanities & Social Sciences Communications*, 8, 298.
- Huc-Hepher, S. & Lyczba, F. (2024) “Sorry, I’m French”: Frenchness as uneasy resource in the construction of home, identity and belonging among French students in London. In C. Wang & T. Lamb (Eds.), *Negotiating Identities, Language and Migration in Global London: Bridging Borders, Creating Spaces* (pp. 52–77). Multilingual Matters.
- Huc-Hepher, S. & Wells, N. (2021). Exploring online diasporas: London’s French and Latin American communities in the UK Web Archive. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web: Exploring Web Archives* (pp. 189–201). Springer.
- James, I. (2020). On open community: Nancy, Laruelle, and the un-determination of the real. In C. Bax & G.-J. van der Heiden (Eds.), *Continental Perspectives on Community: Human Coexistence from Unity to Plurality* (pp. 15–28). Routledge.
- Kitchin, R. (2014). *The Data Revolution*. Sage.
- Kiwan, N. (2020). *Secularism, Islam and Public Intellectuals in Contemporary France*. Manchester University Press.
- Kiwan, N. & Wolfreys, J. (2023) Confronting the politics of denial (Introduction). *Modern and Contemporary France* [Electronic version], 31(2), 139–146.
- Laborde, C. (2016). Elle a perdu son âme. *Libération*, 22 November. Retrieved February 27, 2024, from www.liberation.fr/debats/2016/11/22/elle-a-perdu-son-ame_1530196/
- Laboria Cuboniks. (2015). *Xenofeminism: A Politics for Alienation*. Retrieved February 27, 2024, from <https://web.archive.org/web/20240307215604/https://laboriacuboniks.net/manifesto/xenofeminism-a-politics-for-alienation/>
- Laboria Cuboniks. (2022). *New Vectors from Xenofeminism*. Retrieved February 27, 2024, from <https://web.archive.org/web/20240307215621/https://lab.cccb.org/en/new-vectors-from-xenofeminism/>
- Le Robert. (2023). *Communauté*. Retrieved February 27, 2024, from <https://web.archive.org/web/20240307215647/https://dictionnaire.lerobert.com/definition/communaute>
- Lee, G. (2022). *What Is Chinese (Online)*. Irish Institute for Chinese Studies, University College of Dublin. Thursday 7 April 2022. Retrieved February 27, 2024, from www.youtube.com/watch?v=bzq4X9vAtkU
- Luszczyńska, A. M. (2012). *The Ethics of Community: Nancy, Derrida, Morrison and Menendez*. Bloomsbury.
- Ma, X. (2024). Contesting everyday (food) heritage in London’s Chinatown. In C. Wang & T. Lamb (Eds.), *Negotiating Identities, Language and Migration in Global London: Bridging Borders, Creating Spaces* (pp. 235–262). Multilingual Matters.
- Massey, D. (2005). *For Space*. Sage.

The trouble with community

- Mazzara, F. (2019). *Reframing Migration: Lampedusa, Border Spectacle and the Aesthetics of Subversion*. Peter Lang.
- Meissner, F., & Vertovec, S. (2015). Comparing super-diversity [Electronic version]. *Ethnic and Racial Studies*, 38(4): 541–555.
- Nancy, J.-L. (1986). *La Communauté désœuvrée*. Christian Bourgois Editeur.
- Neal, S., & Walters, S. (2008). Rural be/longing and rural social organisations: Conviviality and community-making in the English countryside. *Sociology*, 42(2), 279–297.
- Office for National Statistics (ONS). (2022). Ethnic group, England and Wales: Census 2021. ONS website, statistical bulletin. Retrieved September 8, 2024, from <https://web.archive.org/web/20240908105512/www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/bulletins/ethnicgroupenglandandwales/census2021>.
- Ogden, J., & Maemura, E. (2021). ‘Go fish’: Conceptualising the challenges of engaging national web archives for digital research [Electronic version]. *International Journal of Digital Humanities*, 1–21.
- Phipps, A. (2019). *Decolonising Multilingualism: Struggles to Decreate*. Multilingual Matters.
- Phipps, A. (2020). *Me, Not You: The Trouble with Mainstream Feminism*. Manchester University Press.
- Pratt, M. (1991). Arts of the Contact Zone [Electronic version]. *Profession*, 33–40.
- Reed-Danahay, D. (2020). *Bourdieu & Social Space: Mobilities, Trajectories, Emplacements*. Berghahn.
- Ridener, J. (2009). *From Polders to Postmodernism: A Concise History of Archival Theory*. Litwin Books.
- Rushdie, S. (2018). Imaginary homelands. In K. Stierstorfer & J. Wilson (Eds.), *The Routledge Diaspora Studies Reader* (pp. 227–229). Routledge.
- Smith, M. P. (2005). Transnational urbanism revisited [Electronic version]. *Journal of Ethnic and Migration Studies*, 31(2), 235–244.
- Spivak, G. (1988). Can the subaltern speak? In L. Grossberg & C. Nelson (Eds.), *Marxism and the Interpretation of Culture* (pp. 66–111). Macmillan.
- Spivak, G. (1996). Subaltern studies: Deconstructing historiography? In D. Landry & G. MacLean (Eds.), *The Spivak Reader* (pp. 203–237). Routledge.
- Taylor, C., & Pitman, T. (2020). Digital culture and post-regional Latin Americanism. In C. Davies & R. O’Brien (Eds.), *Transnational Spanish Studies* (pp. 195–211). Liverpool University Press.
- Thatcher, M. (1987). Interview for *Woman’s Own* (“No Such Thing as Society”). *Margaret Thatcher Foundation: Speeches, Interviews and Other Statements*. Retrieved February 27, 2024, from www.margaretthatcher.org/document/106689
- VNS Matrix (1991). A cyberfeminist manifesto for the 21st century. Retrieved February 27, 2024, from <https://vnsmatrix.net/projects/the-cyberfeminist-manifesto-for-the-21st-century>
- Wark, S., Leung, W.-F., Lury, C., Grey, J., & Yeh, D. (2022). *East and Southeast Asians: Documenting a Category in the Making*. The Public Data Lab. Retrieved February 27, 2024, from <https://web.archive.org/web/20240307215759/https://publicdatalab.org/projects/esea/>
- Waterton, E., & Smith, L. (2010). The recognition and misrecognition of community heritage [Electronic version]. *International Journal of Heritage Studies*, 16(1–2), 4–15.
- Wells, N. (2022). Networking and representing the transnational neighbourhood online: The linguistic landscapes of Latin Americans in London’s Seven Sisters. In S. Ehrig, B. C. Jung, & G. Schaffer (Eds.), *Exploring the Transnational Neighbourhood: Perspectives on Community-Building, Identity and Belonging* (pp. 115–136). Leuven University Press.
- Wolfreys, J. (2018). *Republic of Islamophobia: The Rise of Respectable Racism in France*. C. Hurst & Co.
- Wu, D. Y. (1991). The Construction of Chinese and Non-Chinese Identities [Electronic version]. *Daedalus*, 120(2), 159–179.
- Yeh, D. (2020). Covid-19, anti-Asian racial violence, and the borders of Chineseness [Electronic version]. *British Journal of Chinese Studies*, 10, n.p.
- Yeh, D. (2021). Becoming British East Asian and Southeast Asian [Electronic version]. *British Journal of Chinese Studies*, 11, 53–70.

19

AN INCLUSIVE APPROACH TO WEB ARCHIVING

The case of the Middle East and North African websites in the IIPC Novel Coronavirus collection

Gebeil Sophie

Introduction

This chapter examines the representation of Middle East and North Africa (MENA) countries in the extensive Novel Coronavirus collection developed by the International Internet Preservation Consortium in 2020. By focusing on this empirical case, the chapter aims to delve into the construction of more inclusive web archives at the transnational level and initiate a theoretical discussion on the issue of decolonising the web archive.

The heritagisation of the web is a dynamic driven mainly by Anglo-Saxon and European countries (Musiani et al., 2019), even if China, South Korea, and Malaysia have recently invested in the preservation of their born-digital heritage. The Mediterranean, a Euro-Arab space and the cradle of monotheistic religions, stands as a region marked by clashes of memories and imaginaries, within which heritage commands significant attention from civil societies, states, and transnational entities like UNESCO (Crivello & Dirèche, 2017). This focus, however, sharply contrasts with the nascent state of web archiving in the Mediterranean, particularly in the countries of the MENA. The Arab Springs have highlighted how, in the absence of systematic archiving, the digital heritage of major contemporary events is constructed through diverse forms of choices and interpretations of the past. The consequences become apparent when the digital memory of protestors conflicts with that of states, exemplified by practices such as website erasure by the Syrian regime (Gebeil, 2019). Under these conditions, there is a high risk of losing traces published online during major events such as the Arab Springs or the pandemic, which has particularly affected the Arab countries of the Mediterranean (OECD, 2020). Marked by pre-existing economic and political fragilities, the digital experiences of southern and eastern Mediterranean populations are at risk of becoming invisible, due to lack of systematic collection. In this context, the International Internet Preservation Consortium's participatory data collection has provided an unprecedented opportunity for countries that have virtually no systematic data collection body. As part of the WARCnet project, the IIPC provided access to a spreadsheet containing collected

An inclusive approach to web archiving

URLs and associated metadata. The collection brings together 16,052 web pages covering 137 countries in 51 different languages not limited to IIPC members alone.¹ The collection phase also aimed to include marginalized web spheres, especially non-Western ones.

There are many ways to study this collection, and what interested me were the political implications of the design of the IIPC collection based on the collection itself and the information used in the collecting process (papers, online communication from the IIPC). With this in mind, the MENA web sphere's position in the IIPC collection serves as a case study to contemplate strategies for researchers and archivists, particularly in regions marked by heightened geopolitical tensions where web-archiving initiatives are nearly non-existent. How did the IIPC, a primarily Western organisation, successfully incorporate web content from the MENA region into the Novel Coronavirus collection, and what were the unique aspects of the collected content? From an archival epistemology perspective, this research offers broader insights into navigating complex collecting activities and prompts reflection on the construction of transnational web archive collections without perpetuating Western hegemony. This examination of a minority web presence within an extensive transnational collection sparks considerations for developing more inclusive archiving practices and serves as an empirical case to test the theoretical framework of decolonising archives.

The first section will address the challenges associated with the development conditions of the IIPC collection and the sources used for this study. The second section will demonstrate that the IIPC did indeed have the intention to build a collection dedicated to COVID inclusivity. The third section will specify the role of domain sites from MENA within this collection, and the fourth section will discuss the decolonization of web archives based on this example.

Documenting and questioning the collection development process

In early 2020, when the coronavirus spread in Europe, numerous researchers and institutions already involved in web archiving used their skills to collect online traces related to Coronavirus Disease 2019 (COVID-19) and propose avenues for reflection on the construction and study of the digital heritage of the pandemic, by taking into account different scales and diverse conditions.

The inquiries underpinning this chapter emerged from my participation in these collective transnational projects on web archiving during the pandemic. My involvement was driven by my research interests, which centre around the history of memorial practices (Gebeil, 2021) and anti-racism on the French web, with a focus on the Mediterranean space and a questioning of the management of colonial heritage in contemporary France. Thus, the framework of this article is both transnational and interdisciplinary. The inclusion of items from the MENA in the IIPC collection serves as a rich case study, fostering discussions on three fronts: the study of web archives itself, Mediterranean studies, and the paradigm of decolonising archives.

Firstly, the WARCnet program provided us with the opportunity to access data and engage with archivists on the collection methods during the pandemic. During this experience, I was particularly interested in the international dimension of the IIPC's collection process and its inclusion of sites from countries that are usually under-represented. Nicola Bingham's involvement in the collaborative work played a facilitating role. She is the Lead Curator of Web Archiving at the British Library in the Department of Collections where she manages the Library's web archiving strategy and the ongoing road map for web archiving

capability, ensuring that stakeholders across the Legal Deposit Libraries and other partners have the necessary tools for curating websites according to their own collection development policies. Since 2017, she has been the co-chair, with Alex Thurman (Columbia University Libraries), of the Content Development Working Group (CDG) that spearheaded the Novel Coronavirus collection.² Within the WARCnet project, the Working Group 2 “Analysing Transnational Event” contributed to a better understanding of the construction of the archive as both a trace and a mnemonic lever, but also more broadly as a system, akin to Michel Foucault’s idea that the archaeologist should provide a meticulous description of the archive as a “practice that brings forth a multiplicity of statements, as a myriad of regular events, objects available for processing and manipulation” (1969, p. 178).³ From this perspective, the goal is to comprehend the intentions and conditions of elaboration that guided the identification and selection of materials from the MENA region within the IIPC archive.

Furthermore, studies on the Mediterranean region indicate that historical and heritage issues are the subject of lively debates in a space marked by memory conflicts at different scales, while also being the site of the emergence of minority memories and initiatives aiming to build shared memories that transcend divisions (Gebeil, 2017). The IIPC collecting process thus presents an opportunity to examine how the archive does or does not reflect geopolitical power dynamics in the Mediterranean basin, marked by tension between the increasing individualisation of thought and the assertion of nationalist and fundamentalist currents.

The growing commitment of web archiving institutions to inclusivity aligns with the respect for values of fairness and social justice and has led to notable efforts to diversify collections (Schafer & Winters, 2021). In this perspective, the IIPC has been committed to representing as many countries as possible since its inception and “to collect content from a variety of perspectives when developing its collections.” Despite the end of colonial empires, the Mediterranean region continues to bear the legacy of centuries of European hegemony, with the old continent remaining the main centre of attraction despite geopolitical reconfigurations that have seen the rise of new powers (e.g., China, Russia). On the cultural front, the intervention of transnational organisations, foremost among them UNESCO, to promote the establishment of Mediterranean heritage is accompanied by the dissemination of their own heritage standards and imaginaries, primarily constructed on Western values (Crivello & Dirèche, 2017, p. 16). The born-digital heritage of the Mediterranean remains to be constructed, but the initiative of the IIPC, a Western organization, provides us with a tremendous opportunity to discuss the scope of the decolonization of archives, taking into account the specificities of web archiving.⁴ The concept of decolonizing archives has developed since the 1990s, accompanied by a critical heritage approach that challenges a recognized discourse, where heritage is interpreted through an expert’s perspective that tends to favour prestigious, universal, and grand narratives, thereby neglecting the role of non-dominant groups such as women and ethnic minorities (Berkhofer, 1995). Critical heritage questions the concept of consensual heritage that supports a single interpretation. Plurality of values is crucial for heritage to serve as an expression of social inclusion rather than a means to assimilate and silence different, potentially dissenting voices (Smith, 2006). This critical perspective aligns with the notion of “decolonisation of knowledge,” referring to the historical process of challenging the superiority of European and Western epistemology (Mbembe, 2013, p. 15).⁵ This movement is both a theoretical and practical endeavour, aiming to build a world culture, “a critical humanism founded primarily on the sharing of what differentiates us, beyond absolutes” (Mbembe, 2013, p. 42). In this context, an increasing number of public bodies at local, national, and international levels are being prompted to explore cultural heritage as

An inclusive approach to web archiving

an arena for pursuing social innovation and inclusion objectives (Innocenti, 2016). This has manifested in archive restitutions, the development of practices attentive to the visibility of minorities, and the critical mediation of archives involving community participation. Several manuals geared towards GLAM (Galleries, Libraries, Archives, and Museums) have emerged in this direction, guided by principles of equity, social justice, and the affirmation of diversity. At the same time, voices have risen to emphasize that decolonization should not only be a matter of communication, but should also encompass all means of knowledge production, moving “towards a decolonial archival praxis” (Ghaddar & Caswell, 2019). During a period of global crisis, such as that of 2020, coupled with the governance constraints of web archiving in the Mediterranean, it is essential to study the context of the IIPC’s collecting production, specifically intentions with regard to representativity, and subsequently observe the outcomes related to MENA.

To address these questions, it is essential to comprehend the collection itself while documenting the history of the collection’s creation using various sources.

To achieve this objective, our sources of study were of two types: first, information that provided insights into the very conditions of the collecting process (testimonies, IIPC communications, etc.), and second, the collection itself.

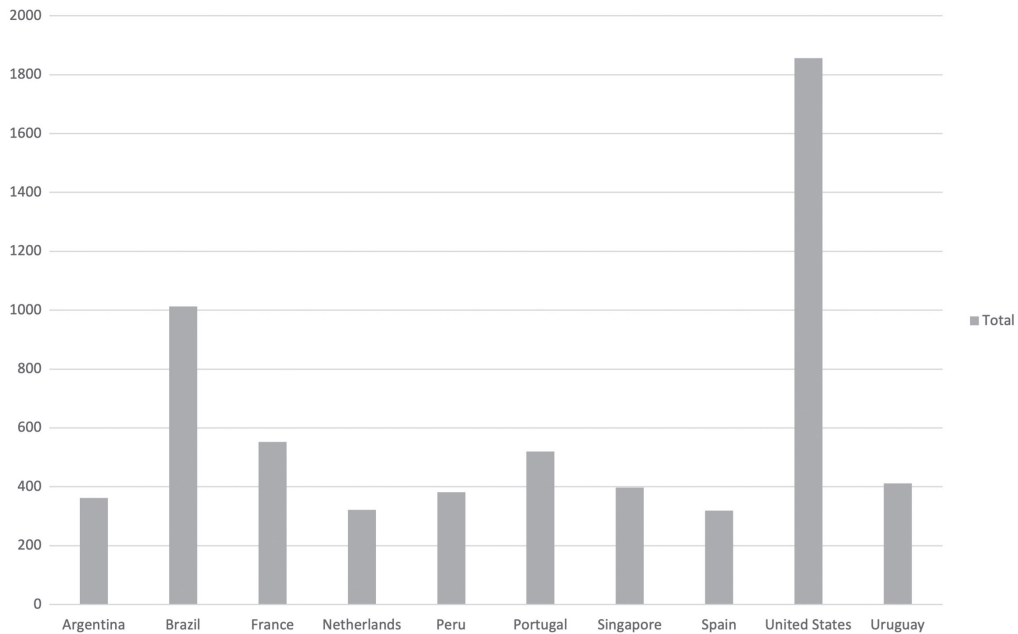
The first type of material included papers and web pages describing the collection process, with a particular focus on the collaborative efforts during the WARCnet program involving Nicola Bingham, who served as co-director of the Novel Coronavirus collection. An interview with her conducted by Friedel Geeraert, published in a WARCnet paper (Geeraert & Bingham, 2020), proved especially informative. The interview sheds light on the circumstances surrounding the collection’s creation, providing precise details about its inception, objectives, achievements, and challenges (see also Bingham et al., 2024). It highlights how the IIPC Content Development Group leveraged past experiences in collecting transnational events to establish an exceptional collection system in collaboration with the Internet Archive (Geeraert & Bingham, 2020). This collection, now accessible on Archive-It, consists of an impressive 21,600,104 documents and files. Additionally, the Internet Archive foundation has increased the storage allocation for the IIPC from 3 TB to 5 TB.⁶

The second type of material pertained to the ongoing collecting process, allowing us to examine the data as of September 2020 and consider the main trends and characteristics of this archive under construction. As part of the WARCnet program, Nicola Bingham shared a metadata spreadsheet with members, containing 10,734 entries. For each entry corresponding to a targeted web page URL, the spreadsheet provided the following information: Page title, Actor categories, Theme, Curator, Selection date, Frequency, Supplementary archiving, Keyword, Supplementary Information, URL history, Language, Provenance. It covered 137 countries in 51 languages. This information facilitated the creation of initial filters, further complemented by statistical analysis and pivot tables within the spreadsheet.

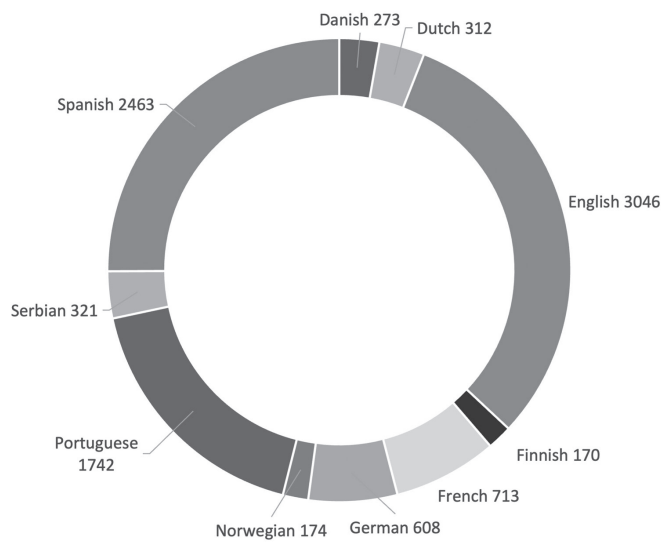
A critical reading of the documents of these two types of sources enables us to describe the collecting context regarding the representation of MENA countries.

An initiative to build a more inclusive web archive

In the Mediterranean region, web archiving remains limited, even on the European shores, where only France, Spain, Greece, Croatia, and Slovenia have systematic web archiving, unlike Italy, for example. Regarding MENA, Egypt is the sole country with a collection system supported by the Alexandria Library in partnership with the Internet Archive. Apart



(a)



(b)

Figure 19.1 (a & b) The 10 most represented countries and languages in the Novel Coronavirus collection (colour version in Zenodo repository).

An inclusive approach to web archiving

from Israel, which has its own preservation project without being an IIPC member, Egypt, France, Spain, Greece, Croatia, and Slovenia are all involved in the consortium. Despite a gradual expansion of the network since its inception in 2006, MENA and the Mediterranean regions remain underrepresented in the IIPC.

Universities and GLAMs (Galleries, Libraries, Archives, and Museums) in the Arab world play a leading role in developing a reflection on digital archiving with the rise of digital humanities. However, these initiatives, focusing on the digitisation and coding of the Arabic alphabet, are less publicized than Western projects conducted in the region. Political instability, the various crises, and armed conflicts in recent years have undermined the preservation of cultural heritage, whether through the deterioration of archives or, in more extreme cases, deliberate destruction: “The Arab world is facing a real challenge: never before has its heritage been exposed to so many different threats” (Bayoumi & Oliveau, 2020, p.16). Under these conditions, the IIPC collection is an opportunity to preserve the traces of countries that are largely excluded from transnational web archiving initiatives.

For the IIPC Content Development Group, when the pandemic began to emerge in early 2020, several members of the IIPC expressed their interest in having the group engage in a collection of this transnational event. The thematic approach was quite open, encompassing strictly medical aspects as well as economic, political, and social facets. Preference was given to “information resources rather than social media feeds or hashtags” (Geeraert & Bingham, 2020, p. 7).

The harvesting of content was based on a dual dynamic. On the one hand, 30 IIPC members received a collection guide focusing on specific topics (origin of the coronavirus, information on symptoms, lockdowns, restrictions, medical, social, economic, and political aspects) to propose relevant URLs identified online. Suggestions were listed in a collaborative spreadsheet where members filled in basic metadata about the selected URLs: title, language, description, domain name extension, country, harvester name, collection perimeter (entire host site or only one page). On the other hand, the IIPC distributed a questionnaire in English on Google Form, open to the public, allowing any internet user or institution not a member of the IIPC to share suggestions by filling in the information cited above for each proposed URL. The process was moderated manually by Nicola Bingham and Alex Thurman (Geeraert & Bingham, 2020). Due to the process as well as storage limitations, the harvesting was not automated, unlike the large collections undertaken by the French National Library (BnF) or the National Audiovisual Institute (INA) as part of the archiving of the national web.

The contributive nature of the process aligns with the aim of broadening the international dimension as much as possible. As explained by Nicola Bingham, moderation required reaching out to some very dynamic and generous archival institutions, asking them to reduce the number of suggestions to provide room for spaces that are typically less represented: “we wanted to represent all regions and all countries” as “China, Russia, Africa, for example, are underrepresented in web archiving” (2021, p.6). This focus aligns with the CDG principle, as stated on its web page, which specifies its commitment “to include relevant websites from many countries and in many languages when developing its collections.” This attention to representativeness reflects the increasing commitment of CDG leaders, like many archivists, to values transmitted through the constitution of archived web collections, as emphasized by Schafer and Winters (2021). Unlike content subject to legal deposit regulations, the collected material in this case is accessible online. This circumstance has compelled managers

to prioritize the quality of information to prevent the propagation of rumours, posing a dilemma as noted by Nicola Bingham. While the group aims to be attentive to a plurality of opinions, the online dissemination of the collection, not solely limited to researchers, has led the team to exclude certain content related to anti-vaccination sentiments or COVID denial (Halboub et al., 2021).

MENA and the Mediterranean in IIPC collection, a discreet presence

The countries represented in the IIPC collection are mainly from Europe (Denmark, France, Norway, Netherlands, Spain, Portugal, UK) and the Americas (United States, Brazil, Peru, Uruguay), with the United States accounting for almost one fifth of the content (Figure 19.1a). This corresponds in part to IIPC member institutions that are particularly active in the network, but there are also non-IIPC members among the 10 most represented countries, such as Brazil and Peru. This European and American overrepresentation is reflected in the languages referenced regardless of the country of publication: 31% of the collection is in English, 25% in Spanish, 18% in Portuguese, and 7% in French (Figure 19.1b).

Within the IIPC collection, the first observation that stands out is the low representation of Mediterranean and MENA countries: 1,733 web pages are identified, representing about 17% of the collection. However, of these 1,733 archived web pages, 97% were published by countries on European shores (1,672) against less than 3% for MENA countries (61), including those not directly bordering the Mediterranean. The European overrepresentation is glaring in the collection (see also Chapter 10 by Geeraert et al. in this volume, Table x.1, p. XX).

How do the different countries' archiving conditions and their membership or not of the IIPC affect these disparities within the collection? Unsurprisingly, the most highly represented countries have archiving institutions and are members of the IIPC, such as Spain, with 319 URLs collected compared to 46 for Italy. However, only three countries that collected URLs are from the Egyptian web even though the Library of Alexandria is the only web archiving institution within the MENA zone and is also a member of the IIPC.

Among the web pages collected from country domains in the MENA zone, the disparities are quite small, ranging from 1 to 7 per country. The United Arab Emirates, Qatar, and Jordan present more than six URLs compared to two or less for Tunisia, Saudi Arabia, Oman, Libya, Iraq, and Algeria. In addition to the country of publication, it is also possible to search for content collected in Arabic. This corresponds to 31 URLs (the collection had 41 in February 2022), or less than 0.3% of the collection, all published on MENA web domains, except 2 from the United States and South Korea (Figure 19.2).

When cross-referencing the country of publication with the language used, we can see that there is a higher representation of languages based on Latin-script alphabets: out of 61 pages collected from MENA web domains, only 27 are in Arabic, against 23 in English, 6 in French, 2 in Hungarian, and 3 in Portuguese. This raises many questions about collection processing. Nicola Bingham explains how language issues posed a significant challenge during the collecting, which ultimately encompasses 136 different languages. It was particularly complex in terms of moderating the selection of URLs submitted for preservation through the online questionnaire. Pages in English could be processed directly by the team. However, Nicola Bingham mentions reaching out to colleagues from member institutions to inquire whether a page indeed dealt with COVID-19 when no one on the team was proficient in the language in question. However, due to the pandemic, the BA (Bibliotheca

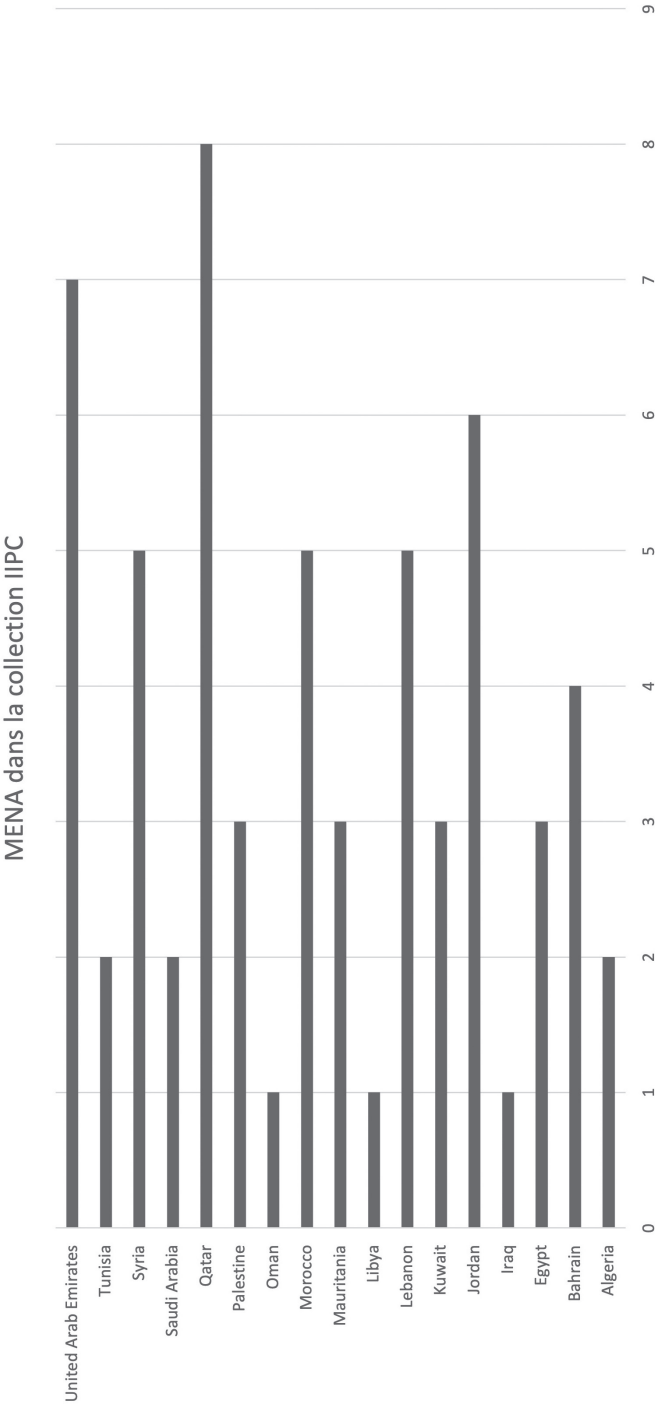


Figure 19.2 Distribution of archived Arabic-language content by country domains.

Alexandrina) could not participate in the IIPC collection process and thus could not assist the team in moderating the content. Consequently, Nicola Bingham implemented a strategy based on automated translation and manual examination of pages for clues:

Particularly in the case of websites with non-Roman scripts that I couldn't interpret, it was challenging. In some cases, dropping a URL into a browser and having a look will tell you if the content is relevant because you can pick up visual clues. [...]

In other cases, I was just a little bit lost. We did make use of Google Translate. So we would drop particular pieces of text into their tool to ensure that a website is relevant. (Bingham in Geeraert & Bingham, 2020, p. 5)

As a result, certain web pages could not be selected because their content was not fully understood, despite the efforts deployed within the CDG. This highlights how the absence of Arabic-speaking personnel on the teams involved in the IIPC collection, both within the CDG and member institutions, was a substantial obstacle to the representation of MENA web domains in the collection. The pandemic context had a notable impact in this regard. While already grappling with challenges related to web archiving, MENA institutions were preoccupied with crisis management during the pandemic, leading to limited availability for focusing on archiving efforts (Chebbi, 2022). This may be why the Bibliotheca Alexandrina was unable to respond favourably to the IIPC's invitation to participate in the project.

In addition to this low linguistic and national presence, a second observation lies in the specific nature of the pages collected from MENA countries, particularly regarding the category of actors represented. The comparison shows an over-representation of institutional and governmental actors in the few MENA web pages. Within the collection as a whole, 35% of pages belong to national (20%) or local (15%) governmental actors, compared to 64% for the MENA sites. Conversely, the total collection offers 43% of media content compared to 28% for MENA sites (Figure 19.3a and 19.3b). This implies that there are more official discourses, including propaganda, than traces of civil society initiatives in often non-democratic countries. The archived pages originate mainly from ministry of health websites, along with dedicated platforms specifically created for the pandemic. These sites are sometimes hosted on government platforms and are designed to inform the population about the mechanisms and symptoms of the virus, preventative measures, and ongoing immunisation campaigns.⁷

Like IIPC, other organisations have compiled international collections dedicated to the coronavirus through Archive-It, focused on more specific topics or areas. By looking at the scope of these different collections, we can identify how archiving policy impacts the nature of the collected content. As of March 15, 2022, Archive-It has 275 collections with the descriptor "COVID" mainly deposited by research institutions or American and Canadian GLAMs, such as the small collection from the University of Massachusetts.⁸ Notably, the Athens Regional Library (*Athens, Georgia Area COVID-19 Response*) has made a set of 134 pages on the pandemic available on the Greek and the Georgian web. Above all, two major initiatives from the United States, led by the National Library of Medicine (NLM) and the University of Illinois, respectively, offer some pages and content from the MENA region. The NLM's Global Health Events web archive was launched in 2014 to document the Ebola outbreak. Regarding the coronavirus, it contains websites and social media from government organisations, NGOs, journalists, health workers, and scientists, including thirty from national websites in the MENA region and seven videos explaining preventative measures

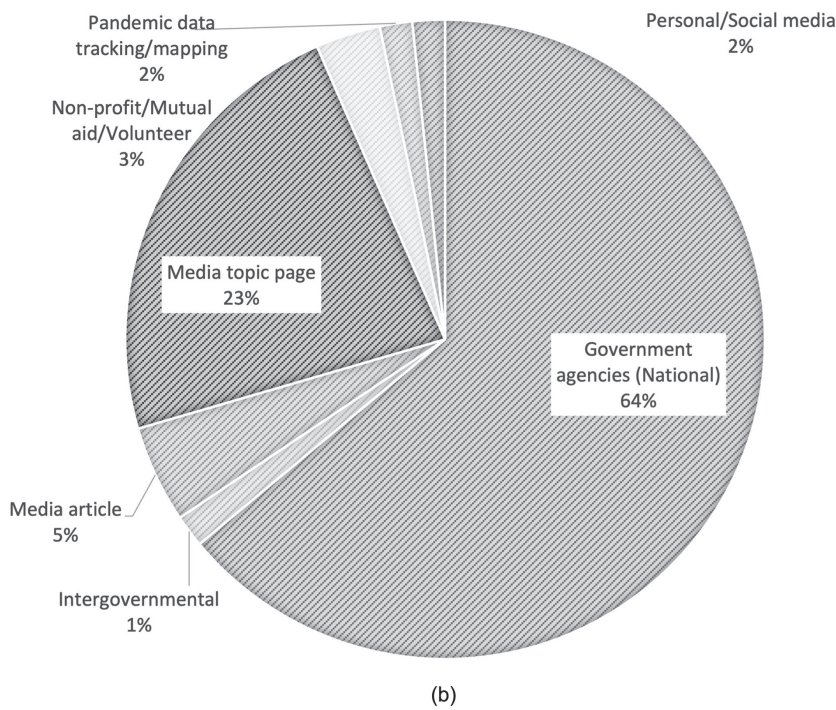
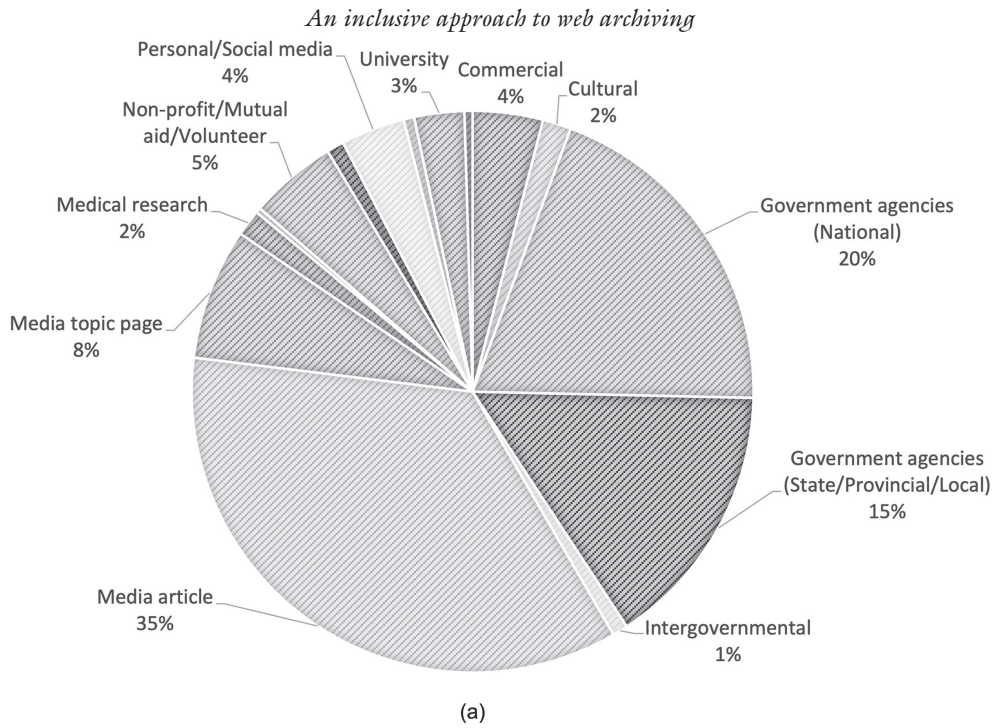


Figure 19.3 (a & b) Comparison of actor categories between the entire IIPC collection (top) and the sites collected from the MENA zone (bottom) (colour version in Zenodo repository).

in Arabic.⁹ At the University of Illinois, the Global Civil Society Organisation Responses to COVID-19 collection includes 29 sites and subsites, presented as a representative sample of responses to the outbreak from NGOs around the world. More than half of the collection targets the MENA area, such as the *Arab Council for Childhood and Development* or the *Palestinian Center for Human Rights*.¹⁰

To finish this overview of Archive-It's collections, also worthy of note are the hundred URLs listed in the *Muftiships Web archive* collection that covers the responses of the main Muslim religious authorities to the coronavirus and illustrates "how Islamic law is administered in the digital age" in the Middle East as elsewhere. Half of the content is in Arabic and a large part is related to the MENA region from an exclusively Islamic perspective. In addition to the U.S.-based foundation, national web archive collections also contain pages from foreign countries, picked up by crawlers (Aasman, Bingham, Brügger, et al., 2021). By consulting the list of COVID-19 collections provided by WARCnet project partners from France, Denmark, Luxembourg, Hungary, the Netherlands, and the United Kingdom, around 50 web pages published from the MENA region have been archived. These pages include content from press agencies and media outlets such as the Tunisian Africa Press Agency (www.tap.info.tn), archived by the National Library of Luxembourg, and the Palestinian online media outlet Metronews (www.metronews.ps), preserved by the Danish web archives.¹¹

Despite the marginal presence of websites from the MENA region in the IIPC collection, the collected URLs still provide visibility to spaces typically absent from web archiving outside of specific collections. The traces collected open new possibilities, shedding light on the history of institutional and vernacular digital practices within Mediterranean societies and the MENA zone (Gebeil, 2022). From a theoretical standpoint, it also raises questions about the power dynamics that inevitably persist within current collections of archived web content, despite efforts to embrace inclusive values within archival institutions (Schafer & Winters, 2021, p. 140).

Decolonising web archives?

Concerning the MENA region, the increased use of digital technologies during the health crisis has not led to a "pandemic effect" accentuating the implementation of web archiving. Marginalized within this transnational dynamic, spatial under-representation is coupled with the marginalisation of vernacular content. Online traces from civil society remain invisible, overshadowed by political communication. Mere fragments of the MENA region emerge in the IIPC collection, which is nevertheless one of the most comprehensive in terms of representing different continents. Several explanations exist for this situation. Firstly, internal factors within the MENA countries, such as political instability, lack of financial resources, and internet censorship practices, hinder the establishment of an archiving policy that is not considered a priority. Secondly, this absence can be attributed to the governance of the IIPC network, which currently has limited reach in the MENA zone. Moreover, these web archive collections serve as valuable sources for historians of the future. Therefore, documenting the archiving process is crucial; otherwise, the collections may not be highly exploitable. Whether the focus is on institutional or media sites or a preference for content in the Latin alphabet, the information documenting the social, cultural, and technical context of archiving imbues the collection with meaning. At IIPC, as in any institution grappling with digital hyper-memory, this "socio-cultural construct" known as archiving results from collective

An inclusive approach to web archiving

compromises guided by the pursuit of the best possible representativeness and considerations of financial or organisational constraints. The acknowledgment of persisting representational biases represents a significant commitment on the part of archivists and scholars.

Because of web archiving, a significant portion of the MENA digital heritage is preserved by the Internet Archive foundation. Paradoxically, several countries in the region are officially in conflict with the United States. This raises ethical questions specific to the preservation of a transnational collection, whose safeguarding reflects the persistence of power dynamics inherited from the present and still relevant, whether through preservation by former Mediterranean colonial powers or through the imperialism of foreign powers such as the United States. In the absence of preservation mechanisms in MENA countries, this native digital heritage saved abroad may someday become the subject of restitution requests from societies in the countries concerned. Similarly, the selection of pages is based on values and practices specific to collecting institutions, which may differ from those of MENA region countries. Due to the lack of collaboration in the pandemic context, this dialogue could not be initiated, but it remains a point worth pursuing in a transnational perspective. This prompts consideration that a transnational collection is not merely an assemblage of sites from different countries, which essentially describes an international collection. The transnational dimension embraced by the IIPC goes beyond that. It offers a rare space for cooperation to envision collaborative work among different countries, considering reciprocal exchanges not only of archived materials but also of expertise, cultures, knowledge, and individuals. It is indeed this dialogue underpinned by the movement of people, including from the South to the North, that carries within it this essential transnational purpose in the face of major events like the COVID-19 pandemic.

In this perspective, MENA web provides an opportunity to develop a more representative and inclusive collection process, offering a platform for examining the concept of decolonising archives. According to Genovese (2016, p. 33), referencing Linda Tuhiwai Smith, “the basic principles of research (and by extension, archival methodology) are inexorably influenced by colonialism and imperialism.” Initially, inquiries by professionals and scholars focused on the archives of former colonised populations, leading to movements for restitution and knowledge-building projects that restored the rights of these populations and promoted empowerment. Initiatives like *The Protocols for Native American Archival Materials* exemplified this approach, contributing to the return of documents to indigenous communities. However, the perspective of the “archival decolonist” (Genovese, 2016, p. 33) has expanded to digital realms as well. The 2022 conference of the Association of Internet Research focused on “practices and processes for decolonising the internet and internet research,”¹² addressing the necessity of breaking away from Western hegemonic narratives to make space for other voices and examining the effects of capitalism through the colonisation of web giants, related to a “colonialism of power” (Schoon et al., 2020, p. 8). In research, this movement also applies to methodologies for studying digital practices and media studies. Regarding research methodologies related to information practices in Africa, Schoon et al. encourage us to “cease to see methodological approaches and concepts developed in the Global South as having a lower ontological density in the hierarchical disciplinary base of knowledge” (2020, p. 2). This stance is stimulating because it urges researchers and professionals to embrace the political dimension of archiving, as Genovese reminds us by quoting Derrida: “there is no political power without control of the archive” (1996, p. 4). Building on these reflections, the examination of a blind spot in the remarkable collection of the IIPC allows us to consider recommendations for enhancing inclusivity and representativeness in transnational web

archiving. While there is no definitive guide, given the social and professional divisiveness of this issue (Janssen, 2023), we can apply the actual reflection focused on decolonisation of the archive to suggest several avenues related to web archives. How can transnational web collections give space to marginalized voices on a global scale?

1. Recognise that collections may reflect hegemonic biases, risking the overrepresentation of wealthier and more advanced countries in web archiving, as they are better connected within international, and predominantly Western, networks. This will require training project teams on these complex issues.
2. Include professionals and researchers from Southern countries in transnational collection projects, which naturally raises questions about dedicated funding.
3. Consider multilingualism by collaborating with professionals, students, and researchers from the Global South or interpreters.
4. Implement a co-constructed collection methodology that leverages the expertise of Southern partners for optimal representativeness. This requires collaboration with institutions, archivists, and even activists to capture a diverse digital expression of civil society, drawing inspiration from projects like the collection of online traces of the Tunisian revolution (Zghidi, 2021). The participative dimension developed by the IIPC also constitutes a basic model that warrants following.
5. Design collaborative interfaces (formats, multilingualism, access conditions) to bridge the digital divide.
6. Develop storage solutions suitable for all partners, ideally with dual recording on servers located on different continents.

Of course, the successful implementation of these measures requires strong political will from funders and leaders. Building a more inclusive transnational web archive also demands sufficient time to forge cooperative relationships, aligning more with funding models like the European Research Council (ERC) rather than national project budgets, especially in the European context.

Conclusion

Despite the different content production and the transformation of documents into archives, “[t]he archive has an organic relationship to what it is an archive of: caused by the event whose memory it records, the archive is the sign, by its very existence, of the event and the fact that it did indeed take place¹³” (Bachimont, 2021). From this perspective, with its fundamental mission being the preservation and transmission of born-digital traces created by human societies, the web archive serves as a “mnemophore,” a custodian, and a bearer of collective memory (Bachimont, 2021). However, the IIPC Novel Coronavirus collection represents only a partial facet of the event’s memory, focusing on the involvement of predominantly Western institutions in an emergency initiative, a pattern observed in several GLAMs during the pandemic. Consequently, the case of data archived from Mediterranean and Arabic web sources within the IIPC collection underscores the enduring challenges faced by organisations committed to the preservation of web content and digital social networks.

Web archiving is also an integral component of heritagisation, contributing to the individuation of a collective identity (Simondon, 2005). Thus, the collections created are tailored for diverse audiences through different types of mediation. Concerning the Novel

An inclusive approach to web archiving

Coronavirus project, the IIPC chose to publish the entirety on the Internet Archive website, which determined the criteria for the collection (e.g., themes and information quality). Providing metadata that includes the date, the URL collected, and various information about the web page facilitates interpretation of these objects that carry memory and meaning.

This case study highlights the importance of documenting the web archiving process. Beyond the biases introduced by technical collection protocols, content selection methods, both within the IIPC and any institution dealing with digital hyper-memory, involve collective compromises to achieve the best possible representativeness while considering financial or organisational constraints. It is also crucial to recognise the cultural bias inherent in striving for representativeness and equity.

The preservation of online content in the MENA region during the COVID-19 crisis is also relevant to the decolonisation of web archives. Paradoxically, because of Archive-It, a significant part of the digital heritage born in the MENA region is being preserved by the Internet Archive foundation, even though several countries in the region are officially in conflict with the United States. Similar ethical questions can be raised for any European country with a colonial past. However, substantial financial resources are essential to break free from this hegemony and establish close collaborations with academic partners, archivists, and members of civil societies. The intention to build transnational collections with the best possible representativeness and inclusion of Southern countries should be encouraged. However, this requires financial means to engage in true co-construction with MENA region countries, fostering the sharing of practices and knowledge around archiving. This discussion can then move beyond archiving practices (Ghaddar & Caswell, 2019) to envision a genuine sharing of the means of knowledge production, contributing to a better understanding of the world. In this context, universities and libraries in the MENA region, particularly as they develop digital humanities (Bayoumi & Oliveau, 2020; Boustany 2023, p. 3), become privileged starting points for such initiatives and notably the ability of research centres to safeguard digital memory.

List of sources

- International Internet Preservation Consortium, *Novel Coronavirus (COVID-19)*, 2020. <https://archive-it.org/collections/13529>.
IIPC Website, <https://netpreserve.org>.
IIPC Weblog, <https://web.archive.org/web/20240304001642/https://netpreserveblog.wordpress.com/>.
IIPC Collaborative Collection: Novel Coronavirus (Covid-19) outbreak, Google form, 2020, https://web.archive.org/web/20240313155225/https://docs.google.com/forms/d/e/1FAIpQLSc4jyson7OGxe8HWaAm2s5BWvAug8mgastWEtNxbFYQbZX62g/viewform?usp=send_form.

Notes

- 1 Novel Coronavirus, IIPC, Archive-It, <https://archive-it.org/collections/13529>.
- 2 Content Development Working Group, IIPC, <https://web.archive.org/web/20210318013025/https://netpreserve.org/about-us/working-groups/content-development-working-group/>.
- 3 Author's translation from French: "pratique qui fait surgir une multiplicité d'énoncés comme autant d'événements réguliers, comme autant de choses offertes au traitement et à la manipulation".
- 4 This reflection is at the heart of the PICCH research program, which studies practices and uses related to colonial and postcolonial audiovisual heritage, <https://web.archive.org/web/20211125025632/https://picch-project.org/>, Dir. Pr. Daniela Petrelli.

- 5 Author's translation from French: "un humanisme critique qui serait fondé avant tout sur le partage de ce qui nous différencie, en deçà des absolus".
- 6 Novel Coronavirus, IIPC, Archive-It, <https://archive-it.org/collections/13529>.
- 7 Website of the Ministry of Health of Morocco on Website of the Ministry of Health of Morocco, April 3, 2020, <https://wayback.archive-it.org/13529/20200403164733/www.sante.gov.ma/sites/Ar/Pages/AccueilAr.aspx>, Dedicated subsite, created by the Ministry of Health of Jordan, version of June 4, 2020, Internet Archive: <https://wayback.archive-it.org/all/20200604221329/https://corona.moh.gov.jo/ar>.
- 8 University of Massachusetts Coronavirus Response Collection, <https://archive-it.org/collections/13736>.
- 9 National Library of Medicine (2020). Global Health Events web archive, Archive-It, Internet Archive, [https://archive-it.org/collections/4887?fc=websiteGroup:Coronavirus+disease+\(COVID-19\)+outbreak](https://archive-it.org/collections/4887?fc=websiteGroup:Coronavirus+disease+(COVID-19)+outbreak).
- 10 International and Area Studies Library, University of Illinois at Urbana-Champaign (2020). Global Civil Society Organization Responses to COVID-19, Archive-It, Internet Archive. <https://archive-it.org/collections/14004>.
- 11 Cf. Metronews Facebook page, www.facebook.com/metronewsps/about/?ref=page_internal.
- 12 Association of Internet Research, <https://aoir.org/aoir2022/aoir2022cfp/>.
- 13 Author's translation: "L'archive possède une relation organique à ce dont elle est archive causée par l'événement dont elle consigne la mémoire, l'archive est le signe, par son existence même, de l'événement et du fait qu'il a bien eu lieu."

References

- Aasman, S., Bingham, N., Brügger, N., De Wild, K., Gebeil, S., & Schafer, V. (2021). *Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections*. WARCnet Papers. https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Aasman_et_al_Chicken_and_Egg.pdf.
- Bachimont, B. (2021). Archive et mémoire: le numérique et les mnémophores. *Signata*, 12. <https://doi.org/10.4000/signata.2980>
- Bayoumi, H., & Oliveau, S. (2020). Digital archiving in the Arab world: Assessment and challenges. *Égypte/Monde arabe*, 22, 13–20. <https://doi.org/10.4000/ema.13141>.
- Berkhofer, R. F. (1995). *Beyond the great story: History as text and discourse*. Harvard University Press.
- Bingham, N., de Wild, K., Geeraert, F., & Nyvang, C. (2024). Surveying the landscape of COVID-19 web collections in European GLAM institutions: An explorative analysis. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge companion to transnational web archive studies*. Routledge.
- Boustany, J. (2023). Machrek-Maghreb: Les professionnels de l'information au défi du quotidien. *I2D — Information, Données & Documents*, 2, 6–16. <https://doi.org/10.3917/i2d.232.0006>.
- Chebbi, A. (2022). Entrevue avec Hajer Sahli, directrice des collections, Bibliothèque nationale de Tunisie. *Revue COSSI*, 11. https://doi.org/10.34745/numerev_1835.
- Crivello, M., & Dirèche, K. (Eds.). (2017). *Traversées des mémoires en Méditerranée: La réinvention du "lien" XIXe-XXe siècle*. Presses Universitaires de Provence. <https://doi.org/10.4000/books.pup.45698>.
- Derrida, J. (1996). *Archive fever: A Freudian impression*. University of Chicago Press.
- Foucault, M. (1969, rééd. 2012). *L'archéologie du savoir*. Gallimard.
- Gebeil, S. (2017). Les politiques de coopération audiovisuelle en Méditerranée (1995–2010). In M. Crivello & K. Dirèche (Eds.), *Traversées des mémoires en Méditerranée: La réinvention du "lien" XIXe-XXe* (pp. 191–200). Presses Universitaires de Provence.
- Gebeil, S. (2019). Archiver les traces numériques en Méditerranée, un défi aux multiples enjeux. *The Conversation France*, <https://theconversation.com/archiver-les-traces-numeriques-en-mediterranee-un-defi-aux-multiples-enjeux-119041>.
- Gebeil, S. (2021). *Website story: Histoire, mémoires et archives du web*. INA.
- Gebeil, S. (2022). Archive-moi si tu peux: À la recherche du patrimoine nativement numérique de la pandémie en Méditerranée. *Revue COSSI*, 11. https://doi.org/10.34745/numerev_1833.

An inclusive approach to web archiving

- Geeraert, F., & Bingham, N. (2020). *Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection*. WARCnet Papers, <https://web.archive.org/web/20230924000852/https://cc.au.dk/en/warcnet/warcnet-papers-and-special-reports>.
- Genovese, T. R. (2016). Decolonizing Archival Methodology: Combating hegemony and moving towards a collaborative archival environment. *AlterNative: An International Journal of Indigenous Peoples*, 12(1), 32–42. <https://doi.org/10.20507/AlterNative.2016.12.1.3>.
- Ghaddar, J. J., & Caswell, M. (2019). “To go beyond”: Towards a decolonial archival praxis. *Arch Sci*, 19, 71–85. <https://doi.org/10.1007/s10502-019-09311-1>.
- Halboub, E., Al-Ak'hali, M. S., Al-Mekhlafi, H., & Alhajj, M. N. (2021). Quality and readability of web-based Arabic health information on COVID-19: An infodemiological study. *BMC Public Health*, 21, 151. <https://doi.org/10.1186/s12889-021-10218-9>.
- Innocenti, P. (2016). Cultural networks and social inclusion of migrants in Europe: Heritage, ICT, and participatory governance. *International Information & Library Review*, 48(4), 274–286.
- Janssen, F. (2023). Engagement with decolonizing archival practices in the UK archives sector: A survey of archives workers' attitudes. *Archives and Records*, 44(1), 95–119. <https://doi.org/10.1080/23257962.2022.2117688>.
- Mbembe, A. (2013). *Sortir de la grande nuit. Essai sur l'Afrique décolonisée*. La Découverte.
- Musiani, F., Paloque-Bergès, C., Schafer, V., & Thierry, G. (2019). *Qu'est-ce qu'une archive du web?* OpenEdition Press. <https://doi.org/10.4000/books.oep.8740>.
- OECD. (2020). COVID-19 crisis in the MENA region: impact on gender equality and policy responses. 10 June 2020. https://web.archive.org/web/20230901000000*/www.oecd.org/coronavirus/policy-responses/covid-19-crisis-in-the-mena-region-impact-on-gender-equality-and-policy-responses-ee4cd4f4/.
- Schafer, V., & Winters, J. (2021). The values of web archives. *International Journal of Digital Humanities*, 2, 129–144. <https://doi.org/10.1007/s42803-021-00037-0>.
- Schoon, A., Mabweazara, M., Bosch, T., & Dugmore, H. (2020). Decolonising digital media research methods: Positioning African digital experiences as epistemic sites of knowledge production. *African Journalism Studies*, 41(4), 1–15. <https://doi.org/10.1080/23743670.2020.1865645>.
- Simondon, G. (2005). *L'individuation à la lumière des notions de forme et d'information*. Éditions Jérôme Millon.
- Smith, L. (2006). *Uses of heritage*. Routledge.
- Zghidi, S. (2021). Enjeux de la narration d'une révolution à partir d'archives vidéo: Instant tunisien. *Bulletin de l'AFAS*, 47, 54–71.

20

THE MANY LIVES OF WECHAT

Curating histories of the web in museum environments

Simone Natale

Efforts to archive and preserve the history of the web face enormous challenges since they deal with objects that are ever-changing, volatile, and inherently situated in specific locations, times, and experiences of use. As Brügger (2018) notes, “what is missing in a web archive is specifically the kind of information about ‘the web’ that the web itself usually offers us when we study the online web, such as search results and other information about what ‘the web’ looks like here and now” (p. 757). Much of the materials and the practical experiences that users have of the web are irremediably lost, due to reasons including the impossibility to record every activity, the complicated boundaries between private and public information, and the specific affordances of digital platforms where a large part of the online activities are situated but that remain largely outside the historical record (Ben-David, 2019).

Museum practice can provide a useful insight into some of the ways to engage with (not, of course, to solve) this conundrum. At different points across the last two centuries, museums have been forced to ask questions about what should be collected and preserved, and consequently, what counts as an “object,” where the boundaries between the tangible and the intangible can be drawn, and to what extent these boundaries should inform historical research and preservation (Graham and Cook, 2010; Foti, 2018). Today, the case of museums has the potential to contribute to approaches that help tackle broader questions about digital media and history, as curators, museum practitioners, and their institutions are forced to reconsider many of the assumptions that have underpinned museum collections and exhibitions for decades.

Drawing from findings of the Arts and Humanities Research Council (AHRC)-funded Circuits of Practice project, which explored the role of museums in constructing a historical heritage about computing and digital media (Natale, Parry, and Foti, 2022), this chapter interrogates how museums preserve histories of the web and reflects on how museum practice can provide insights relevant to wider efforts in web archiving and historiography. From 2019 to 2021, the Circuits of Practice project brought together curators from leading museums in the UK (Bletchley Park, the Centre for Computing History, The National Museum of Computing, the National Science and Media Museum, the Science Museum, the Victoria and Albert Museum) and international institutions (the Computer History Museum in the USA, the National Museum of Emerging Science and Innovation “Miraikan” in Japan,

The many lives of WeChat

the National Museum of Science and Technology “Leonardo Da Vinci” in Milan, Italy) to address the following overarching question: how do museums narrate modern computing? Findings from this project, particularly regarding the case of the acquisition of the social media platform WeChat by the Victoria and Albert Museum (V&A) help reflect on the importance of making web histories, and more broadly digital histories “alive” to museum audiences as well as to present and future web users.

In a series of research workshops, action research methods were used to carry out a series of practical interventions within the museum partners (Natale, Parry, and Foti, 2022). Action research brings together practical action and theoretical reflection in the pursuit of practical solutions to issues of pressing concern (Reason and Bradbury, 2001). Set within the overall design-thinking structure, the action research allowed Circuits of Practice to move from practical problems and issues faced by each of the museum partners to work creatively and responsively with practice-based methods. The workshops involving V&A were framed within the second research group of Circuits of Practice, which addressed the question of how artefacts and objects are mobilized to narrate the histories of modern computing. Based in London, V&A was founded in 1852 and is today the world’s largest museum of applied arts, decorative arts, and design. With a permanent collection of over 2.27 million objects, the museum preserves and exhibits the cultural heritage of design, covering the history of applied arts in the UK and in the world.

In this chapter, I first present the theoretical framework developed through the project. Then, I move on to the examination of a specific case: V&A’s curation of the social media platform WeChat (Kane, Gardner, and Park, 2024). Furthermore, this example provides insights that can empower the ongoing efforts of museums to curate digital objects, including web-based content and software, as well as wider approaches to transnational web historiography.

The many lives of digital objects

Rosemary Joyce and Susan Gillespie (2015) have criticized the use of the “life” metaphor to consider material objects, contending that such a notion replicates cultural bias that projects the dynamics of human lives (e.g. birth, death, etc.) into objects, and proposed to use instead the more neutral notion of “itineraries.” While it is true that things and especially objects related to computing are often the subject of anthropomorphization, the idea that things have “lives” can be helpful especially if it helps to acknowledge and critically examine how such projections and appropriations shape the ways material objects are appropriated and narrated across time. As things are inseparable from the social and cultural values and ideas that people attach and project onto them, examining their life trajectories framework illuminates the trajectories that shape our perception, representation, and engagement with material objects (Natale, 2016).

Using the plural form is crucial here since digital objects, which include web pages, software, hardware, and other materials related to web histories and archiving have not just one but multiple biographies. As Domínguez Rubio (2016) underlines, any consideration of objects and artefacts should pass through the acknowledgement that things are subject to material processes that unfold over time. It is impossible to understand objects without a consideration of their fragile and temporal realities (Bell and Geismar, 2009; Ingold, 2012). Although digital media are sometimes treated as immaterial, they are never “pure information,” as any single bit is stored and processed through and within material technologies and networks (Kirschenbaum, 2008; Rinehart and Ippolito, 2014).

Secondly, besides their material trajectories in time, digital objects are also integrated into wider social structures and interactions. A crucial tradition of thought in anthropology and art theory has unveiled that things, like people, have a social life (Appadurai, 1986; Gell, 1998). In Gell's own words, "social agency can be exercised relative to 'things,' and social agency can be exercised by 'things' (and also animals)" (Gell, 1998, pp. 17–18); therefore, artefacts can appear as agents in particular social situations. The meaning of objects, including the web (de Wild, 2019), is continually renegotiated within a process that informs their social status and their material circulation (Lesage, 2016), as well as the feelings and emotions people project onto them (Turkle, 2007).

Finally, objects also have discursive lives: they are not only material things but also things that people perceive, imagine, and talk about. Societies and individuals attribute ideas, appreciations, and values to specific objects at any moment. This level of discourse is not irrelevant to understanding objects' trajectories, as it informs how people, but also groups and institutions, approach and interact with them (Reynolds, 2018). If anything, this dimension is especially relevant to digital objects, due to the degree of technical opacity that often characterizes digital objects. As the functioning of hardware and software artefacts remains obscure to many, the narratives that circulate in the public sphere about digital objects play a strong role in directing their representation as well as their pragmatic uses: think, for instance, of the narratives and constant controversies about AI technologies and how they inform users' engagement with these technologies (Natale, 2019).

It is important to underline that these dimensions when it comes to examining the trajectories of any specific object or artefact, cannot be treated as separate or distinct. All digital objects have material, social and discursive "lives," and the relationship between the three dimensions is subject to constant transformations and modifications across time. The advantage of considering the three dimensions, therefore, is that they help acknowledge the dynamic character of digital objects — the fact, in other words, that digital objects cannot be defined or described in univocal or fixed ways, but need to be examined and understood against complex social, material and discursive environments.

Explorations into web archiving and historiography have shed light on the fact that only the application of a wide range of different approaches, methodologies, and practices can help fill at least some of the gaps left by existing projects and processes capturing online materials at different movements of time (Musso and Merletti, 2016; Ben-David and Amram, 2018). If it is true that a better understanding of the web of the past is a condition for gaining a more complete understanding of the web today (Brügger, 2018), only a perspective that is attentive to the multifaceted dimensions of digital objects' trajectories or "lives" has the potential to truly enhance such understanding. Preserving the past of the web, in this sense, entails more than just the material preservation of web pages: it also needs to embrace the social and the discursive dimensions that are an indissoluble part of what gives "lives" to the web and digital media in general.

An online platform is always more than one: The case of WeChat at the Victoria and Albert Museum

As cultural heritage institutions, museums aim at both the preservation of objects of the past and their exhibition to the public; therefore, their practices are particularly focused on unveiling the different dimensions of digital objects' lives (Keramidas, 2015; Foti, 2018). In both exhibiting and collecting, for instance, the Software History Center of the Computer

The many lives of WeChat

History Museum in California, US, developed a holistic approach to software preservation which included its technical and material dimension (e.g. preservation of source code, but also the material culture of franchising and packaging objects associated to specific pieces of software), its social dimension (e.g. oral testimonies, the re-enactment of hardware and software interfaces that shape users' experiences) and its discursive dimensions (e.g. commentaries and reports shedding light on the software's reception and impact, see Weber, 2016).

The acquisition of a social media platform operated by a major cultural heritage institution, the V&A in London, UK, provides an ideal example of how this approach can be mobilized to preserve and exhibit content and software that constitutes the history of the web. Three research workshops in the Circuits of Practice project focused on the case of V&A's acquisition of WeChat (Cormier, 2017), a digital platform with over one billion users, most of them in China (Negro, 2017, pp. 193–208). The workshops were conducted with the participation of three V&A staff members, Corinna Gardner, Natalie Kane, and Juhee Park. The first workshop, called "REFLECT," aimed to identify and elaborate key concepts and a common vocabulary to be activated in the collaborative research. The second workshop, called "READ," aimed to apply the conceptual tools developed in the first workshop on the case study of WeChat's acquisition at the V&A. Finally, the third workshop was called "RESPOND" and aimed to mobilize the work done in the first two workshops to feedback with practical suggestions and inputs that contribute to curatorial knowledge and practices at V&A. The complex strategies and practices that V&A developed and applied to complete this acquisition represent an exceptionally useful case study of efforts to integrate into the museum collection, a web-based platform whose ever-changing, volatile character makes it particularly challenging to preserve (Ben-David, 2019).

WeChat (*Weixin* in Chinese) is a social media platform launched in 2011 by Tencent, a major actor in China's tech industry. Originally marketed as a messaging app, WeChat has taken up more and more functions, such as booking systems for different kinds of public and private services as well as electronic payments (Peng and Wang, 2021). Due to the breadth of available functions and its pervasive role in the everyday experience of many of its users, WeChat has been indicated as one of the most evident examples of the infrastructuralization of digital platforms. As Plantin and de Seta argue (2019, p. 262), "Because of the ever-increasing number of WeChat's functions, it has become increasingly hard to live in China without a WeChat account."

The acquisition of WeChat was proposed in 2015 by V&A curators Luisa Mengoni, Corinna Gardner, and Brendan Cormier (Kane, Gardner, and Park, 2024). The opportunity to add WeChat to V&A's collection was first provided in the context of the exhibition "Values of Design" for the V&A Gallery at the Design Society in Shenzhen, China. Curators from V&A met for the first time with members of the Tencent design team in 2015, expressing interest in adding WeChat to the museum's permanent collection. As V&A Senior Design Curator Brendan Cormier (2017) recounts, V&A curators at this stage "weren't sure ourselves how that might work," due to the particular nature of this object. Curating software presents specific challenges to museums, not only for its allegedly immaterial nature but also especially for its situated and performative character (Foti, 2018). As historian of computing Michael S. Mahoney points out, what makes the history of software hard is that it is not only or even not primarily about computers: software reflects the histories of the communities that created them, and the cultural, social, and practical circumstances underpinning its adoptions and uses (Mahoney, 2008). An additional complication, moreover, is that the very

functioning of a social media platform such as WeChat relies on the information shared by its community of users. This not only opens the question of the difficulty of distinguishing between software and data but also problematizes the possibility of drawing clear boundaries between digital objects and networks.

As an institution whose mandate revolves around the cultural heritage of arts and design, V&A was strategically placed to tackle such challenges (Atkinson, 2018). In collecting and exhibiting artefacts related to histories of design, V&A throughout its history has often faced the need to consider elements such as the experience of users and the performative dimensions of objects (Burton, 1999). The accumulated expertise and cultural identity of the institution informed the approach of V&A curators and helped them respond to the challenges posed by the acquisition of WeChat (Cormier, 2017).

Approaches that consider reproducibility of the technical dimension as crucial to the conservation of digital objects (Agar, 1998) tend to consider the preservation of source code, and thus, the capacity to reproduce the technical functioning of the object, as essential to preserving software (Di Cosmo, Gruenpeter, and Zacchiroli, 2018). Such an approach presents, however, significant problems, including the need to adapt to different standards and programming languages and to preserve adequate hardware and software infrastructures where the software can run. Moreover, the difficulty — often, the impossibility — of having access to proprietary software code makes the preservation of source code impossible in many cases.

The design-oriented approach that characterizes the V&A, in this regard, provided the context to develop a holistic and flexible approach that brought the software's social and discursive dimensions to the centre stage. Since acquiring a copy of a social network app would have been of little use without its insertion in the original networked space, the museum acquired not only a copy of the software but also an offline demonstration, in the form of a video, which illuminates crucial aspects of the software's user experience. The acquisition also included materials such as interviews and testimonies of the design process and development of the WeChat interface, as well as a selection of contents such as stickers and GIFs and a 2011 demo .apk file of the first release of WeChat, which had been used to pitch the network to Apple's AppStore (Kane, Gardner, and Park, 2024). In V&A exhibition spaces, moreover, WeChat was framed alongside other objects and narratives that illustrated wider trajectories within the history of design. This contextualization in the museum's gallery further emphasized issues such as user experience and interface design for the presentation of WeChat.

The acquisition of WeChat and its exhibition in V&A galleries, therefore, were the result of the recognition that WeChat, like every other software, is not a discrete object that can be described and defined in univocal terms. It is a multifaceted object that escapes fixed and rigid attributions of meaning and cannot be strictly separated from the wider technological, social and discursive networks that surround it. As a result, its acquisition within the museum's collection required the assemblage of not one but a plurality of objects that gave access to its multiple dimensions, emphasizing those that were deemed of special relevance to the museum, its visitors, and its mandate of preserving the cultural heritage of design.

WeChat's acquisition, in this sense, shows how practices for curating software can emerge not only and not predominantly by tackling the question of "what is software" in absolute terms, but rather by interrogating what software means from the particular point of view of the institution that operated the acquisition. The framing of WeChat as an object relevant

The many lives of WeChat

to the history of design functioned as a lens that helped privilege specific dimensions of the object, such as the user experience, which appears central to the contextualization of software within the V&A.

At a discursive level, the exhibition of WeChat activated not a single narrative but a plurality of narratives. On the one side, exhibiting a social media platform evokes the personal experiences of visitors with software and social networks, opening up the opportunity for mobilizing private and collective memories regarding life online (Papacharissi, 2002). On the other side, the fact that this specific platform is strongly characterized as Chinese (Negro, Balbi, and Bory, 2020) highlights the fact that digital media are both universal as well as culturally specific (Bell et al., 2018). The acquisition of WeChat, finally, produced a meta-narrative about the agency of V&A itself, demonstrating the museum's capacity to shape new approaches that will impact the politics of curating and exhibiting digital objects (Cormier, 2017).

V&A's flexible, design-oriented approach to software objects is enabling the institution to further advance its politics of acquisition in the area. The experience with WeChat demonstrated that acquisitions of software in the museum's collection can be rigorous and productive even in the absence of elements that other cultural heritage institutions might consider central, such as the acquisition of software's source code. This is important also because it means that the museum might look for future acquisition of other networked, proprietary apps that shape the everyday experiences of large masses of people around the world.

More broadly, the case under the exam shows that rather than one definition of software, multiple definitions are situated and contextualized alongside specific approaches, contexts, and points of view. While particular dimensions of WeChat were central to its acquisition at V&A, other institutions whose scope relates, for instance, to the history of science and technology rather than design, would have privileged different dimensions of the object and thus, might have acquired different materials to be preserved and exhibited at the museum. The case of V&A's acquisition, in this sense, reminds us that digital objects are not to be defined in absolute terms but rather against the contextual and the circuitry of relations and attributions that surround and add "lives" to them.

Conclusion

The case examined shows how in preservation practices within a museum environment, digital objects are never defined or understood in univocal ways. Their definitions, meanings and values emerge through and within a network of relations between the objects, the other objects in the collection, the curators and practitioners, the institutions, and the visitors of the museum. In preparing WeChat for both the museum's exhibition and collection, curators selected the "lives" of the platform that were more relevant to the institution, its missions and its target public, but at the same time mobilized elements of all three dimensions highlighted above, i.e. the material, the social, and the discursive lives of WeChat. In this context, new practices on how to curate web-based objects developed not only and not predominantly by posing the question of "what is software," but rather by interrogating what software means from the point of view of a design history museum.

The practice-based research conducted within the Circuits of Practice project provides not just insight into museum practice but also, more broadly, a gateway to consider the multidimensional nature of web content and its impact on web archiving and preservation. As Domínguez Rubio (2016) put it, "Objects are anything but given or self-evident." The

museum functions in this regard as a laboratory in which meanings, uses and definitions of digital objects are self-reflectively negotiated and where the relational circumstances that characterize digital objects become manifest.

Importantly, rather than diverging from the characteristics of digital objects outside of the museum, the trajectories of digital objects in the museum's collections and exhibitions are in a relationship of continuity with the material, social and narrative trajectories of digital objects outside the museum (Keramidas, 2015). Much like digital objects become repositories of multiple social uses, meanings, and exchanges as they are created, used, circulated, and eventually discarded in everyday life, so do digital objects establish a relational and iterative social "circuit of meaning" within the museum environment. The relational character of digital objects resonates, moreover, with the multiple ways in which visitors navigate the exhibitions, engaging with digital objects just like the ones they encounter in their everyday lives and projecting their own previous experiences and ideas about digital media.

One aspect that deserves additional emphasis is the transnational character of V&A's acquisition. Online platforms may trespass national and linguistic contexts, yet, at the same time they often bear traces of the cultures in which they emerge and are used (Peng and Wang, 2021). WeChat's acquisition worked, in this sense, also as a strong acknowledgement of the need for perspectives that are sensitive to the global dimension of the web and of digital media in general (Arora, 2019), and therefore, of digital heritage. Moreover, considering how tech corporations customarily struggle to shape histories of the digital placing their role and agency at centre stage (Natale, Bory, and Balbi, 2019), WeChat decision to collaborate with V&A should also be read alongside Tencent's wider efforts to expand in the Western market. In fact, Tencent made in the 2010s some significant marketing investments to expand its platform across its traditional areas of use. In contrast to other social media platforms which were launched originally in China, such as TikTok, this strategy was, however, only partially successful for WeChat (Negro, 2017).

Museum practice, in conclusion, invites us to emphasize the need to consider the multiple dimensions of web contents and software as a useful pathway to question what needs to be preserved on the web and how to do so (Parry, 2010). As Brenda Laurel (1993) noted, a "computer-based representation without a human participant is like the sound of a tree falling in the proverbial uninhabited forest." The lively community of historians, archivists, librarians, curators, practitioners, and volunteers who contributed to this endeavour have developed very different approaches and strategies to archive the web (Brügger, 2018). The diversity of the professional and institutional contexts to which such efforts are based — from museums to archives, from public libraries to user-based practice — cannot but add to the breadth of perspectives through which the different dimensions of the web and digital content are introduced and preserved in the historical record.

While museums and archives have different mandates and approaches, they share their broader mission as repositories of public knowledge in the information age (Usherwood, Wilson, and Bryson, 2005). Scholars have called for the need to diversify institutions such as libraries, archives and museums, as each of these operationalizes very different ideas and practices about memory and preservation (Robinson, 2012). Much, however, can be gained also from perspectives that consider the common challenges that these institutions address and take into account the different responses and choices that they make toward a common effort.

The many lives of WeChat

At V&A, decisions regarding the acquisition of a complex digital object such as a social media platform were informed not only by the need to make WeChat into an element of both the collection and its exhibition but also by considerations about the institution's mission as a museum of design. Other institutions would have made very different decisions about which materials to acquire to introduce a record of WeChat in the museum (Natale, Foti, and Parry, 2024). Similar choices, of course, limited the range of dimensions that were collected about WeChat in this acquisition; but at the same time, they ensured that aspects of WeChat that would have not made it into the record at other institutions — for instance, video recordings, providing a dynamic testimony of the platform's interface — were included in the museum's collection.

Transposed into wider debates about web preservation, therefore, the case examined here shows how much can be gained from considering the challenge of web archiving as a collective effort. Because digital media are multifaceted objects that have not one but multiple “lives,” the diversity of institutions (as well as groups and individuals) that contribute to the preservation of elements of the web has special value. The challenge of preserving the web, therefore, entails the involvement of a variety of actors, each shedding light on specific dimensions of objects related to the histories of the web.

Acknowledgements

The author is grateful to the Arts and Humanities Research Council (AHRC) UK, for its generous support of the “Circuits of Practice” project 2020–2021 (AH/T00276X/1). I would like to thank our project partners and particularly the V&A curators who participated in the workshops and helped shed light on the case presented in this chapter: Corinna Gardner, Natalie Kane and Juhee Park.

References

- Agar, J. (1998). Digital patina: Texts, spirit, and the first computer. *History and Technology*, 15(1–2), 121–135.
- Appadurai, A. (1986). *The social life of things: Commodities in cultural perspective*. Cambridge University Press.
- Arora, P. (2019). *The next billion users: Digital life beyond the West*. Harvard University Press.
- Atkinson, P. (2018). The Role of Design History in the Museology of Computing Technology. In O. Moret (Ed.), *Back to the future: The future in the past: Conference Proceedings* (pp. 464–468). Edicions de la Universitat de Barcelona.
- Bell, J. A., & Geismar, H. (2009). Materializing Oceania: New ethnographies of things in Melanesia and Polynesia. *The Australian Journal of Anthropology*, 20(1), 3–27.
- Bell, J. A., Kuipers, J., Hazen, J., Kemble, A., & Kobak, B. (2018). The Materiality of Cell Phones Repair: Re-making Commodities. *Anthropological Quarterly*, 91(2), 603–632.
- Ben-David, A. (2019). 2014 not found: A cross-platform approach to retrospective web archiving. *Internet Histories*, 3(3–4), 316–342.
- Ben-David, A., & Amram, A. (2018). The Internet Archive and the socio-technical construction of historical facts. *Internet Histories*, 2(1–2), 179–201.
- Brügger, N. (2018). *The archived web: Doing history in the digital age*. MIT Press.
- Burton, A. (1999). *Vision & accident: The story of the Victoria and Albert Museum*. V&A Publications.
- Cormier, B. (2017). *How we collected WeChat*. Available online at <https://web.archive.org/web/20230715000000/www.vam.ac.uk/blog/international-initiatives/how-we-collected-wechat> (retrieved 24 February 2021).
- De Wild, K. (2019). *Internet art and agency: The social lives of online artworks*. PhD Dissertation, University of Dundee, Dundee.

- Di Cosmo, R., Gruenpeter, M., & Zacchioli, S. (2018). Identifiers for digital objects: The case of software source code preservation. In *iPRES 2018: 15th International Conference on Digital Preservation* (pp. 1–9). Amsterdam, The Netherlands.
- Domínguez Rubio, F. (2016). On the discrepancy between objects and things: An ecological approach. *Journal of Material Culture*, 21(1), 59–86.
- Foti, P. (2018). *Collecting and exhibiting computer-based technology: Expert curation at the museums of the Smithsonian institution*. Routledge.
- Gell, A. (1998). *Art and agency: An anthropological theory*. Clarendon Press.
- Graham, B., & Cook, S. (2010). *Rethinking curating: Art after new media*. MIT Press.
- Ingold, T. (2012). Toward an ecology of materials. *Annual Review of Anthropology*, 41, 427–442.
- Joyce, R. A., & Gillespie, S. D. (Eds.). (2015). *Things in motion: Object itineraries in anthropological practice*. SAR Press.
- Kane, N., Gardner, C., & Park, J. (2024). Social media enters the museum: Collecting WeChat at the Victoria & Albert Museum. In S. Natale, P. Foti, & R. Parry (Eds.), *Museums and the history of computing: Objects, narratives and practice*. Routledge.
- Keramidas, K. (2015). *The interface experience: A user's guide*. Bard Graduate Center.
- Kirschbaum, M. G. (2008). *Mechanisms: New media and the forensic imagination*. MIT Press.
- Laurel, B. (1993). *Computers as theatre*. Addison-Wesley.
- Lesage, F. (2016). A cultural biography of application software. In C. Paterson, D. Lee, & A. Saha (Eds.), *Advancing media production research: Shifting sites, methods, and politics* (pp. 217–232). Palgrave Macmillan.
- Mahoney, M. S. (2008). What makes the history of software hard. *IEEE Annals of the History of Computing*, 30(3), 8–18.
- Musso, M., & Merletti, F. (2016). This is the future: A reconstruction of the UK business web space (1996–2001). *New Media and Society*, 18(7), 1120–1142.
- Natale, S. (2016). Unveiling the biographies of media: On the role of narratives, anecdotes and storytelling in the construction of new media's histories. *Communication Theory*, 26(4), 431–449.
- Natale, S. (2019). If software is narrative: Joseph Weizenbaum, artificial intelligence and the biographies of ELIZA. *New Media and Society*, 21(3), 712–728.
- Natale, S., Bory, P., & Balbi, G. (2019). The rise of corporational determinism: Digital media corporations and narratives of media change. *Critical Studies in Media Communication*, 36(4), 323–338.
- Natale, S., Foti, P., & Parry, R. (Eds.). (2024). *Museums and the history of computing: Objects, narratives and practice*. Routledge.
- Natale, S., Parry, R., & Foti, P. (2022). *Circuits of Practice Research Report: Narrating Histories of Computing and Digital Media in Museum Environments*. Centre for Research in Communication and Culture, Loughborough University.
- Negro, G. (2017). *The Internet in China: From infrastructure to a nascent civil society*. Palgrave Macmillan.
- Negro, G., Balbi, G., & Bory, P. (2020). The path to WeChat: How Tencent's culture shaped the most popular Chinese app, 1998–2011. *Global Media and Communication*, 16(2), 208–226.
- Papacharissi, Z. (2002). The virtual sphere: The internet as a public sphere. *New Media & Society*, 4(1), 9–27.
- Parry, R. (2010). *Museums in a digital age*. Routledge.
- Peng, W., & Wang, W. Y. (2021). Buying on Weixin/WeChat: Proposing a sociomaterial approach of platform studies. *Media, Culture & Society*, 43(5), 945–956.
- Plantin, J. C., & de Seta, G. (2019). WeChat as infrastructure: The techno-nationalist shaping of Chinese digital platforms. *Chinese Journal of Communication*, 12(3), 257–273.
- Reason, P., & Bradbury, H. (Eds.). (2001). *Handbook of action research: Participative inquiry and practice*. Sage.
- Reynolds, D. (2018). *Media in mind*. Oxford University Press.
- Rinehart, R., & Ippolito, J. (2014). *Re-collection: Art, new media, and social memory*. MIT Press.
- Robinson, H. (2012). Remembering things differently: Museums, libraries and archives as memory institutions and the implications for convergence. *Museum Management and Curatorship*, 27(4), 413–429.

The many lives of WeChat

- Turkle, S. (2007). *Evocative objects: Things we think with*. MIT Press.
- Usherwood, B., Wilson, K., & Bryson, J. (2005). Relevant repositories of public knowledge?: Libraries, museums and archives in “the information age.” *Journal of Librarianship and Information Science*, 37(2), 89–98.
- Weber, M. (2016). Self-fulfilling history: How narrative shapes preservation of the online World. *Information & Culture*, 51(1), 54–80.

21

PARTICIPATION, PLATFORMS, AND CULTURAL HERITAGE

Web archiving challenges

Marta Severo

Introduction

Over the past few decades, there has been a remarkable surge in citizen participation in the field of cultural heritage (Giaccardi, 2012). Citizens are now recognized as key actors in the production, acknowledgment, and preservation of cultural heritage. Such a trend concerns all types of heritage, from monuments to landscapes, from oral traditions to manuscripts, until more recent products of cultural industries (comics, films, etc.). Intangible cultural heritage constitutes an exemplary case of this trend. Designated as a heritage category by UNESCO in 2003 (Aikawa, 2004), intangible cultural heritage encompasses various elements, including practices, representations, expressions, knowledge, and skills, that communities, groups, and individuals recognize as integral to their cultural identity. Intangible cultural heritage is not only created by citizens but also must be institutionally acknowledged and preserved by citizens to guarantee intergenerational retransmission (Bortolotto, 2007).

This participatory trend has been significantly amplified by the advent of digital technologies, which have opened up new avenues for citizen expression and collaboration (Bonacchi et al., 2019). Digital culture has empowered citizens, promoting individual creativity (Jenkins, 2006; Leadbeater & Millet, 2004). Moreover, the rise of Web 2.0 and then of platformization has facilitated interactive and collaborative practices, allowing the exchange of ideas and goods (Gillespie, 2010). The concept of platformization is here used in its extensive sense to underline the variety of digitally developed environments, their level of organization, and their impact on social relations. Platformization embodies the transformative process through which digital platforms reshape social, economic, and cultural interactions. As Van Dijck (2013) asserts, platforms serve as infrastructures facilitating connectivity, fundamentally altering how individuals, communities, and institutions engage with information, goods, and services. Platforms, ranging from social media to e-commerce marketplaces or even institutional sites, act as intermediaries facilitating interactions among users, content, and services (Kenney & Zysman, 2016) and as arenas for self-expression, community formation, and political discourse (boyd & Ellison, 2007). Consequently, platformization engenders complex negotiations between empowerment and exploitation, participation and control, highlighting the dialectical nature of digital connectivity (Couldry & Mejias, 2019).

Participation, platforms, and cultural heritage

In the domain of cultural heritage, this shift has led to a proliferation of digital participatory platforms, enabling contributors, often referred to as amateurs, to share content related to their cultural objects, monuments, traditions, and places (Severo, 2022). Even established institutions such as archives, libraries, and museums have joined the movement, launching their own participatory platforms, often based on crowdsourcing tools, to engage people in constructing knowledge about cultural objects (Ridge, 2014). These digital tools, whether institutional, commercial, or community-based, are now viewed as powerful facilitators of new forms of engagement, knowledge-sharing, and dissemination in the cultural heritage field.

However, with the widespread use of digital participatory platforms concerning cultural heritage, comes the imperative to consider their long-term preservation. Hence, this chapter seeks to examine the theoretical, technical, and ethical challenges associated with archiving these digital participatory tools. It aims to contribute to the literature on web archives by exploring the importance and variety of this phenomenon through a diachronic perspective. From a broader viewpoint, it intends to study the complexity of participation, notably digital participation, preserved in web archives. Taking the researcher's standpoint, the chapter will explore the potential of web archives to investigate this transnational phenomenon (Brügger, 2018; Brügger, 2013; Ben-David & Huurdeman, 2014). The first part will present an overview of the socio-technical solutions underlying the umbrella term “participatory platforms” for cultural heritage, with a focus on distinguishing between institutional and amateur platforms. Then, the second part will offer an empirical analysis of six participatory platforms, three institutional and three amateur ones, detailing their organization and exploring the possibilities that web archives offer to study their evolution over time. Finally, the last part will engage in a discussion based on the empirical analysis, addressing the challenges and prospects concerning web archiving digital participation in the realm of cultural heritage, encompassing theoretical, empirical, and ethical considerations.

**Participatory platforms for cultural heritage:
Institutional vs. amateur tools**

When examining digital participation in the cultural heritage realm, scholars typically distinguish two types of digital tools: institutional participatory platforms, created by cultural institutions (e.g., archives or libraries) to gather knowledge about their collections, often following a top-down model of participation; and amateur platforms, established by non-institutional actors, facilitating knowledge exchange related to cultural heritage through a decentralized approach.

Within an institutional context, “participatory platforms” are commonly associated with crowdsourcing platforms, digital tools provided by institutions, where non-professional volunteers can perform simple tasks (e.g., transcribing handwritten texts or indexing documents) to aid in building cultural heritage databases for safeguarding purposes (Ridge, 2014). Such platforms are often utilized in citizen science projects to involve citizens in co-constructing official knowledge (Haklay, 2015). Conversely, amateur platforms, rooted in participatory culture (Jenkins, 2006), initially emerged as spontaneous initiatives by contributors, as seen in the iconic case of Wikipedia. Enthusiasts of music, arts, or cinema utilized various digital spaces (e.g., blogs, forums, social media) to share their creations or critique cultural objects (Hennion, 2001). Presently, commercial actors manage amateur platforms that leverage users' passions, ranging from photography on Instagram to music

on Spotify, or writing and reading on Wattpad and LibraryThings. In the cultural heritage domain, Facebook groups, Instagram hashtags, YouTube accounts, and Wikipedia projects have been established to share content (e.g., photos, videos, texts, ideas, emotions) about monuments, oral traditions, and other cultural items or practices.

Scholars tend to discern contrasting characteristics between these two types of platforms, primarily based on the type of content, actors' organization, and the focus of actions. Firstly, concerning the content, institutional platforms enforce standardized expressions to ensure data quality and compatibility with existing scientific knowledge, employing tools like pre-filled forms, checkboxes, or other specific formats. In contrast, amateur platforms offer users the freedom to create, share, and interact with content unrestrictedly, engaging in interactions such as likes, comments, and reactions, while still adhering to platform-specific guidelines (e.g., character limits on Twitter). Secondly, in terms of actors' organization, institutional platforms predominantly define rules from the perspective of the institution, emphasizing individual empowerment but not placing significant emphasis on social networking. Conversely, amateur platforms are perceived as democratic and horizontal spaces, where contributors have the autonomy to establish their own rules through social interactions and mutual sharing. These platforms facilitate connectivity, fostering relationships among users across geographical boundaries, creating diverse networks, encouraging social interactions, and fostering a sense of community. Thirdly, concerning the focus of actions, institutional platforms prioritize the product generated on the platform over the producers themselves. The platform's primary objective is to generate knowledge, and as such, contributions are often anonymous or not prominently linked to individual users (Iranowska, 2019). In contrast, actors in amateur platforms participate primarily for personal motivations, seeking to strengthen their digital identity. These platforms' interfaces provide users with tools to create profiles that can be assessed by others (boyd & Ellison, 2007; Papacharissi, 2010).

Importantly, while these tools may appear separate and opposed from a theoretical standpoint, their coexistence and interconnectedness on the web are vital as users frequently navigate between institutional and amateur ecosystems to collect or share knowledge. For instance, genealogists actively contribute to both institutional platforms like the crowdsourcing portals of national or local archives for relevant documents and non-institutional platforms like commercial websites (e.g. Ancestry) for co-constructing their genealogical trees (Tasker & Liew, 2020). Therefore, studying these platforms over time entails not only understanding their individual properties but also considering the surrounding ecosystem, where contributors interact and where users' participation is fostered. In the next section, we conduct an empirical study of six platforms to investigate how web archives can support researchers in comprehending this intricate transnational phenomenon.

Studying participatory platforms: An empirical analysis

Methodology

To explore the challenges related to web archiving participatory platforms, we conducted an empirical analysis on a selected sample of platforms. Initially, we established an observatory of cultural participatory platforms¹ during a prior project, providing valuable insights into the geographical and typological diversity of this phenomenon. Based on this, we identified six platforms, three amateur, and three institutional, which offer significant representation of

Participation, platforms, and cultural heritage

our research object due to their historical or geographical significance, as well as the configuration of actors and tools they encompass. For the institutional platforms, we chose: *Transcribe Bentham*,² one of the earliest and most renowned platforms of its kind; *AnnoTate*,³ a prominent campaign on the Zooniverse website within the field of humanities; *Memory of Men*,⁴ a French platform established by the Ministry of Armed Forces, hosting one of the most successful crowdsourcing campaigns related to the First World War. Among the amateur platforms, our selection includes: *Goodreads*,⁵ the largest transnational social cataloging website (a subsidiary of Amazon) that enables individuals to create and search for book reviews, annotations, and quotes; *I.N.D.U.C.K.S.*,⁶ a worldwide database dedicated to Disney comics, originally created by a spontaneous volunteer and now populated by hundreds of users; *Don't Lose Your Way*,⁷ a UK platform initiated by The Ramblers' Association, aiming to map all paths that risk being lost rights of way.

To analyse these platforms, we developed a comprehensive grid based on the three principal properties of participatory platforms: content, actors' organization, and actions' focus. For each property, we determined the available sources to gather information on its evolution over time. Regarding content, our focus centred on the format, creation, and collection of data. We aimed to reconstruct the process of content production over time and identify changes in the interface of content creation and related formats. Concerning actors' organization, we drew our attention toward the mechanisms of interaction between participants. Our aim was to discern changes in the organization over time, identify the entities establishing the platform's rules, and explore the social networks related to the platform, including their development and impact. Lastly, concerning the focus of actions, we delved deep into the motivations of the involved actors, both institutions and amateurs. We sought to identify these motivations and examine whether contributors could build their digital identity and acquire new competencies through their participation.

To collect this information about the platform's history, the involved actors, and the generated content, our analysis integrated both digital methods and digital ethnography (Rogers, 2019; Kozinets, 2010). We extended our investigation beyond the platform itself, considering surrounding tools like social media accounts and forums that actively contribute to the organization of participation. Due to space limitations, a comprehensive exploration of case studies within this chapter is not possible. Therefore, we will primarily focus on the main findings regarding web archives, as they substantiate the discussion in the third part.

Institutional platforms

We have selected three distinct institutional platforms due to their significant transnational importance, each serving unique purposes. *Transcribe Bentham*, established in 2010 by University College London, stands as one of the earliest platforms of its kind. In its 13 years of existence, it has amassed approximately 60,000 unpublished manuscripts written by the philosopher and reformer, Jeremy Bentham. Of these, more than 30,000 manuscripts have been transcribed by volunteers. This ongoing project continues to attract thousands of contributors from around the world every day (Causer & Wallace, 2012). In contrast, *AnnoTate* was a closed project that operated from 2015 to 2017. It holds a dual interest. Firstly, it is based on the Zooniverse platform, an international crowdsourcing environment hosting projects from institutions in a wide variety of countries. Secondly, it is the inaugural project of its kind created by an art gallery, Tate Britain, in collaboration with Zooniverse (Tate, 2015; Barber, 2018). Its objective was to annotate a substantial corpus of documents

from the Tate collection, including diaries, sketchbooks, letters, and more, with the aim of “revealing the secret lives of artists.” Lastly, *Memory of Men* (known as “Mémoire des hommes” in French) is a national project with international significance due to the impactful participation it has garnered. On this platform, hundreds of amateurs transcribed more than 1.5 million entries over a span of five years, thereby completing the database of the “Dead for France,” an honorary acknowledgement attributed to soldiers of the First World War.

In terms of content, these three projects share the common feature that the globality of transcribed documents is readily accessible through institutional repositories (Causser & Wallace, 2018). Transcribe Bentham’s repository is housed in the UCL Library’s digital repository, while AnnoTate’s documents are stored in the Tate Archive. Similarly, the French project’s transcriptions are maintained within the Ministry of Armed Forces’ database. It is important to highlight two key aspects. Firstly, these data collections are published in separate spaces from the platforms where transcription activities occur. Secondly, they solely retain the generated data, lacking information about transcription actions and contributors. Put differently, distinguishing between pre-existing content and participant contributions is unfeasible. Some insights into content evolution during crowdsourcing campaigns are retrievable from web archives. In the case of Transcribe Bentham, the UK Web Archive encompasses two captures, whereas the Wayback Machine, with approximately a thousand captures of various project pages, offers a glimpse into the platform’s content evolution. Regarding AnnoTate, the situation is more complex due to the project’s initiation on Zooniverse and subsequent migration to a new URL⁸ in 2015. Presently, evidence of the crowdsourcing campaign is scarcely online, with minimal captures (6 in the UK Web Archive and 96 in the Wayback Machine, exclusively from the new URL since 2016). As such, insights into content evolution mainly derive from third-party sources like institutional documents or media articles found in both live web and web archives. The French project follows a similar trajectory, with limited captures in web archives and reliance on third-party sources. However, a noteworthy facet of this project lies in the informal volunteer group known as “1 Day 1 Soldier” (in French, “1 Jour 1 Poilu”), which initiated a parallel challenge on social media, posting their edits on Twitter. This inventive approach partially enables the tracking of content evolution through tweets containing the hashtag “#1J1P.” Fortunately, this hashtag has been archived and is accessible within the web archives of the INA (the French Institute for Audiovisual Content) (Severo, 2021).

Regarding the organization of actors, there are notable distinctions among the three projects. In the case of Transcribe Bentham, the crowdsourcing platform employs MediaWiki, an open-source wiki software. This wiki encompasses various pages that facilitate the tracking of platform organization and actor roles. Notably, the “Transcription Guidelines” page charts the evolution of rules, the “Benthamometer” monitors contribution progress, and the “Leaderboard” showcases top users. The “history” page of the Wiki emerges as a valuable resource, allowing reconstruction of page evolution (though not applicable to the Leaderboard). For example, based on this tool, it is possible to trace the change of rules over time. However, the Wiki lacks the capacity to capture interactions between actors and social networks. Pertinent insights related to this feature can be gleaned from project-affiliated social media accounts (e.g., Twitter, a designated hashtag, and a Facebook page, see Causser & Terras, 2014). Nonetheless, accessing such data may necessitate the use of the commercial Twitter API and personal Facebook archives (if accessible to researchers). Turning to AnnoTate, which generally mirrors the scenario of Zooniverse projects, information remains scant. Zooniverse provides an “About” page for each project; regrettably, this page is seldom

Participation, platforms, and cultural heritage



#1J1P 17/09/ 2016 : Georges CARON, 25 ans, maçon dans l'#Oise, est tué à Berny-en-Santerre tinyurl.com/z6wz3q4

PARTIE À REMPLIR PAR LE CORPS.

Nom CARON
 Prénoms Georges
 Grade Soldat
 Corps 120^e REG^o D'ARTILLERIE
 N^o 120.5 an Corps. — Cl. 1901
 Matricule 120 an Recrutement Compagnie
 Mort pour la France le 24 septembre 1916
 à Berny-en-Santerre (Somme)
 Genre de mort Eue à l'ennemi
 Né le 12 juillet 1891
 à Beiry-sur-Oise Département Oise
 Arr^o municipal (1^{er} Paris et Lyon), à défaut rue et N^o
 Jugement rendu le 22 novembre 1916
 par le Tribunal de Beiry-sur-Oise
 acte ou jugement transcrit le 22 novembre 1916
 N^o du registre d'état civil Beiry-sur-Oise (Oise)
 534-708-1021. (36354)

Mémoire des Hommes / Ministère de la Défense
Fiche transcrite le 17/09/2016

1 Jour - 1 Poilu @1J1Poilu · 1 août
 Derrière chaque fiche, un homme... Ne les oublions pas !
 Indexons nos Poilus Morts pour la France sur Mémoire des Hommes !

Figure 21.1 Tweet dedicated to the Georges Caron entry (colour version in Zenodo repository). Printed by permission from Jean-Michel Gilot.

archived in web archives, potentially obscuring any edits made over time. Additionally, a “Talk” page serves as a forum for user discussions, and although often accessible (as with AnnoTate), it tends to contain limited exchanges. Furthermore, Zooniverse projects typically lack robust social media activity, thereby limiting supplementary information availability. Conversely, Memory of Men stands apart due to the vigorous social media engagement of its volunteers, furnishing a rich data source complementing the details found within captures on the institutional platform in the Wayback Machine. Additionally, the substantial activity surrounding the project’s Twitter hashtag facilitates the identification of social networks connecting volunteers and, notably, linking volunteers with the institution. This connection is exemplified through the administration of the official project Twitter account (@MDHDefense), incidentally also serving as the webmaster of the institutional platform.

Lastly, concerning the focus of actions, a variety of materials shed light on the institutions’ motivations behind the three projects. These include the platform itself, official documents, a handful of scientific publications, and, in the cases of Transcribe Bentham and AnnoTate, respective project blogs. Furthermore, insights into the institutions’ empowerment strategies can be gleaned from readily available guidelines and tutorials accessible on the live web or within web archives. However, the picture

becomes less uniform when examining the motivations of the volunteers. Regrettably, for AnnoTate, pertinent information is non-existent, and the identification of potential interviewees poses a significant challenge due to the absence of interfaces that once listed participant names — interfaces that are no longer accessible, even in web archives. Conversely, in the case of Transcribe Bentham, personal pages within the Wiki emerge as invaluable resources for gaining deeper insights into participant motivations and for facilitating direct contact. Finally, Memory of Men presents an intermediate scenario. While the institutional platform provides scarce information about volunteers (featuring only a dozen portraits published during the project), the project’s social media accounts offer an ideal field for understanding contributors more comprehensively. Nonetheless, access to this information is subject to the familiar constraints imposed by commercial social media platforms (Zimmer & Hoffmann, 2014; Ruppert et al., 2013).

Amateur platforms

The selected amateur platforms are particularly noteworthy due to the diverse range of actors steering these projects. *Goodreads* exemplifies the plethora of commercial actors that offer a platform for users’ expressions while capitalizing on their data for profit. Established in 2006 and acquired by Amazon in 2013, Goodreads boasts over 90 million members worldwide, who have generated more than 50 million book reviews. In contrast, I.N.D.U.C.K.S. (International Network of Disney Universe Comic Knowers and Sources), also known as Inducks, stands as an openly accessible database aiming to catalog all Disney Universe comics

Table 21.1 Sources of information for studying institutional participatory platforms

	<i>Content</i>	<i>Actors’ organization</i>	<i>Actions’ focus</i>
Transcribe Bentham	Data collection freely available on the institutional platform, but no information about the participation; Less information about the transcription on Twitter	Rich information on wiki pages and their history; no information about social networks	Institutional documents allow to identify the motivations of the institution; amateurs’ motivations are expressed on the wiki personal pages
AnnoTate (on Zooniverse)	Data collection freely available on the institutional platform, but no information about the participation	Very less information on institutional sources	Very little information on institutional sources
Memory of Men	Data collection freely available on the institutional platform, but no information about the participation; Rich information about the transcriptions on Twitter (available in INA archive)	Rich information on institutional sources and Twitter exchanges; information about social networks in tweets (available in INA archive)	Institutional documents allow to identify the motivations of the institution; amateurs’ motivations are expressed in exchanges on Twitter (available in INA archive) and Facebook (only personal archive)

Participation, platforms, and cultural heritage

ever published globally (Breckon, 2015). Although initiated by amateurs in 2003, previous research reveals that the platform is not only fuelled by worldwide volunteers but also supported by professionals working for Disney who willingly forego copyright regulations on the comics (De Togni, 2022). Currently, Inducks offers comprehensive indexes for approximately 130,000 Disney comic releases worldwide, operating under this unique licensing arrangement. The third case is *Don't Lose Your Way*, an initiative carried out by an association. To address an institutional void, The Ramblers, the largest hiking association in the UK, launched a crowdsourcing project to accelerate the recognition of public right-of-way rights for all hiking trails across the territory, whether in current use or forgotten. What sets this platform apart is its evolutionary nature. In its initial phase, users could compare recent maps with older ones to identify hiking trails still present on the older maps and those that have vanished. After identifying forgotten trails, the second phase aimed to prioritize the identified trails. Lastly (this phase has not started yet), the project will focus on providing contributors with support in navigating the administrative procedures with local authorities to add trails to the institutional maps, also known as 'definitive maps' (The Ramblers, 2018; Claghorn, 2020).

In terms of content, the three platforms exhibit significant differences. Goodreads, in line with its commercial nature, does not intend to make crowdsourced content available for download or as open data. While an API permits quantitative access to platform content, it has faced limitations in recent years, and new API keys are no longer provided. User-generated content, when it is not made private by users, is accessible to everyone. Moreover, the Wayback Machine contains thousands of captures, allowing researchers to study the evolution of platform content over time. Ethnographic methods or live web scraping can be utilized to analyse this information (Thelwall & Kousha, 2017; Thelwall, 2019). Nonetheless, all rights to the content remain solely with the commercial actor. As for Inducks, similar to Transcribe Bentham, the Wiki system hosting the platform provides complete accessibility and traceability of the content. However, the history pages on Inducks are not accessible to common users, distinguishing it from Transcribe Bentham. One unique feature of Inducks pertains to the specific content license. While the content of the wiki is offered under an open license, volunteers do not possess the right to alter the legal status of the Disney comics. This restriction ensures that the legal ownership and copyright of Disney comics remain unchanged despite the open licence for other content on the platform. As stated on the website: "The information in the Inducks database is available under the conditions as mentioned in the Inducks licence. The pictures on this website are the property of Disney and are not part of Inducks." This legal violation is tolerated because the amateur-provided information about comics on this platform is valuable to the commercial actor (De Togni, 2022). Regarding The Ramblers, the goal is to provide all content as open data. However, information about contributions and contributors, as well as their evolution over time, tends to be lost, similar to what is observed with institutional platforms. The project's objective is data collection, with participation being merely functional toward this goal. When we are writing, the interface and data related to the priority attribution phase can be consulted, but the content related to previous participation remains inaccessible. Additionally, maps are only reachable for registered users, rendering them unavailable in web archives as well.

Regarding the organization of actors, the three platforms exemplify the range of possibilities that amateur platforms can show. In the case of Goodreads, the actors are at the core of the platform, with the platform functioning more as a social networking site for book readers

rather than a mere book review website, as highlighted by Thelwall and Kousha (2017). Goodreads offers various tools to facilitate digital identity and social network building among its users. However, retrieving information about actors' organizations over time is not straightforward. Web archives are the primary source of historical information, as the commercial platform has little incentive to provide such data. Yet, many user accounts are private, and given the platform's extensive size, web archives can only store a small fraction of its content. *Don't Lose Your Way* presents a somewhat contrasting scenario. The website does not provide explicit information about contributors and contributions, as data is collected without tracking individual authors. However, an intriguing aspect is the significant activity on social media, primarily on Twitter and Instagram. The association has created dedicated accounts and hashtags, motivating and tracking participation and participants outside the platform. Researchers seeking access to this information face the usual challenges associated with retrieving social media data, as mentioned earlier. Inducks falls somewhere in between the two extremes. While the platform's primary focus is on content, the wiki system offers various details about contributors, particularly their activity, such as the number of indexed publications, votes, posts, etc. Although there is no explicit information about social relations among users, such connections can be inferred from rating activity. Users have the option to rate comics, and the platform allows access to the list of individuals who have voted and commented on specific comics. Furthermore, the rules of contribution can be traced over time through the section "Bolderbast," which is meticulously archived by the Wayback Machine. This feature aids in understanding the platform's evolving rules and guidelines for contributors.

Regarding the focus of actions, on all commercial participatory platforms such as Goodreads, user motivations are readily visible and expressed through comments, lists, and interactions with platform content. These motivations serve the dual purpose of engaging existing users and attracting new ones. For Goodreads, researchers can access this information in the live web or web archives, provided it is not privatized by individual contributors. However, the strategy of Amazon is not explicitly expressed on the platform or the web in general, as the goal is to downplay the commercial nature of the platform and prioritize user expression.

In the case of Inducks and *Don't Lose Your Way*, their configurations are quite similar. On these platforms, the motivation behind their existence is central and clearly presented. The evolution of the platform's motivation can also be traced through web archives. User empowerment is an essential aspect, and both platforms provide tutorials and guides to support user participation. However, the platforms themselves do not offer a space for users to express their motivations. For *Don't Lose Your Way*, this information can be retrieved from social media. For Inducks, social media activity is limited, and the previous chat room (IRC Inducks) is inaccessible in archives. Instead, user motivations can be inferred from the comments left on the comics.

Discussion: Web archiving challenges

The proposed empirical analysis of institutional and amateur participatory platforms provides valuable insights into the transnational dynamics of digital participation in the cultural heritage field. It is particularly interesting to explore the challenges related to the historical examination of this phenomenon and the information sources that can support such studies.

Participation, platforms, and cultural heritage

Table 21.2 Sources of information for studying amateur participatory platforms

	<i>Content</i>	<i>Actors' organization</i>	<i>Actions' focus</i>
Goodreads	Commercially owned content, visible information collected in web archives	Rich information accessible in web archives (only public content)	Rich information about users' motivation in web archives; no information about platform motivation
I.N.D.U.C.K.S.	Content accessible through the wiki; complex copyright status	Rich information is stored in the wiki but no history available; it is accessible in web archives; few information about social network (rating)	Information in the wiki personal pages; no social media activity
Don't Lose Your Way	Content accessible as open data, no information about the contribution	Information in third sources and web archives; information on social networks in social media	Information on the platform and in social media

Upon initial examination, in alignment with the distinctions outlined in the existing literature, the analysis tends to emphasize different issues concerning web archiving on institutional platforms and amateur ones. For institutional platforms, the necessity of web archiving might appear secondary, given that these platforms primarily aim to provide extensive access to the collected data. As demonstrated in the three cases, organizations behind these platforms generally prioritize not only preserving the accumulated knowledge (structured in databases) but also ensuring its accessibility. Following the principles of the commons, institutional crowdsourcing platforms tend to promptly share the collected data within their digital spaces. Essentially, it can be asserted that institutional participatory platforms consider archiving as an integral part of their mission, incorporating web archiving systems within the platform itself. However, this archiving mission predominantly centres on the content (the data) and frequently overlooks information related to the participants. Conversely, for amateur platforms, particularly those based on commercial tools like social media, the situation becomes more intricate. These platforms encounter the well-known challenges associated with web archiving this type of source (Brügger, 2015; Helmond & van der Vlist, 2019). On the one hand, for commercial entities managing such platforms, the collected data is a primary value aimed at generating profit, making them disinclined to provide public access to their archives. On the other hand, information within these commercial spaces, particularly on social media, is usually accessible only through account creation and password protection. Consequently, capturing their content over time via public web archives is particularly complex (Vlassenroot et al., 2019; Vlassenroot et al., 2021). Certain platforms like Twitter offer APIs that allow public web archives or individual researchers to create specialized collections or micro-archives incorporating data from these platforms, although this comes with significant technical limitations and costs (Littman et al., 2018; Espley et al., 2014; Pehlivan et al., 2021; Acker & Kreisberg, 2020). Nevertheless, an advantage of amateur platforms compared to institutional ones is that they often provide more information useful for historical analysis of contributors and contributions, including profile pages, social networks, and content interactions.

Upon closer examination, this distinction reveals some evidence, but it also uncovers transversal issues that warrant exploration. Our argument is structured around three levels: theoretical, technical, and ethical. First, the historical analysis of participatory platforms raises a significant theoretical question pertaining to the differentiation between amateur and institutional platforms. This question has two facets. On the one hand, empirical analysis demonstrates that this distinction, primarily grounded in the project's piloting actor, isn't particularly relevant for web archiving. A comparison between Transcribe Bentham and Inducks unveils that these two platforms, one institutional and one amateur, share a remarkably similar approach to web archiving. In fact, the wiki infrastructure on which they are built is highly suitable for storing a wide range of information related to the participatory platform, covering aspects such as content, contributors, and contributions. Conversely, the case of Don't Lose Your Way illustrates that an amateur platform may provide very limited information about participation. On the other hand, analysed case studies highlight that institutional and amateur platforms can be closely intertwined. Participatory projects not only can be based on different types of digital spaces, combining both institutional and amateur solutions, but contributors interested in specific cultural objects can participate across several platforms based on the content they wish to share. For instance, in the Transcribe Bentham project, digital participation mainly occurs on the institutional crowdsourcing wiki. However, interactions among contributors also extend to social media. Similarly, in the case of Memory of Men, a significant amount of information related to the "Dead for France" project circulates on Twitter and Facebook. This information pertains not only to the organization of actors but also to the content of the crowdsourcing campaign itself. This multiplicity of digital spaces is also evident in the case of amateur platforms like Don't Lose Your Way, where the importance of social media accounts tied to the project is observed. Therefore, what do web archiving participatory platforms and digital participation in cultural heritage more broadly, entail? Are the annotated data corpora accessible on institutional websites sufficient to safeguard the project and the related digital participation? Should the activities of amateur platforms also be archived? And if so, should archives of amateur and institutional platforms be interconnected?

This leads us to the second level of transversality, which concerns the technical challenges linked to web archiving digital participation. Revising the theoretical approach by discarding the distinction between amateur and institutional platforms necessitates finding methods to support the historical exploration of the amateur-institutional relationship across diverse digital spaces. This presents three types of technical challenges. Firstly, while information about crowdsourced content is generally accessible in the live web and web archives, data about participation and participants are often confined to private sections of websites, rendering them inaccessible to both researchers and web archivists. Secondly, social media serve as crucial spaces for interactions among participants (amateur-institution, amateur-amateur), yet gaining access to their historical data presents significant technical and financial challenges. Thirdly, even when data can be gathered from diverse digital spaces, assembling them into a unified database that allows tracking the flow of information and users across these spaces is frequently near-impossible. Content often lacks metadata ensuring interoperability, while users may have varying pseudonyms or even act anonymously, making it challenging to trace them across spaces. Hence, how can we investigate the historical evolution of digital participation given these technical constraints? How can web archives facilitate such analyses?

Participation, platforms, and cultural heritage

This brings us to the final transversal level, the ethical aspect. In earlier research (Severo, 2021), we demonstrated that the only viable method for analysing digital participation through a comprehensive approach that considers interactions among spaces is to focus on the circulation of actors rather than content. In other words, it is necessary to construct a complex methodology using a bricolage assemblage of techniques to combine data from various web archives: public web archives, collections curated by researchers, or data collections made available by participants of the studied platform. However, this approach raises several ethical concerns. It not only renders any form of anonymization unfeasible but also allows the researcher to potentially infer a unique identity for users who lead entirely separate digital lives across different online spaces. The main implication of this is that all participants must provide informed consent to participate in the research. However, is it technically possible to contact such users? And if so, is this sufficient? In this chapter, space constraints prevent a thorough exploration of the extensive literature on the ethics of citizen science and participatory research in the cultural heritage realm (Kidd & Cardiff, 2017; Purdam, 2014), as well as the ethical considerations tied to web archives (Lomborg, 2018). Nevertheless, it is important to acknowledge this crucial aspect that researchers working with diverse web archives must grapple with, along with the potential impact this practice might have on website owners themselves (Hegarty, 2023).

Conclusion

In today's context, digital participation in relation to cultural heritage is a phenomenon that spans across national boundaries. This chapter has demonstrated the intricacies involved in conducting a historical analysis of this phenomenon. Drawing from a sample of six case studies, we have challenged the conventional distinction between amateur and institutional platforms as outlined in the existing literature on cultural participatory platforms. Furthermore, we have explored the benefits of integrating information from various web archives. This approach allows us to delve into not only the collected data on participatory platforms but also the dynamics of participation that facilitate the data collection process. On a surface level, it might be concluded that institutional platforms are generally subject to web archiving, while amateur platforms are not. Activities and data on the former are preserved and accessible, while interactions and data on the latter face the risk of being lost or at least inaccessible to external users. However, upon closer examination, it becomes evident that institutional and amateur platforms give rise to several interconnected issues.

When investigating digital participation, we advocate moving away from the notion of a platform as a discrete digital environment separated from others. Instead, we suggest analysing participatory platforms as ecosystems that bring together actors dispersed across numerous digital spaces, encompassing both amateur and institutional platforms. This collaborative effort among various actors contributes to the construction of shared knowledge concerning cultural heritage objects. This shift in perspective enables a deeper exploration of the relationships formed among social actors (such as citizens, institutions, and commercial entities) across these platforms. This approach also enables the observation of agency changes as actors shift from one space to another, combining information from institutional web archives and collections of commercial data (Akrich, 2010). The convergence of institutional and amateur perspectives occurs in diverse ways and serves different purposes depending on the specific space in which it unfolds. Consequently, it is crucial to underscore the technical

and ethical challenges brought about by this holistic approach. Researchers delving into this phenomenon must analyse various web archives — public, personal, commercial, etc. — to gather pertinent information. Additionally, public web archives themselves face the challenge of collecting information not only related to platform content but also concerning the organization of actors and the motivation of contributors' actions.

In conclusion, it is pertinent to address the ethical dilemmas that this comprehensive approach raises and to propose potential strategies to navigate them. The study of digital participation involves an exploration of participants and their trajectories across different digital spaces. As detailed in the previous section, to trace participants across spaces, it is imperative not to anonymize web-archived information, allowing for the observation of contributors' behaviors as they traverse from one space to another. In response, we propose that adopting this transversal perspective not only as an analytical method but also as a design philosophy for participatory platforms could be a viable solution. For institutions and researchers intending to establish such platforms for scholarly purposes, it becomes crucial to acknowledge and embrace the multi-space nature of these projects from the outset. Designers should contemplate constructing participatory platforms as multi-space environments, blending institutional crowdsourcing tools with more flexible expressive spaces like social media or wikis. This approach could also influence the archiving policies of these diverse digital spaces right from the inception. The Transcribe Bentham Wiki exemplifies this approach by offering a transparent and traceable structure, allowing for the retention of extensive information about digital participation over time. Such an approach caters to the varied needs and motivations of contributors, fostering active engagement across different digital spaces while mitigating the risk of dispersion into less accessible environments for researchers. Although empirical testing of this hypothesis falls beyond the scope of this chapter, we present it as a call for future research exploration.

Notes

- 1 Website of the *Observatory of the Cultural Platform*, <https://web.archive.org/web/20240318024446/https://anr-collabora.parisnanterre.fr/observatoire/>.
- 2 Transcribe Bentham platform, https://web.archive.org/web/20240318024739/http://transcribe-bentham.ucl.ac.uk/td/Transcribe_Bentham.
- 3 Annotate project on the Zooniverse platform, <https://web.archive.org/web/20240318025914/www.zooniverse.org/projects/drrogg/annotate>.
- 4 Platform *Mémoire des hommes*, managed by the French Ministry of Army, <https://web.archive.org/web/20240318030326/www.memoiredeshommes.sga.defense.gouv.fr/>.
- 5 Platform *Goodreads*, <https://web.archive.org/web/20240318030711/www.goodreads.com/>.
- 6 Platform *InDucks*, <https://web.archive.org/web/20240318030922/https://inducks.org/>.
- 7 Platform *Don't lose your way* by Ramblers association, <https://web.archive.org/web/20240318031024/https://dontloseyourway.ramblers.org.uk/>.
- 8 <https://web.archive.org/web/20240318031150/https://anno.tate.org.uk/>. Website of the project *Annotate* managed by the Tate.

References

- Acker, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science*, 20, 105–123.
- Aikawa, N. (2004). An historical overview of the preparation of the UNESCO International Convention for the Safeguarding of the Intangible Cultural Heritage. *Museum International*, 56(1–2), 137–149.

Participation, platforms, and cultural heritage

- Akrich, M. (2010). From communities of practice to epistemic communities: Health mobilizations on the internet. *Sociological Research Online*, 15(2), 116–132.
- Barber, S. T. (2018). The ZOONIVERSE is expanding: Crowdsourced solutions to the hidden collections problem and the rise of the revolutionary cataloging interface. *Journal of Library Metadata*, 18(2), 85–111.
- Ben-David, A., & Huurdeman, H. (2014). Web archive search as research: Methodological and theoretical implications. *Alexandria*, 25(1–2), 93–111.
- Bonacchi C., Bevan A., Keinan-Schoonbaert A., Pett D., & Wexler J. (2019). Participation in heritage crowdsourcing. *Museum Management and Curatorship*, 34(2), 166–182.
- Bortolotto, C. (2007). From objects to processes: UNESCO's intangible cultural heritage. *Journal of Museum Ethnography*, 19, 21–33.
- Boyd, D., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210–230.
- Breckon, M. (2015). INDUCKS: Worldwide database about Disney comics. *Reference Reviews*, 29(3), 43–44.
- Brügger, N. (2013). Web historiography and Internet studies: Challenges and perspectives. *New Media & Society*, 15(5), 752–764.
- Brügger, N. (2015). A brief history of Facebook as a media text: The development of an empty structure. *First Monday*, 20(5), n.p.
- Brügger, N. (2018). *The archived web: Doing history in the digital age*. MIT Press.
- Causser, T., & Terras, M. (2014). Crowdsourcing Bentham: Beyond the traditional boundaries of academic history. *International Journal of Humanities and Arts Computing*, 8(1), 46–64.
- Causser, T., & Wallace, V. (2012). Building a volunteer community: Results and findings from Transcribe Bentham. *Digital Humanities Quarterly*, 6(2), n.p.
- Causser, T., & Wallace, V. (2018). 'Making such bargain': Transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription. *Digital Scholarship in the Humanities*, 33(3), 467–487.
- Claghorn, J. (2020). Using spatial network analysis to recover England's and Wales' lost footpaths and rights of way. *Journal of Digital Landscape Architecture*. Digital Landscape Architecture Conference, 03–04 June 2020, Harvard University, Cambridge, MA, USA. Wichmann Verlag/VDE VERLAG GMBH, 23–32.
- Couldry, N., & Mejias, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4), 336–349.
- De Togni, I. (2022). The Inducks index, editorialized by Disney comics amateurs and professionals: New dynamics, new participatory models. *Hybrid. Revue des arts et médiations humaines*, 8.
- Espley, S., Carpentier, F., Pop, R., & Medjkoune, L. (2014). Collect, preserve, access: Applying the governing principles of the national archives UK government web archive to social media content. *Alexandria*, 25(1–2), 31–50.
- Giaccardi, E. (Ed.). (2012). *Heritage and social media: Understanding heritage in a participatory culture*. Routledge.
- Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, 12(3), 347–364.
- Haklay, M. (2015). *Citizen science and policy: A European perspective*. The Woodrow Wilson Center/ Commons Lab.
- Hegarty, K. (2023). Imagining permanence on the web: Tracing the meanings of long-term preservation among the subjects of web archives. *New Media & Society*, <https://doi.org/10.1177/146144482311870>.
- Helmond, A., & van der Vlist, F. N. (2019). Social media and platform historiography: Challenges and opportunities. *TMG — Journal for media history*, 22(1).
- Hennion, A. (2001). Music lovers: Taste as performance. *Theory, Culture & Society*, 18(5), 1–22.
- Iranowska, J. (2019). Greater good, empowerment and democratization? Affordances of the crowdsourcing transcription projects. *Museum and Society*, 17(2), 210–228.
- Jenkins, H. (2006). *Fans, bloggers, and gamers: Exploring participatory culture*. NYU Press.
- Kenney, M., & Zysman, J. (2016). The rise of the platform economy. *Issues in Science and Technology*, 32(3), 61–69.
- Kidd, J. & Cardiff, R. (2017). 'A space of negotiation': Visitor generated content and ethics at Tate. *Museum and Society*, 15(1), 43–55.
- Kozinets, R. V. (2010). *Netnography: Doing ethnographic research online*. SAGE Publications.

The Routledge Companion to Transnational Web Archive Studies

- Leadbeater, C., & Millett, P. (2004). *The Pro-Am revolution. How enthusiasts are changing our economy and society*. Demos.
- Littman, J., Chudnov, D., Kerchner, D., et al. (2018). API-based social media collecting as a form of web archiving. *International Journal of Digital Libraries*, 19, 21–38.
- Lomborg, S. (2018). Ethical considerations for web archives and web history research. In N. Brügger & I. Milligan (Eds.), *The SAGE handbook of web history* (pp. 199–219). SAGE.
- Papacharissi, Z. (2010). *A networked self: Identity, community, and culture on social network sites*. Routledge.
- Pehlivan, Z., Thièvre, J., & Dugeon, T. (2021). Archiving social media: The case of Twitter. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The past web*. Springer, Cham.
- Purdam, K. (2014). Citizen social science and citizen data? Methodological and ethical challenges for social research. *Current Sociology*, 62(3), 374–392.
- Ridge, M. M. (Ed.). (2014). *Crowdsourcing our cultural heritage*. London, Ashgate Publishing, Ltd..
- Rogers, R. (2019). *Doing digital methods*. SAGE.
- Ruppert, E., Law, J., & Savage, M. (2013). Reassembling social science methods: The challenge of digital devices. *Theory, Culture & Society*, 30(4), 22–46.
- Severo, M. (2021). *L'impératif participatif: institutions culturelles, amateurs et plateformes*. Éditions de l'INA.
- Severo, M. (2022). Collaborative platforms: Cultural heritage and participation. *Public History Weekly*, 10(1).
- Tasker, G., & Liew, C. L. (2020). 'Sharing my stories': Genealogists and participatory heritage. *Information, Communication & Society*, 23(3), 389–406.
- Tate. (2015). Archives & access project: AnnoTate — crowdsourcing to transcribe the archives. <https://web.archive.org/web/20221225190829/www.tate.org.uk/about-us/projects/transforming-tate-britain-archives-access/archives-access-project-annotate>.
- The Ramblers. (2018). Don't Lose Your Way: The guide to finding lost rights of way. <https://web.archive.org/web/20240318031441/https://wroughton.gov.uk/wp-content/uploads/2019/11/Rights-Of-Way-Guide.pdf>.
- Thelwall, M. (2019). Reader and author gender and genre in Goodreads. *Journal of Librarianship and Information Science*, 51(2), 403–430.
- Thelwall, M., & Kousha, K. (2017). Goodreads: A social network site for book readers. *Journal of the Association for Information Science and Technology*, 68(4), 972–983.
- Van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press.
- Vlassenroot, E., Chambers, S., Di Pretoro, E., et al. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, 1, 85–111.
- Vlassenroot, E., Chambers, S., Lieber, S., et al. (2021). Web-archiving and social media: An exploratory analysis. *International Journal of Digital Humanities*, 2, 107–128.
- Zimmer, M., & Hoffmann, A. L. (2014). Large-scale networked data analysis in social media research: The case of Twitter. *International Journal of Communication*, 8(21).

22

BUILDING AN ARCHIVE OF HISTORICAL WEB DEFACEMENTS

Michael Kurzmeier

Introduction

Web defacements are unauthorised alterations or overwrites of existing web pages. While web pages might be defaced for a number of reasons ranging from cyber war campaigns to competitive hacking, web pages are also defaced for the expression and propagation of political content in the widest sense. These political web defacements, especially their preservation and accessibility for research, are the topics of this chapter. This chapter describes the creation of a web defacement archive using published data from one of the largest community-maintained cybercrime archives.

The mentioned community archive, attrition.org, collected material from 1995 until 2001. While the website remained online, and while research access was granted to the author, no new submissions were processed. In March 2021, a complete copy of the archived defacements was made available on GitHub.

Regarding preservation, all types of web defacements are at a double disadvantage. On a technical level, defaced web pages are unwanted alterations that are likely to be quickly deleted or restored to their original versions. It is therefore unlikely that web archiving crawls would archive defaced pages while they are live. Where they do, collection scope and guidelines might again lead to the removal of defaced content from the collection. Especially with national collections, a defaced page might be outside the collection scope. Yet, the inverse might also be true and a website might move into scope through being defaced by a group speaking to a specific national issue. For example, a defaced UK website with a statement regarding a specific Danish issue might be in focus for Danish web archiving where prior to the defacement it was not. Additionally, web defacements are more likely to be seen as unwanted intrusions and nuisances than artefacts worth preserving. This is why, despite their significance, political web defacements remain an under-utilised resource for research on both the history of activism on the web¹ and subaltern political communication.² From a theoretical perspective, defacements are descriptions of technical objects (web pages) by the defacer who takes up agency to interact with the object in a way not originally envisioned by the designer of the technical object:

users may define quite different roles of their own. If this happens, the objects remain a chimera, for it is in the confrontation between technical objects and their users that the latter are rendered real or unreal.

(Akrich, 1992, p. 208)

This conflict between user roles qualifies web defacements as a site of struggle between technical objects and user-defined custom roles. It also helps to explain the inherent *in-between-ness* of web defacements in terms of archiving and preservation efforts.

Speaking on general web archiving considerations, Niels Brügger states that edits and deletions are common even among more *legitimate* web pages:

Web content may have been changed, moved, or deleted for various reasons, and sometimes in the hope of not being found out.

(Brügger, 2019)

This is most certainly true for web defacements, which are usually deleted and the original page restored as quickly as possible.³ While understandable from the perspective of the admin or domain owner, these defacements nevertheless form a part of web history. Web defacements may be seen as an avant-garde form of engagement with web pages, a *cracking* of existing media forms (Kelly, 2009, p. 286) to find novel ways of interaction. To preserve that history, and to make web defacements available for research, special archival work is necessary. This issue hints at questions of usability and contextualisation of defaced web pages, which will be explored in more detail in the coming sections.

Because of the mentioned disadvantages in preservation, web defacements as a form of hacktivism are rarely archived and thus mostly lost for systematic study. Besides large collections such as the Internet Archive, which might pick up a few hacked pages during a crawl, there also exists a small scene of community-maintained cybercrime archives that archive hacked websites, some of which are hacked in a hacktivist context and may involve web defacements; however, most of these community archives have since ceased operations. Web defacements and the community surrounding them may also be studied through interviews with defacers. See, for example, Hopkins (2014) and Samuel (2014) who conducted interviews with defacers on artistic and technical practices. While community archives exist, no comprehensive source is available to serve academic research. Community archives usually do not feature metadata, have unknown collection practices, and hardly ever focus exclusively on political defacements. As many community archives have themselves disappeared from the web, in the context of creating a more complete historical record (Milligan, 2016), filling the described gaps in web archiving is important.

Additionally, research methods for web defacements are anything but standardised. The community archives Zone-H, used in the described projects by Hopkins and Samuel holds copies of upwards of 10.000 defaced domains. To make this collection usable for research, methods to categorise and engage with the data have to accompany the data itself. A recent study found that lack of metadata is one of the key barriers for scholars when engaging with archived web material (Healy et al., 2022). As in many web archiving efforts, the problems arising from a large collection of unrelated, unstructured data are problems of identifying relevant material within the corpus and problems of making adequate and data-based statements about themes and sentiments within that corpus (Milligan, 2022).

Building an archive of historical web defacements

Considering the transnational dimension of this research, the preservation and study of web defacements present immense opportunities to explore the interconnected nature of hacktivism across borders. Defaced websites, serving as platforms for expressing political messages with transnational relevance, underscore the importance of adopting a global perspective in archival practices. Defacements that transcend national boundaries, such as a UK website defaced to address a specific Danish issue, highlight the need for international collaboration in web archiving efforts. Understanding the transnational dynamics of web defacements not only enriches the historical narrative but also underscores the importance of a coordinated approach in preserving and studying these digital artefacts.

The second disadvantage in regard to the preservation of defacement is their interconnectedness with the original page at the time of defacement. Political web defacements often have a connection to the original web domain. This relation might be superficial, such as when web defacements target pages by the top-level domain (TLD; a personal website on a national domain is defaced to convey a message regarding that country's foreign policy). Going through many intermediate layers, this relationship might also be very close, such as when a website is defaced and replaced by a parody that mimics the look and content of the original page. This relationship means that for the full analysis of political defacements, two archived copies are necessary: The original page at the time of defacement, and the defacement itself. Only then can the full relationship between the original page and the defacement be fully understood.

These two disadvantages lead to the necessity to archive and present web defacements in such a way that research ideally has access to both versions and can compare them side-by-side. This requires a combination of sources, combining an archive of defaced web pages with a larger, established, web archive. The following section will outline the workflow used to convert the data published by attrition.org into an archive that serves research needs, together with an overview of analysis models.

Workflow

This workflow starts with the download of the published archive and ends with the presentation of the SolrWayback⁴ instance, together with an interactive network graph for analysis. The workflow has been described in detail (Kurzmeier, 2023) and will only briefly be presented here.

As an overview, the workflow was designed with the following objectives in mind:

- Stabilise the data from individual files to WARC files
- Create metadata to enrich the WARC files
- Contextualise, where possible, through providing a copy of the original page
- Create a stable archive of web defacements
- Create tools for the analysis of this unusual data

The individual steps reflect these objectives.

Data source

The data source used for this project is the defacement mirror provided by attrition.org. The site has been collecting defacement from 1999 until 2001 and in March 2021 has published

its archive and some of its archiving tools on GitHub (Attrition.org Web Page Defacement Mirror, n.d.). Regarding the history of the collection, the page reads:

In January, 1999, the attrition.org staff at the time decided to create a mirror for defaced web pages. By early June, we had 1300 defacements mirrored. Some of those from before January were copied from other defacement mirrors that ran before us. [...] The mirror ran for almost two and a half years and ultimately cataloged 15,2013 defacements including 63 compromised NASA servers and 19 police department pages. [...]

(Attrition.org Web Page Defacement Mirror, n.d.)

In March 2021, the collection of defaced pages was released to the public. What makes this collection of defaced pages especially suited for building a web archive of defacements is the age of the pages, with the oldest defacements dating back to 1995, and the accessibility of material under a CC-BY-NC-SA license.⁵

Web defacements are generally understood to be unauthorised alterations to the appearance of a website (Maggi et al., 2018; Samuel, 2014). This alteration must be noticeable to the user. In contrast to a phishing attempt, a web defacement is only successful when noticed by the user. Regarding collection scope and practice, the attrition.org defacement archive simply states that:

A web defacement happens when the content of a public web page is altered by someone other than the legitimate person or company responsible for the content. [...] At the time, these were sometimes significant events carried out as a form of digital protest or “hactivism”. Other times they were simple “digital vandalism”, essentially graffiti that was a nuisance at best.

(Attrition.org Web Page Defacement Mirror, n.d.)

While this technical definition of web defacements acknowledges political motives as one potential motivation for defacers, it does not specify how attrition acquired learned about defaced websites, nor how the team decided which pages to include and which to ignore.

In contrast to attrition’s own definition as a collection of hacktivist material, Woo et al. classified only about 30% of defacements as political in nature (2004, p. 71). A survey of Zone-H, another defacement archive, revealed that dedicated political motivation accounts for no more than 10% of defacements (Romagna & van den Hout, 2017b). The authors further acknowledge that a general shift in strategy has taken place amongst defacers:

Hacktivism now seem more likely to deface any websites with poor security measures (no matter the topic of the site) in order to spread a message. It therefore appears that the target per se is no longer so important, while the visibility (calculated on the number of defaced websites) represents the main element.

(Romagna & van den Hout, 2017a)

Having spent some time working with the material, and based on NLP-based clustering of defaced pages (Kurzmeier, 2023), it can be said that of all defacement mirrors encountered

Building an archive of historical web defacements

during the research, attrition.org was mostly focused on political defacements in the widest sense.

What is entirely absent from the provided data, however, is any kind of descriptive meta-data. The files provided via GitHub structure defacements are as follows:

{Year}/{Month}/{Day}/{URL}/{individual files}

As described by Ruest et al. (2022), derivative datasets together with a robust finding aid are of high value for web archive research. Step one of the workflow therefore includes the creation of rudimentary metadata, which is almost entirely absent in the attrition download.

Data processing

The goal of this workflow step is to produce WARC files from the downloaded collection of local files, and enrich these files with relevant metadata including time and date of defacement, and original URL. In the first step, this metadata can be extracted from the folder structure described above. In the second step, using WARCit, the local files are converted into WARC files and the extracted metadata is added to the WARC file header.

As Table 22.1 shows, WARC-Date is set to the defacement data as it was extracted from the folder structure (for example, if the folder structure was 2000/10/05/example.com, 2000-10-05 was extracted as the date). The WARC-Source-URI points to the local files the WARC was created from, while the WARC-Target-URI refers to the extracted domain — in this case, example.com.

While the files are being processed, both internal and external links are being extracted and stored in a DataFrame together with the other metadata. This data will be used later to create a visualisation of the external links defacers use in their defacements. For the purpose of creating a web archive of defaced pages, the WAR files with added metadata are an ideal starting point.

Indexing

The goal of this workflow step is to import the created WARC files into a SolrWayback instance for search, analysis and playback. This process takes some time, given that over 15,000 WARC files are created, but is achievable on everyday hardware. Once the files have

Table 22.1 Sample WARC header with reconstructed metadata

WARC/1.0
WARC-Date: 2000-10-05T00:00:00Z
WARC-Creation-Date: 2017-12-05T18:30:58Z
WARC-Source-URI: file:///./path/to/somefile.html
WARC-Type: resource
WARC-Record-ID: ...
WARC-Target-URI: www.example.com/to/somefile.html
Content-Type: text/html
Content-Length ...

been indexed, the server installation can be migrated to a low-cost web server as will be described later on.

Regarding the use of SolrWayback to create a web defacement archive, it must be stated that SolrWayback, for all its usefulness and impressive out-of-the-box functionality is not a software designed for replay of web defacements. This means that, while some of its functions such as content search and domain graphs are very useful, others such as the visualisations of content over time are less useful due to the relatively short time span covered by the attrition archive. Other features, which are very specific to the research on defaced websites, such as external link targets and side-by-side comparisons of the original and the defaced page, are obviously not part of SolrWayback. In addition to these points, web defacements have unique features that are not accounted for by SolrWayback or any other web archive browsing application, such as attribution of attacks for defacer groups or individuals.

Analysis

After the files have been indexed, and using the extracted metadata, a preliminary analysis of the corpus allows us to gain insight into the archived defacements.

Figure 22.1 shows the number of defaced domains for each month from August 1995 to July 2001. Months without defaced domains are omitted. Figure 22.1 confirms the narrative by the attrition.org team, which is that of a volunteer team being overwhelmed by the growing workload. Looking at the timescale, the pre-1999 material (which was described as being incorporated from an unnamed older archive) makes up a small fraction of the overall content. The years 2000 and 2001 clearly dominate the corpus of archived defacements by volume. A sharp drop after June 2001 indicates the end of archiving activity by attrition.org. Overall, the historical record of web defacements is fragmented, and low figures for any given period of time cannot be understood as sole evidence of low defacement activity.

That attrition eventually stopped archiving defacements is in line with a general change in the defacement scene that happened during the early 2000s. A 2014 dissertation on defacers found most defacement mirrors either offline or no longer accepting submissions (Samuel, 2014), while a 2018 study found that web defacements have lost some of their political appeal (Maggi et al., 2018). The reasons for this decline in archiving activity — not in overall defacement activity — are unclear. Possible explanations include volunteer teams being increasingly overwhelmed with submissions, and a general platformisation of the web from the mid-2000s on. In contrast, the largest still operating defacement archive, currently holding upwards of three million defacements, has turned towards a paid access model for researchers (Maggi et al., 2018).

It is important at this point to be aware of how uneven the temporal distribution of material is throughout the archive. As web defacements tend to follow geopolitical events (Balduzzi et al., 2018), the geopolitical events of 1999 until May 2001 will be extensively covered by the defacements in the attrition archive. At the same time, having access to pre-2000 defacements can offer insight into a time period hardly covered by institutional web archives (Weber, 2022).

Another useful way to visualise the collection is through a link network. However, for this special content, a custom approach is necessary. Generally, the use of hyperlink network analysis for the analysis of large datasets has been described by Waldherr et al. (2017). In relation to political ideology, the application of hyperlink analysis has been described by de Bakker

Building an archive of historical web defacements

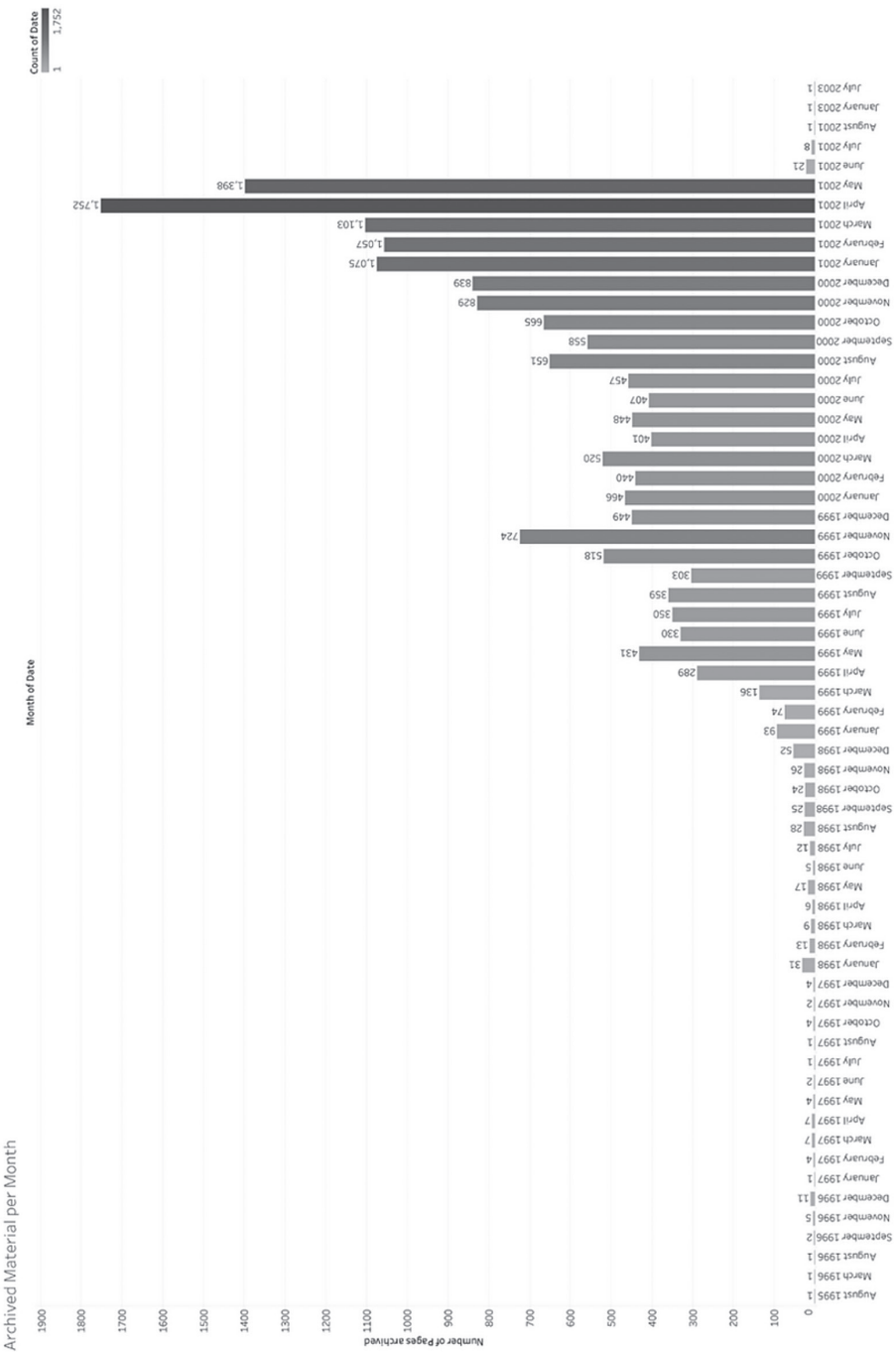


Figure 22.1 Distribution of archived material in the attrition archive per month.

and Hellsten (2013). Relating explicitly to archived web content, Brügger (2013) describes the use of link network graphs as an appropriate tool.

However, though the above refers partially to datasets in general and partially to general web archives, the situation is different in regard to defaced web pages. As described in the introduction, defacements are usually ephemeral web content that is quickly restored or removed. As such, defacers usually do not link to other defaced pages as these are not stable web entities. A link network of archived content does not yield very much useful information, as most defacements stand for and by themselves and do not link to other defaced pages. Mapping the archived pages by domain is also not a productive approach, as generally defacements would only affect the index.html. Where the original site was not deleted, but renamed (a common practice amongst defacers during the time), the attrition crawler would capture many HTML documents belonging to the original page, not the defaced page.

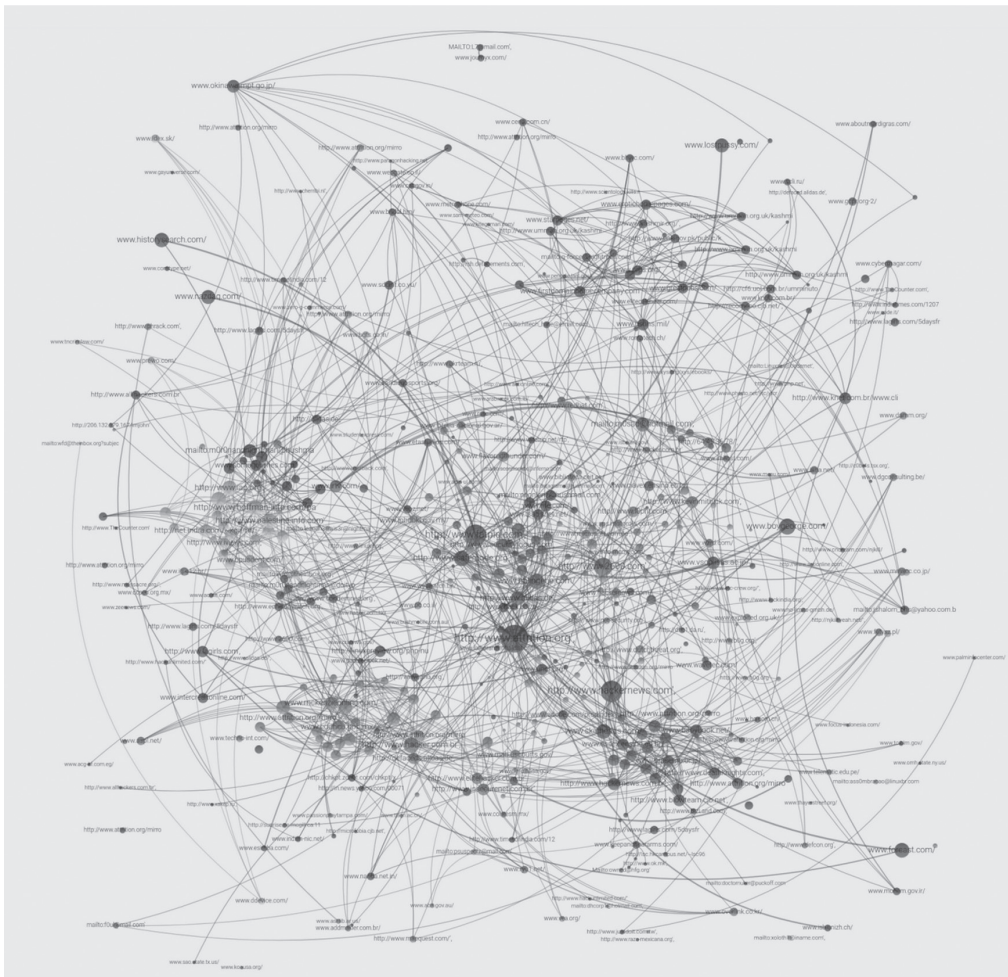


Figure 22.2 Directed network of link targets used in defacements, with a minimum of five incoming links, available on defacementarchive.com.

Building an archive of historical web defacements



Figure 22.3 Email address used by a defacer.

As has been described in the data processing section, during the initial file processing, all external and internal links have been extracted from the defaced pages. Using this data it is possible to build a link network of *link targets* used by defacers. The assumption is that defacers, in communicating political messages, will often refer to the same external sites.

This network visualisation of Figure 22.2 allows users to identify link targets used in defacements, and through those link targets, draw conclusions about topics defacers talk about. For example, the cluster on the middle left is mostly composed of links relating to Palestine, likely in reaction to the 2000 Camp David Proposals (Camp David Proposals for Final Palestine–Israel Peace Settlement, n.d.) and Second Palestinian Intifada (Al-Aqsa Intifada Timeline, 2004). Other link targets hint at scene-internal communication, such as www.hackernews.com or defacers' personal websites, where they offer further information about their work (and occasionally link back to defacement archives to show off past defacements).⁶ In other cases, the external link targets reveal authorship through email addresses left by defacers as can be seen in Figure 22.3.

Where email addresses are used in defacements, it is commonly done in conjunction with an invitation to either the general public or the site administrator to contact the defacer to learn more about their work. Political defacers would occasionally offer their help in restoring the original site and in preventing further attacks,⁷ though the sincerity of these offers remains unknown.

Creating this link network of external sites helps the analysis of political web defacements in two ways. First, it provides an overview of potential topics found in this corpus of defaced web pages. For research use, this allows us to gain insight into the defacer's political messages, and internal and external communication, and to attribute authorship of defacements. Regarding its integration with SolrWayback, this graph complements the existing tools provided. Using the full-text search, all defacements referring to one specific external site can be found. It was said earlier that SolrWayback is not a tool specifically for

the replay of archived defacements, and that its design does not in all cases meet the research requirements for this special material. Using a link graph such as the one described here, however, can augment SolrWayback's functionality and provide a valuable finding aid. The published version of the defacement archive can then feature the link graph and full-text search to connect an external link to all local pages that refer to it.

The described network of external link targets provides a finding aid for SolrWayback's full-text search, as SolrWayback allows queries for both displayed text in archived web pages and HTML tags in HTML documents. This means that even though the link targets are not part of the defacement archive, they can be used to find pages that refer to them. In Figure 22.4, the email address from Figure 22.3 is used to find all pages referring to it.

The latest SolrWayback release (4.4.1. at the time of writing) allows users to define a custom time frame to run queries against. This is a welcome addition for any archive that holds older content and is used in the above figure to cover the relevant time frame. The timescale can be set to months, weeks, days, or years which allows for fine-tuned queries. In the above example, the query returns 13 total occurrences of this email address between December 2000 and May 2001. Clicking on any of the data points opens a list view of results and allows further filtering of results by domain and file type.

The export options shown in Figure 22.5 are another example of how the generic functions of SolrWayback can be used to complement the specific tools provided for the analysis of defacements. Results obtained through the external link network analysis can be exported into derived datasets for further analysis, e.g. all defacements that feature the email address from Figure 22.4. In this way, it is possible to export datasets focussed on geopolitical events.

To provide an example of these case studies of geopolitical events, literature has confirmed that web defacers often act in response to geopolitical events (Balduzzi et al., 2018; Denning, 1999; Samuel, 2014). Based on the network analysis, and previous into hacktivism as a means of political expression, the keywords "Kashmir", "India" and "Pakistan" are used in SolrWayback's n-gram tool.⁸

The keywords used relate to the Kashmir conflict. The Kashmir conflict is a territorial dispute between India, Pakistan, and China over the Kashmir region. This conflict arose from the 1947 partition of India and has, with varying intensity, continued until today, leaving its marks in the cultural history of Pakistan and India. Aside from military confrontations, the conflict has also led to an ongoing cyber war between the two countries, a part of which features web defacements. Reflecting on the value of event-based web archives (Rogers, 2019), the availability of material relating to the Kashmir conflict offers the opportunity to complement the historical record of the accompanying cyberwar.

The results of the query in Figure 22.6 ample findings of the three keywords throughout the archive, starting in 1998. While the term "Kashmir" is used from July 1999 onwards, mentions of India and Pakistan can also be found in older defacements. The first peak corresponds to the Kargil War of 1999 (Lavoy, 2009), supporting the thesis that political web defacements tend to follow geopolitical events.

Web defacements operate on more than just the level of content visible on the screen. In appearing unexpectedly, they are at the same time proof of the vulnerability of infrastructure which is in constant need of repair and maintenance. Generally, defacement targets may be chosen based on the technical challenge presented (easy or hard to compromise the target system) and the target's relationship to the message. Where hackers choose targets based on the technical challenge, so-called mass defacements of outdated servers are often the result.⁹

Building an archive of historical web defacements

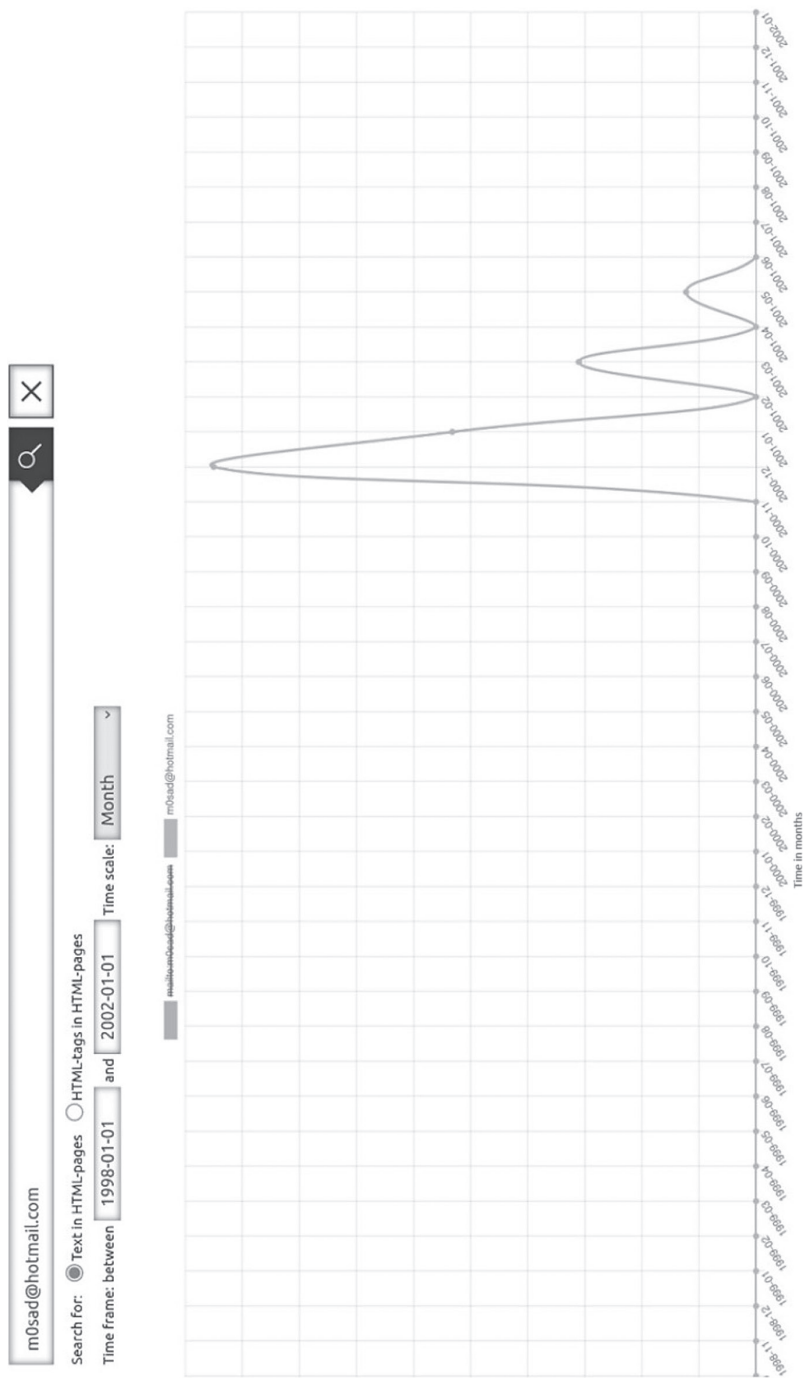


Figure 22.4 Search results for a defacer's email address.

The Routledge Companion to Transnational Web Archive Studies

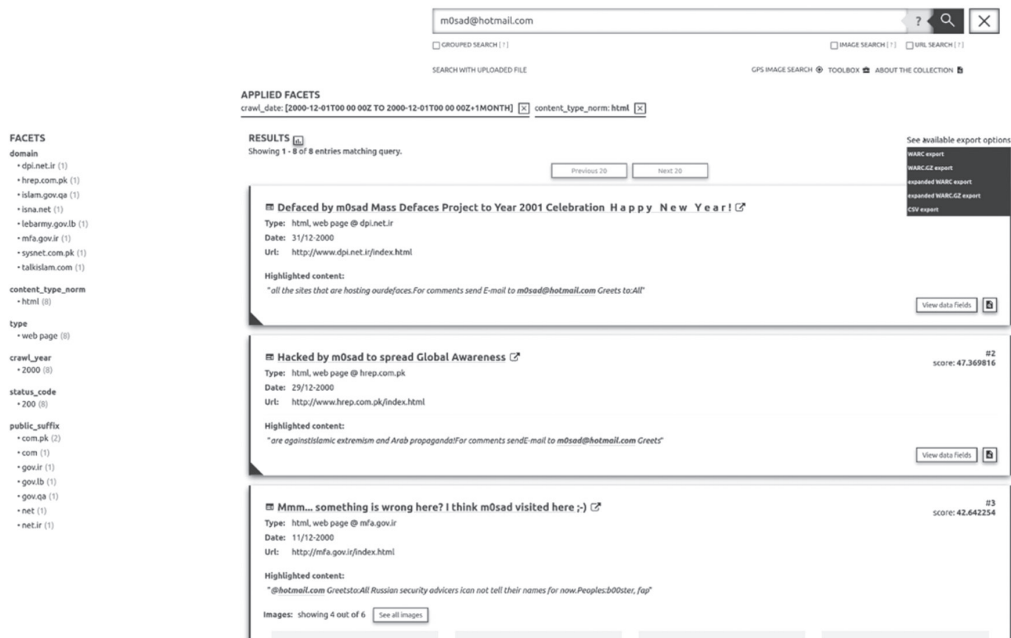


Figure 22.5 Results list view with available export options.

On the other end of the spectrum, websites with strong security setups might help build the attacker's reputation within the hacking community. Regarding the relationship between the target and the defacer, three general categories emerge:

- The target is entirely unrelated to the message and its selection was either the result of opportunistic behaviour or the pursuit of a reputation as a skilled hacker. Defacements in this category may also attempt to reach new audiences.
- The target's TLD is related to the defacer's message (such as the website of an Indian business in the context of the Kashmir conflict), while the content is not. These targets are often chosen because of their assumed allegiance with the perceived enemy.
- The target's TLD and content are clearly related to the message, such as Indian or Pakistani government websites in the context of the Kashmir conflict. Here, the defacement serves a double purpose: It shows the vulnerability of perceived enemy infrastructure and propagates a message.

Using SolrWayback's function to visualise search results by domain, the existence of different target groups can be shown. Using the highlighted examples in Figure 22.7, the following strategies are visible:

- Defacements of an unrelated target domain with the intention of reaching a new audience (economictimes.com).
- Defacement of an "enemy" domain, where content is largely unrelated to the message (nike.co.in).

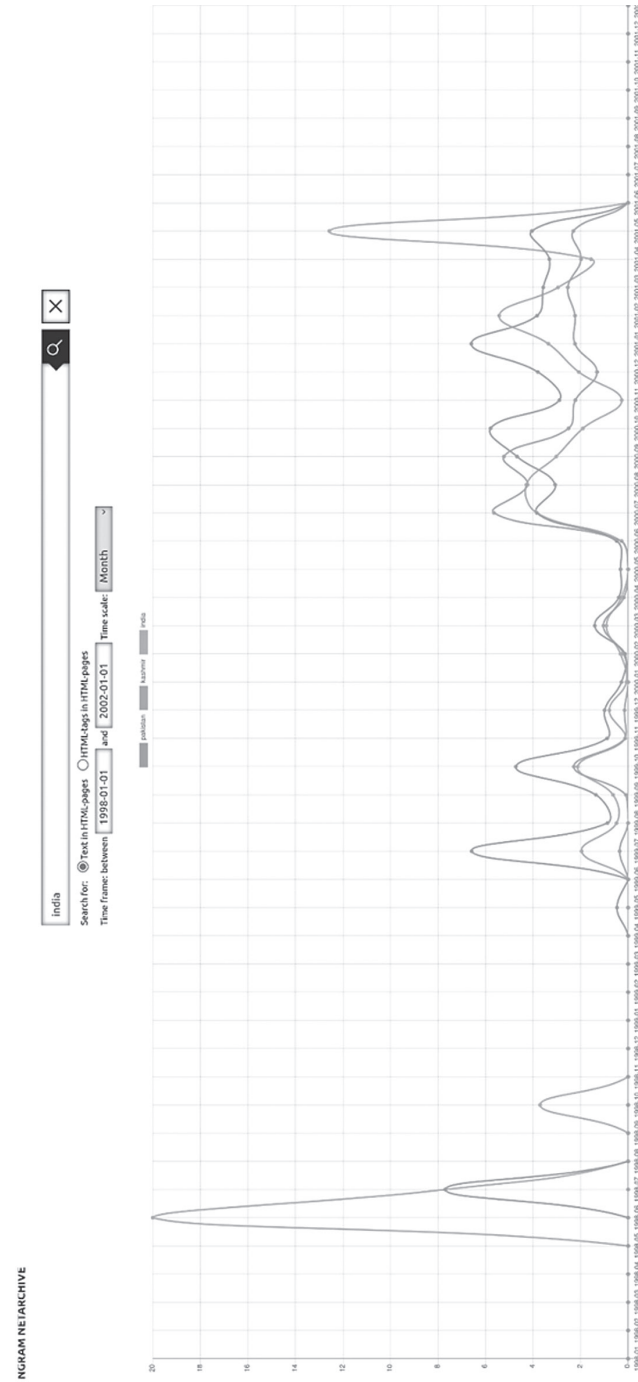


Figure 22.6 Distribution over time for keywords ‘pakistan’, ‘kashmir’, ‘India’.

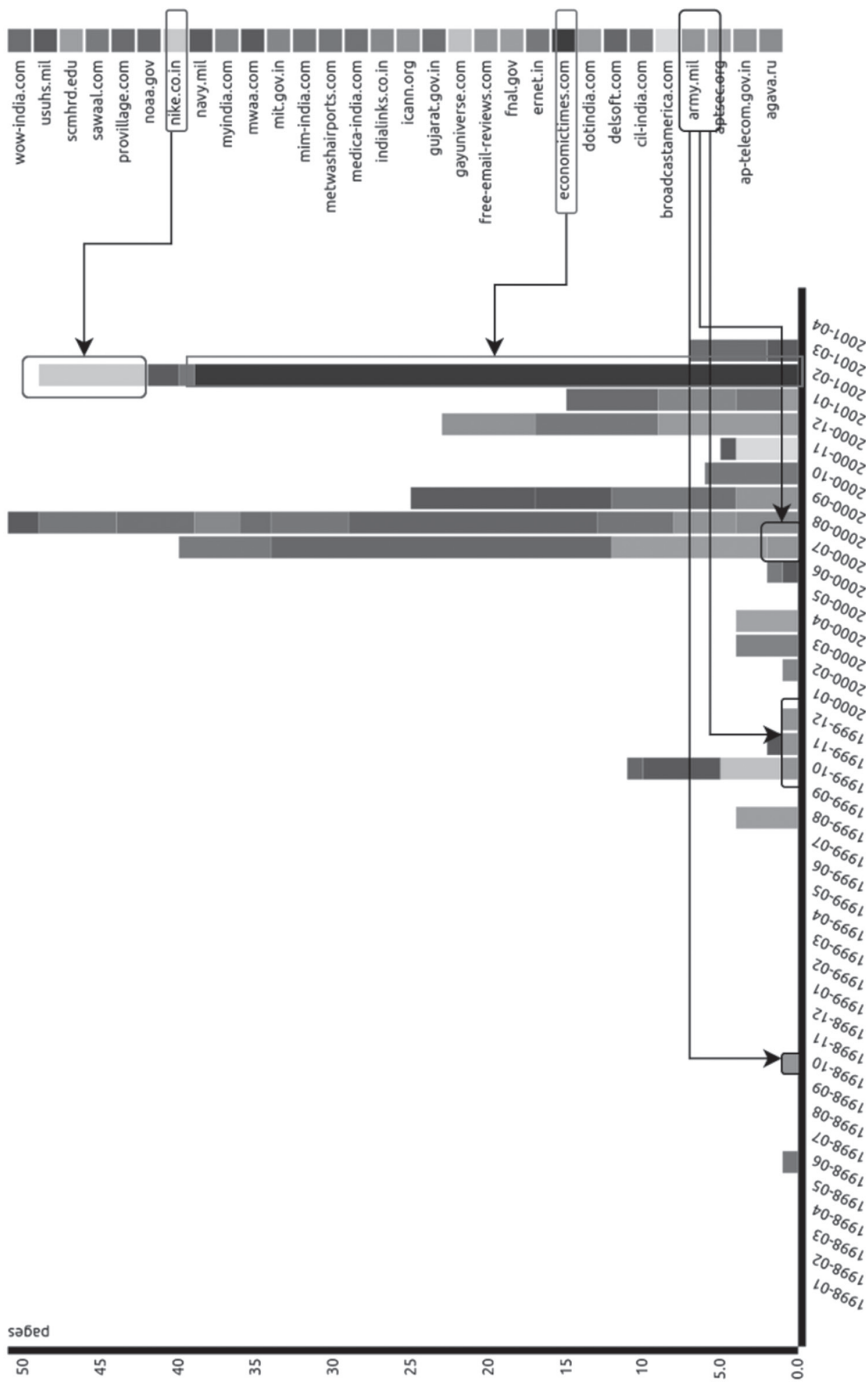


Figure 22.7 Visualisation of affected domains per month with selected domains highlighted.

Building an archive of historical web defacements

- Repeated defacements of high-profile domains (army.mil) to both increase profile as a skilled hacker and propagate political messages.

It is important to note at this point that the number of pages is not always indicative of the scope of a defacement. Rather, it seems that in later defacements the attrition.org crawler harvested the entire domain while in earlier defacements only the index.html was harvested. While complex defacements are present (where more than the index.html is altered), this collection practice leads to later defacements featuring more pages. This inaccuracy in the visualisation of results is at the same time a reminder that the research design in this case uses a generic web archiving tool. Counting pages per domain is perfectly reasonable for SolrWayback, but in this specific case not always appropriate for the source material.

Presentation and contextualisation

Presentation and contextualisation of web defacements require further effort, and can only partially be provided by an out-of-the-box web archive playback solution. In this final section, key requirements for presentation and contextualisation of web defacement are presented. This approach is largely based on the combination of different web archive data to provide the original and defaced page side-by-side.

The previous section showed that defacement can engage with the target page and domain in different ways. However, to confirm the relationship between the defacement and the original page, it is usually necessary to know the scope and content of both. While the defacements are in the archive that has been described in this section, the originals are not. The defacements described here range from 1995 to 2001 so many original pages have disappeared from the live web. While in some cases, the URLs alone can help to understand the nature of the relationship between original and defacement (such as army.mil), this is not true in all cases. Generally, the more a defacement engages with the original — for example when a defacement parodies the original — the more is having both copies important.

Regarding the potential use of a second web archive for cross-reference, I undertook an exploratory survey into the overlap between the Danish web archive and the Danish domains in the attrition archive. This brief study can help to understand the potential use of cross-referencing between web archives to contextualise defacements. The results are shown in Figure 22.8.

This brief study is by no means comprehensive and only gives a small overview of the overlap between Netarkivet and the attrition archive. It shows that in no case did Netarkivet have a copy of the defaced page in their archive, and that time differences average between +300 and -300 days. Of course, the age of the material is another factor here, as some of the older defacements predate the start of Netarkivet collection activity.¹⁰

What this data shows in the context of this chapter is that a national web archive is unlikely to have many copies of defaced pages. This may be because defacements are more ephemeral than most other types of web pages, and as such are not reliably captured by web crawls. Another possible factor is that web defacements are likely to be out of scope for most national web archives, especially those operating without a legal mandate. At the same time, this emphasises the value of the defacements which are being made available.

Returning to the idea of contextualisation through cross-referencing, the brief analysis of Netarkivet data indicates a high probability that the original of a defaced page is also archived

Difference between Defacement and Capture Date, in Days

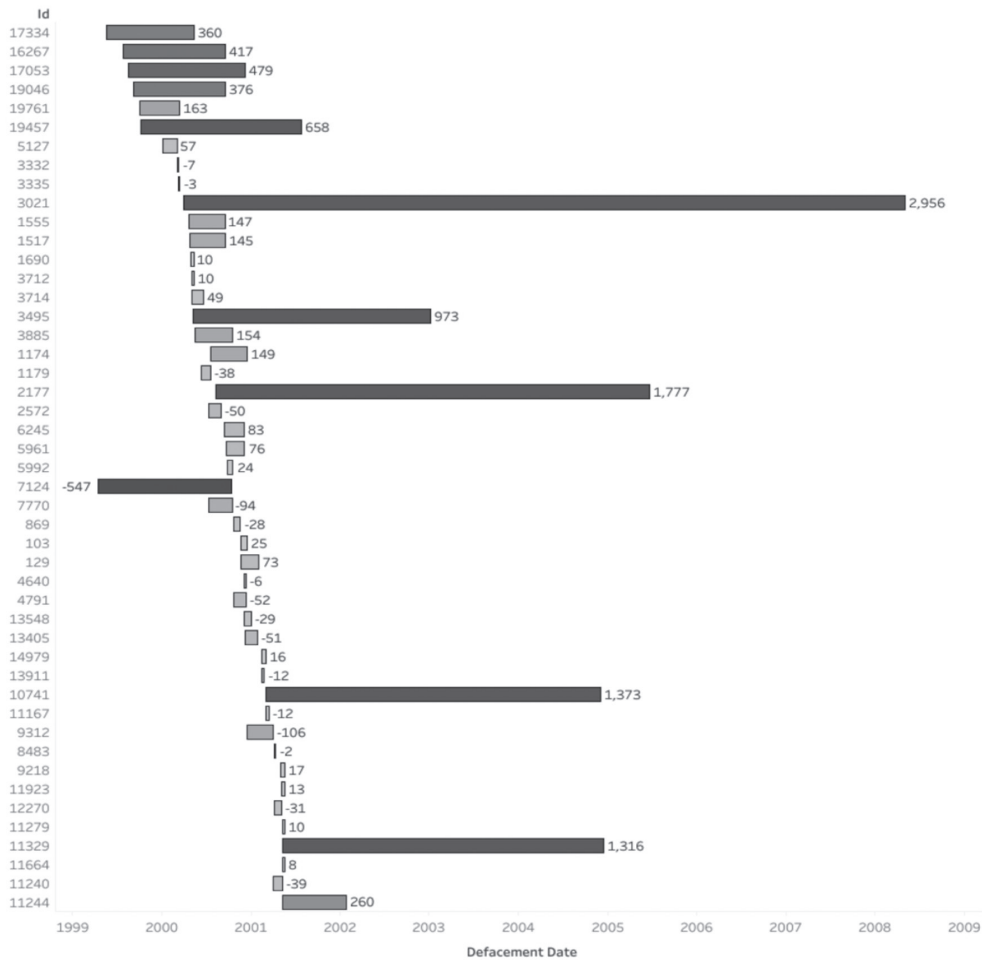


Figure 22.8 Time difference in days between defacement and nearest capture.

in a second web archive. SolrWayback allows this cross-referencing through an “alternative playback engine” function. This function searches for the URL in a second web archive and opens the results in a new tab. As default, this alternative playback engine is set to use the Internet Archive’s Wayback Machine, which is also the configuration the defacement archive is going to use. An application, and the value of cross-referencing to contextualise defacements, is described in the following.

Part of the defacement archive is a 1996 defacement of the British Labour Party website. The defacement with the banner “new Labour. Same Politicians. Same Lies” features a caricature of Tony Blair and multiple sub-pages, such as:

- The Budget Response: More of those lies all parties feed you close to an election.
- The Road to nowhere? READ THIS

Building an archive of historical web defacements

This elaborate defacement is significant as it features sub-pages expressing in detail the defacer's dissatisfaction with Labour policy. Using the full-text search, we can also find a defacement of conservative-party.org.uk where the author refers to the Labour defacement:

I am a hacker, and I am starting where the Labour hacker left off a couple of weeks ago. As he, whoever he was said, we, the people of Britain are putting our country in the hands of these parties, who can't even secure a web page, so how do we expect them to run a country?¹¹

This both confirms the authenticity of the Labour defacement and indicates that defacements are observed by other hackers, who might join the conversation through their own work.

By itself, both Labour and Conservative defacements are valuable early defacements; however, their true context can only be understood when the original page is present. Using the alternative playback engine, we can automatically retrieve the nearest copy from the Internet Archive which confirms the labour.org.uk defacement to be a parody of the original page. The original headline reads “new Labour new life for Britain”, and the sub-pages mentioned above link to “The Road to the Manifesto” and “New Information”. Contrasting the original and defacement, we can identify further parts that were altered in the defacement such as the picture of Tony Blair and the Labour logo.

Only through this contextualisation can we confirm the defacement to be an elaborate parody of the original site. This includes minute details such as the Labour logo, the photo of Tony Blair, and even the designer's logo. This contextualisation helps to understand the nature of the defacement. None of this contextualisation would have been possible through either the defacement archive or the Internet Archive alone, but only the combination of sources enables us to compare the two versions side by side.

The use of the alternative playback engine function is another example of a repurposed function taken from SolrWayback. While the original design idea of the function was to create sameness by confirming the archived content across multiple sources, the way the function is being used by the defacement archive is to create different versions of a web-site. Through these different versions, the content of the defaced page can be compared against the content of the nearest snapshot. This creates an edition-like view where multiple witnesses are compared to create a critical perspective on a transmitted text.

In the current configuration, the Internet Archive is being used as an alternative playback engine primarily for two reasons: The Internet Archive is openly accessible and its collection scope is not limited to a singular country or language. For most defacements, the Internet Archive is thus likely to hold a copy close to the defacement date. A possible next step in the development of the defacement archive might be to work with national web archives to investigate how this cross-referencing approach might be used with their collections. As the brief analysis of the Danish web archive has shown, capture dates exist within reasonable proximity of many defacements. Specifically, where topics of national interest are mentioned in defacements (for example, a few Danish domains were defaced protesting the introduction of the Euro in the country), national web archives might understand these defacements as being in scope for their collection.

Conclusion

This chapter has demonstrated the workflow for the development of an archive of defaced websites. The chapter described the data source, and the steps taken to process the data, enrich them with extracted metadata, and index the data in SolrWayback. Following the deployment of the SolrWayback instance, different approaches to the analysis of this unusual material were described. Building on a discussion about the suitability of out-of-the-box web archiving playback solutions in this special case, a method was introduced whereby defaced pages are contextualised by cross-referencing them with material from other web archives. Through these steps, the chapter provides a novel data source for the research of web histories, social movements online, and the history of political hacking in the pre-2002 era. Complementing this data source are custom finding aids such as the external link network. With this defacement archive publicly available, the archive may help further research and also offers the potential to be expanded in the future.

The available data source — essentially the complete attrition.org website, including the defacement archive — presented a rare opportunity for the development of an archive such as described in this chapter. The defacements were old, in many cases predating the start of national collections, but defacements were also mostly political in content. Combining these two attributes, the defaced pages offered the opportunity to provide valuable insight into the history of political expression through illicit means. Adding to the value of the data was its accessibility. Before the data was published on GitHub, attrition administrator Jericho granted research access to the archive based on a shared understanding of the potential value of these defacements. That the data was eventually published under a CC-BY-NC-SA (Attribution-NonCommercial-Share-Alike) licence greatly helped this project.

What was quickly found, however, was that the acquired data was a collection of loose files, without descriptive metadata. The first part of the workflow was therefore dedicated to the creation of WARC files that contained the relevant metadata. Especially when converting local files into WARC files, the original URL and time of harvest must be manually added to the WARC header. Once the files had been converted from what was essentially a collection of HTML, CSS, and image files in a variety of folders, the files could be indexed in a web archive playback solution. The most prominent solutions for this would be SolrWayback, OpenWayback, and Pywb. SolrWayback was eventually chosen for its range of analysis features (some of which were shown in the chapter) and its ready integration of other web archives as alternative playback sources.

While SolrWayback has its clear advantages, no web archive playback solution is designed specifically for defaced websites. The analysis part thus had to develop custom strategies for the analysis of defacement tools that had to be either developed or reappropriated. The external link network is an example of a custom tool developed to group defacements based on the link targets used on pages. Assuming that linking to the same external source indicates some level of alignment between defacements, this external link network allows the identification of themes in defacements quite effectively. Using the full-text search, the link targets can then be used to retrieve defacements from SolrWayback. An example of a reappropriated tool is found in the use of the Internet Archive as an alternative playback engine. Originally used to confirm sameness between snapshots of websites, the defacement archive uses the alternative playback function to create edition-like comparisons of the original and defaced websites. That this approach can greatly improve the understanding of a defacement was shown in the case of the labour.org.uk and conservative-party.org.uk defacements.

Building an archive of historical web defacements

In its current state, the defacement archive provides the first corpus of defaced pages from a singular source. The time period covered by the material spans from 1995 to 2002 and features defacements concerned with geopolitical events such as the Kashmir and Palestine conflicts, but also scene-internal topics such as the arrest of hacker Kevin Mitnick (Mitnick & Simon, 2011). Possible development of this project in the future may include the expansion of the source data by incorporating more defaced pages from a variety of sources, as well as moving toward a unified interface solution that combines both custom finding aids and existing functions in SolrWayback.

While this archive provides a crucial foundation, the future evolution of this project holds the potential for significant expansion by incorporating defaced pages from diverse sources, introducing a transnational dimension. This broader inclusion could capture a more comprehensive narrative of global hacktivism, thereby highlighting how defacements resonate across borders, reflecting interconnected political concerns.

Overall, this chapter has presented a workflow used to develop an archive of defaced websites. It has demonstrated the steps taken to convert a collection of HTML, CSS, and image files into indexed WARC files and how custom analysis tools can be used to understand and contextualise defacements. The defacement archive presented is the first of its kind, offering valuable insight into the history of political expression through illicit means, and is publicly available for further research and expansion.

Acknowledgements

This work would not have been possible without the patient support and advice from many colleagues, especially Thomas Egense and Anders Klindt Myrvoll from the Royal Danish Library. Dr. Karin de Wild recommended SolrWayback for exactly the features described in the chapter and thereby greatly helped the development process of the defacement archive.

Notes

- 1 Defacements have been described in the tradition of adusting and detournement (Lievrouw, 2011, p. 72).
- 2 See, for example, web defacements in the context of anti-globalisation protests where defacements were used alongside other forms of repurposed media (Vegh, 2003, p. 86).
- 3 Websites are in constant need of repair and maintenance, not only to protect them from defacers. The response to a defacement, whether to restore the page, to delete it altogether, or not to react at all, speaks to the value placed on web pages as technical infrastructure (Krebs & Weber, 2021, p. 21).
- 4 SolrWayback is a web application for browsing historical harvested ARC/WARC files similar to the Internet Archive Wayback Machine.
- 5 This is not the place to discuss the actual ownership of these defacements. For a discussion on the ethical considerations of working with defaced websites, see Kurzmeier (2021).
- 6 For more information on the public image defacer groups maintain, including interviews and personal websites, see Kurzmeier (2021).
- 7 Some defacers use the claim that ‘no (permanent) harm was done’ to the target system. The claim serves to establish a group identity by publicly embracing the group’s values of seeing hacking as a creative act and implies this defacer is pursuing a higher goal where the means justify the end, especially because allegedly no harm was done to anybody (Kurzmeier, 2021, p. 133).
- 8 For a general history of the region and conflict, see Parashar (2018), Schofield (2010) Bhat (2019). For the role of digital communication in the conflict, see Bunt (2003).

- 9 For example, outdated Content Management Systems with known vulnerabilities are ideal targets for such mass defacements (Abela, 2020).
- 10 Material pre-2005 was ingested into Netarkivet from the Internet Archive.
- 11 Quoted here using the original grammar.

References

- Abela, R. (2020). Statistics Show Why WordPress Is a Popular Hacker Target'. *Wp White Security*. <https://web.archive.org/web/20240311070928/https://melapress.com/wordpress-security-survey-results-2022/>.
- Akrich, M. (1992). The De-Description of Technical Objects. In W. E. Bijker & J. Law (Eds.), *Shaping Technology/Building Society: Studies in Sociotechnical Change* (pp. 205–224). MIT Press.
- Al-Aqsa Intifada Timeline. (2004, September 29). https://web.archive.org/web/20240311071156/http://news.bbc.co.uk/2/hi/middle_east/3677206.stm.
- Attrition.org Web Page Defacement Mirror. (n.d.). Retrieved October 4, 2022, from <https://web.archive.org/web/20240311071254/https://attrition.org/mirror/>.
- Balduzzi, M., Flores, R., Gu, L., & Maggi, F. (2018). *A Deep Dive into Defacement: How Geopolitical Events Trigger Web Attacks*. Trend Micro Forward-Looking Threat Research (FTR) Team.
- Bhat, S. A. (2019). The Kashmir Conflict and Human Rights'. *Race & Class*, 61(1), 77–86. <https://doi.org/10.1177/0306396819850988>.
- Brügger, N. (2013). Historical Network Analysis of the Web. *Social Science Computer Review*, 31(3), 306–321. <https://doi.org/10.1177/0894439312454267>.
- Brügger, N. (2019). Understanding the Archived Web as a Historical Source. In N. Brügger & I. Milligan (Eds.), *The SAGE Handbook of Web History* (pp. 17–29). SAGE.
- Bunt, G. R. (2003). *Islam in the Digital Age: E-Jihad, Online Fatwas and Cyber Islamic Environments*. Pluto Press.
- Camp David Proposals for Final Palestine–Israel Peace Settlement. (n.d.). Retrieved May 9, 2023, from <https://web.archive.org/web/20240311071428/www.mideastweb.org/campdavid2.htm>.
- de Bakker, F. G. A., & Hellsten, I. (2013). Capturing Online Presence: Hyperlinks and Semantic Networks in Activist Group Websites on Corporate Social Responsibility. *Journal of Business Ethics*, 118(4), 807–823. <https://doi.org/10.1007/s10551-013-1962-1>.
- Denning, D. E. (1999). Activism, Hacktivism, and Cyberterrorism: The Internet as a Tool for Influencing Foreign Policy. *Global Problem Solving Information Technology and Tools*. <https://web.archive.org/web/20240311071524/https://nautilus.org/global-problem-solving/activism-hacktivism-and-cyberterrorism-the-internet-as-a-tool-for-influencing-foreign-policy-2/>.
- Healy, S., Byrne, H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., & Jansma, R. (2022). *Skills, Tools, and Knowledge Ecologies in Web Archive Research*. WARCnet Papers. https://web.archive.org/web/20230924000852/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Skills_Tools_and_Knowledge_Ecologies.pdf.
- Hopkins, T. (2014). *Virtual Graffiti: Describing Humans*. Carleton University. <https://web.archive.org/web/20240311071614/https://repository.library.carleton.ca/>.
- Kelly, C. (2009). *Cracked Media: The Sound of Malfunction*. MIT Press.
- Krebs, S., & Weber, H. (Eds.). (2021). *The Persistence of Technology: Histories of Repair, Reuse and Disposal* (1st ed.). transcript Verlag. <https://doi.org/10.14361/9783839447413>.
- Kurzmeier, M. (2021). *Political Expression in Web Defacements*. Maynooth University. <https://web.archive.org/web/20240311071717/https://zenodo.org/records/6308125>.
- Kurzmeier, M. (2023). Unlocking Political Web Defacements for Research. *Journal of Digital History*. <https://doi.org/10.1515/jdh-2021-1006>.
- Lavoy, P. R. (Ed.). (2009). *Asymmetric Warfare in South Asia: The Causes and Consequences of the Kargil Conflict*. Cambridge University Press.
- Lievrouw, L. A. (2011). *Alternative and Activist New Media*. Polity.
- Maggi, F., Balduzzi, M., Flores, R., Gu, L., & Ciancaglini, V. (2018). Investigating Web Defacement Campaigns at Large. *Proceedings of the 2018 on Asia Conference on Computer and Communications Security — ASIACCS '18*, 443–456. <https://doi.org/10.1145/3196494.3196542>.

Building an archive of historical web defacements

- Milligan, I. (2016). Historiography and the Web. In I. Milligan, *Exploring Big Historical Data: The Historian's Macroscope* (pp. 3–15). Imperial College Press.
- Milligan, I. (2022). *You shouldn't Need to be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure*. WARCnet Papers, 107–110. <https://doi.org/10.1145/2910896.2910913>.
- Mitnick, K. D., & Simon, W. L. (2011). *Ghost in the Wires: My Adventures as the World's Most Wanted Hacker* (1st ed). Little, Brown and Company.
- Parashar, S. (2018). Competing Masculinities, Militarization and the Conflict in Kashmir. *International Feminist Journal of Politics*, 20(4), 663–665. <https://doi.org/10.1080/14616742.2018.1532179>.
- Rogers, R. (2019). Periodizing Web Archiving. In N. Brügger & I. Milligan (Eds.), *The SAGE Handbook of Web History* (pp. 42–56). SAGE.
- Romagna, M., & van den Hout, N. J. (2017a). Hacktivism and Website Defacement: Motivations, Capabilities and Potential Threats. In 27th Virus Bulletin International Conference (vol. 1, pp. 1–10).
- Romagna, M., & van den Hout, N. J. (2017b, October 11). Hacktivism and Website Defacement: Motivations, Capabilities and Potential Threats. In 27th Virus Bulletin International Conference, Madrid. https://web.archive.org/web/20240311071825/www.researchgate.net/publication/320330579_Hacktivism_and_Website_Defacement_Motivations_Capabilities_and_Potential_Threats.
- Ruest, N., Fritz, S., & Milligan, I. (2022). Creating Order from the Mess: Web Archive Derivative Datasets and Notebooks. *Archives and Records*, 43(3), 316–331. <https://doi.org/10.1080/23257962.2022.2100336>.
- Samuel, A. (2014). *Hacktivism and the Future of Political Participation*. Harvard University. <https://web.archive.org/web/20240311071929/www.alexandrasamuel.com/dissertation/pdfs/Samuel-Hacktivism-entire.pdf>.
- Schofield, V. (2010). *Kashmir in Conflict: India, Pakistan and the Unending War* (Fully revised ed.). Tauris.
- Vegh, S. (2003). Classifying Forms of Online Activism. In M. McCaughey & M. D. Ayers (Eds.), *Cyberactivism: Online Activism in Theory and Practice* (pp. 81–97). Routledge.
- Waldherr, A., Maier, D., Miltner, P., & Günther, E. (2017). Big Data, Big Noise: The Challenge of Finding Issue Networks on the Web. *Social Science Computer Review*, 35(4), 427–443. <https://doi.org/10.1177/0894439316643050>.
- Weber, M. S. (2022). *Web Archives: A Critical Method for the Future of Digital Research*. WARCnet Papers, https://web.archive.org/web/20230902214753/https://cc.au.dk/fileadmin/user_upload/WARCnet/Weber_Web_Archives_A_Critical_Method.pdf.
- Woo, H., Kim, Y., & Dominick, J. (2004). Hackers: Militants or Merry Pranksters? A Content Analysis of Defaced Web Pages. *Media Psychology*, 6(1), 63–82. https://doi.org/10.1207/s1532785xmp0601_3.

23

CONVERSATION 4

Institutional challenges

*Friedel Geeraert, Karin de Wild, Moisés Rockembach,
Jane Winters, Michael Kurzmeier, Nadezhda Povroznik,
Helena Byrne, Juan-José Boté-Vericad, and Sharon Healy*

How do you see the role of web archives in the wider heritage environment of GLAM (Galleries, Libraries, Archives, and Museums) institutions?

Friedel Geeraert: Collections of archived web content can be a connecting factor between GLAM institutions. There is a large variety in the kinds of collections these institutions manage. However, with the increasing awareness about the importance of web archiving, web archives can become a common collection type between GLAM institutions, thereby further stimulating the development of shared standards and practices.

Karin de Wild: There are still opportunities for further collaboration among GLAM institutions to preserve born-digital heritage. In her studies on “Networks of Care”, Annet Dekker (2022) demonstrates how digital artworks are often preserved within various institutions, each adhering to its own standards and expertise. For instance, she discusses the artwork “Brandon”, which consists of a website now part of the Guggenheim’s permanent collection. Simultaneously, other digital traces of the artwork can be found in the Internet Archive, while ephemera from offline events are still located at De Waag in Amsterdam and Fales Libraries & Special Collections at NYU. How can the various institutions collaborate to preserve the artwork as a whole? How can they provide access and create connections between the various technical and material elements of the artwork? Dekker argues for a more collaborative approach to preserving born-digital heritage to achieve more sustainable solutions.

Moisés Rockembach: Web archives, with their rich repositories of digital culture and history, offer unique materials that can be integrated into dynamic online exhibits or used in digital storytelling projects. This not only broadens the audience reach for GLAM institutions but also deepens engagement

Conversation 4: Institutional challenges

by allowing users to explore historical contexts, narratives, and connections in a more immersive and interactive manner. Leveraging web archives in this way can transform the traditional perception of archives, making them lively hubs of culture and learning that extend well beyond their physical premises.

Jane Winters:

The GLAM framing encompasses a wide and highly diverse range of institutions, some of which are already deeply engaged with web archives (most often libraries and archives) while others are only just beginning to grapple with the challenges of born-digital collecting. It will be interesting to see how institutions with a focus on exhibition and display start to engage with innovative methods of presenting web archives and other born-digital data alongside analog materials in gallery spaces.

How do you see the role of non-institutional/political activist/fan-based or other amateur/volunteer-based practices as an addition to institutional approaches to web archiving?

Michael Kurzmeier:

Noninstitutional actors have provided important additions to the body of archived web content available today. Their advantage over institutional web collection practices is generally a higher degree of organisational flexibility and the ability to employ community resources for archiving. Their disadvantage is, broadly speaking, a lack of resources including curatorial work and network infrastructure. These characteristics make non-institutional actors well-suited for focussed archiving efforts, which give the added bonus of easier community formation around this focus or shared agenda.

From the perspective of institutional actors, caveats generally exist around the understanding of good archival practices, including ethics, collection scope and metadata. Where these caveats exist, they must be resolved before an institutional actor can meaningfully engage with archived material from a non-institutional source. This process becomes more difficult as time passes and communities dissolve, data is lost and collection practices are harder to reconstruct.

As non-institutional archives can be of special interest to institutional web archiving due to the age or nature of the material, they can be important additions to institutional web archives if the source data can be enriched with metadata and questions of collection practice can be resolved.

Nadezhda Povroznik:

The various non-institutional/political activist/fan-based or other amateur/volunteer initiatives can be central initiatives in the countries where the national programs of web preservation do not exist or these programs are biased and politically engaged due to the specifics of the regime that can be autocratic and oppressive. In Russia, the only program of web preservation is initiated by the autonomous non-profit organisation “Information Culture” (“Informatsionnaya Kultura”¹). The National Digital Archive of Russia (Natsionalniy

Tsifrovoy Archiv Rossi²) is a project on web preservation of Russian websites conducted by the “Information Culture”. The main focus of this web archiving project is on the governmental websites but also the vast list of websites for regular crawling is published on the web-site. It includes the websites of some oppressed organisations such as Memorial. International Memorial (or International Historical Educational Charitable and Human Rights Society “Memorial”) is a non-profit and non-governmental organisation aimed at “studying political repressions in the USSR and in present-day Russia and promoting moral and legal rehabilitation of persons subjected to political repressions” (What is International Memorial³). The mentioned initiative by “Information culture” is a wonderful example of activism aimed at preserving web history and the voices at risk. The current political situation makes the functioning of the “Information culture” rather difficult and for these reasons, there should be space for dialogue, open discussions, and finding measures to support relevant initiatives and sustain their valuable activities.

How to include archival practices of software as a standardised element of the curation of web archives?

Michael Kurzmeier: In the wider context of social sciences and humanities, catalogues such as the SSH Open Marketplace allow researchers to document their workflows, including tools and data sources used. Through these workflows, research software and file formats used can be described in detail so that the research process becomes transparent and reproducible. This general approach can be adapted to the specific needs of web archive curation practices to increase transparency and reproducibility of archival workflows, making more space for the description of the effects of software on the harvested data.

Moisés Rockembach: I agree with Michael Kurzmeier’s approach, and I would like to emphasise the importance of standardisation and the use of metadata in web archiving practices further. Standardising archival practices and employing detailed metadata ensures that each piece of archived content is not just preserved but also searchable and analysable in context. Hence, adopting standardised practices and metadata enriches the archival ecosystem, making it more accessible and interpretable.

What are the possible implications of having to study a nation’s online culture if the nation in question does not have an extended web archive (or any web archive at all), thus forcing one to use other existing collections (national/transnational)?—and what to do about this?

Michael Kurzmeier: The fewer sources available, the more restricted the access to a nation’s online culture. Relying on collections where the research object was not in the focus of their collection practices can potentially skew this

Conversation 4: Institutional challenges

perspective even further as the research object may be treated as marginalia. Especially in regards to older web content, the absence of a national web archive often means losing access to large parts of the nation's online culture whereas more recent material may still be preserved through the platforms that hosted the content.

Additionally, researchers should explore unconventional sources such as social media, forums, and community-driven platforms. These platforms often serve as informal archives of cultural expressions, public discourse, and evolving trends. While they may not replace a formalised web archive, they can provide valuable glimpses into the nation's digital landscape.

To mitigate the risk of marginalising the research object within existing collections, scholars should advocate for the inclusion and recognition of their subject within the archival community. Engaging in conversations with archivists, librarians, and digital preservationists can help raise awareness about the importance of documenting and preserving the nation's online culture.

Ultimately, the absence of a national web archive necessitates a creative and collaborative approach to studying a nation's online culture. By combining transnational collections, fostering partnerships, and exploring alternative sources, researchers can work towards a more nuanced and comprehensive understanding despite the limitations imposed by the lack of a dedicated archival infrastructure.

Friedel Geeraert: Being active in the field of web archiving in a country where no web archive exists on the national level puts things into a rather harsh perspective. Researchers wishing to study the Belgian web from the 1990s for example would have to search for web archives that have a worldwide collection scope such as the Wayback Machine of the Internet Archive or use web archives set up at archival institutions with a more regional or thematic collection scope to find archived the early Belgian web. Most small-scale Belgian initiatives only started in the 2000s however, so the Belgian web of the 1990s has mostly disappeared.

Other sources comprise publications. In the early days of the web, registries were published for example, containing URLs of websites that were deemed interesting to a Belgian audience. Manuals on how to construct your own website can also provide insight into the early Belgian web. However, compared to archived web content, such sources only offer limited insights.

It is therefore necessary to raise awareness about the importance of web archiving and obtain the necessary legal framework to establish a sustainable web archive at the national level to safeguard more of the Belgian web being lost. A copyright exception for web archiving at the European level would also be beneficial.

Nadezhda Povroznik: Thinking about underrepresented cultures and communities on the web archives should be aligned with the discussions about bias, diversity, equality, and inclusivity. Some crucial factors can affect all the

potential projects aimed at preserving culture for other communities. Undoubtedly, these projects on web preservation have great value in capturing the resources that are under threat. However, fragmented preservation can bring bias and shape the gap in reconstructing the real picture. The contextual misinterpretation is such a bias. The captured collections from outside can be wrongly treated and interpreted due to being disconnected from the contexts and cultural nuances of their mutual interactions with other actors. It can cause potential ethical and other implications including embedding digital colonialism issues. Conducting relevant projects at a high level of proficiency requires cross-disciplinary and transcultural dialogue to promote these conversations and discover the solutions optimal for all the parts. I would warn from bringing political and other biases to hit up the interest in the collection marked as “national culture” if it is in fact marginalized. It distorts real situations and blurs actual understanding of the diverse national culture.

*Helena Byrne,
Juan-José
Boté-Vericad,
Sharon Healy:*

The case of the island of Ireland offers a real-world example of the challenges for studying Irish online culture using web archives due to the fact the island is administered under two separate sovereignties, and therefore have different legal mandates for the collection of web heritage. The six counties of Northern Ireland are under the sovereignty of the United Kingdom of Great Britain and Northern Ireland, and the remaining twenty-six counties are under the sovereignty of the Republic of Ireland.

At present, there are three main web archiving initiatives, which capture websites as part of their efforts for the preservation of digital heritage for the island of Ireland. Since 2011, the Public Record Office of Northern Ireland (PRONI) has been capturing websites related to Northern Ireland through a selective permissions-based collection approach and provides online access to their collections via the PRONI Web Archive (<https://webarchive.proni.gov.uk/#!>). Since 2013, the UK Web Archive has also been capturing websites in Northern Ireland through their annual national web domain crawl, under The Legal Deposit Libraries (Non-Print Works) Regulations 2013 (NPLD) (www.webarchive.org.uk/ukwa). The UK Web Archive has also captured websites in Northern Ireland through a selective permissions-based approach since 2005. Also, the UK Parliament Web Archive (<http://webarchive.parliament.uk/atoz>) and UK Government Web Archive (www.nationalarchives.gov.uk/webarchive) collect materials relevant to Northern Ireland.

In the Republic of Ireland, the National Library of Ireland (NLI) has been capturing websites since 2011 through a selective permissions-based approach, which is accessible online through the NLI Web Archive (<https://archive-it.org/home/nli>). Because digital legal deposit legislation in this jurisdiction does not account for the collection of the .ie domain, the NLI are prohibited from conducting an annual domain crawl. Nonetheless, the NLI Web Archive conducted two .ie web domain crawls in 2007 and 2017 (via a tender to the Internet Archive), however, these collections are currently inaccessible to researchers or the public because at the time of the crawls, digital legal deposit legislation was not in place to allow for the same.

Conversation 4: Institutional challenges

Therefore, the study of web heritage for the island of Ireland presents challenges due to the disparity of the collection regimes across jurisdictions. In this instance, in order for a researcher to study Irish web heritage, researchers must rely on multiple web archive collections, inclusive of the UK Web Archive, the PRONI Web Archive, and the NLI Web Archive, and are dependent on web archives such as the Internet Archive (Wayback Machine, <https://archive.org/web>), Common Crawl (<https://commoncrawl.org>), and Archive.today (<https://archive.ph>) to try and account for some of the gaps. Moreover, in the case of the web heritage of the 26 counties, prior to 2011, researchers are wholly reliant on the use of non-national web archive collections. This presents a clear indication for the importance of transnational collections for the study of Irish online culture and heritage.

Moisés Rockembach: I agree with my colleagues and emphasise the significance of addressing the potential loss of historical and cultural context when a nation lacks a comprehensive web archive. Currently, there are more nations without web archives than those with web archives. Beyond web archiving done by institutions, encouraging the use of alternative preservation methods, like community-driven archiving and web scraping technologies, is crucial. It is also vital to advocate for recognizing web archives as essential resources, just as we do with traditional paper archives. This approach will help ensure that digital heritage is valued and preserved for future research.

What lessons have you learnt from your experience working with transnational web archives? Put differently, do you have any advice on how not to be deterred by the challenges involved in this endeavour?

Helena Byrne, Juan-José Boté-Vericad, Sharon Healy: The WARCnet report *Skills, Tools, and Knowledge Ecologies in Web Archive Research* (Healy et al., 2022) highlights that the challenges in working with web archives do not necessarily decrease over time, and suggests the need for training across all levels of capability. Researchers new to working with web archives should take solace in the fact that they are not the only ones struggling to work with this content. This is why it is important to get involved in networks and events within the web archiving communities to share these challenges. Sometimes what may feel like a failed experiment can give valuable insights that support other researchers.

Friedel Geeraert: One of the main lessons I have learnt is that having a multidisciplinary approach is very worthwhile. The cross fertilisation between the different profiles involved in the studies I had the pleasure to work on was very beneficial. Everyone brought different types of expertise to the table and had different insights and perspectives on the research results because of their different backgrounds. Being able to lean on the expertise of colleagues when it comes to data visualisation or distant reading techniques for example was a fantastic experience as it helped the respective research teams to overcome certain hurdles or to solve certain challenges more quickly.

Notes

- 1 Informacionnaja kul'tura, Wayback Machine. Snapshot on March 20, 2024, <https://web.archive.org/web/20240318024446/https://anr-collabora.parisnanterre.fr/observatoire/>.
- 2 Ruarxive. Russian national digital archive, Wayback Machine. Snapshot on May 28, 2024, <https://web.archive.org/web/20230528160203/https://ruarxive.org>.
- 3 International Memorial, Mission and Statute, Wayback Machine. Snapshot on February 2, 2024, <https://web.archive.org/web/20240202205213/www.memo.ru/en-us/memorial/mission-and-statute>.

References

- Dekker, A. (2022). Networks of Care. In K. Sluis & A. Dewdney (Eds.), *The Networked Image in Post-Digital Culture* (pp. 189–207). Routledge.
- Healy, S., Byrne, H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., & Jansma, R. (2022). *Skills, Tools, and Knowledge Ecologies in Web Archive Research*. WARCnet Special Report. Retrieved from https://web.archive.org/web/20230902214716/https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Skills_Tools_and_Knowledge_Ecologies.pdf.

PART V

Institutional challenges

24

SCREENS IN STRUGGLE

From archived web corpus to readable data for history research

Sophie Gebeil and Jérôme Thièvre

Introduction

This chapter focuses on collaborative and experimental research aimed at identifying and analyzing a corpus from the web archives of the National Audiovisual Institute (INA) on webTV between 2010 and 2015. Focused on the history of French activist WebTV sites, its objective was to examine the ways in which the memories of social movements (e.g., May 1968) were portrayed online in the 2000s. Because of a secure data-sharing framework set up between Aix-Marseille University (AMU) and the INA legal deposit team, this project provided an opportunity to explore the files of archived web pages. As we will learn in this chapter, such a research design requires team science, meaning a close collaboration with archivists (Jérôme Thièvre, head of the web archive team at the National Audiovisual Institute), and IT specialists (Cyrille Le Menestrel). Based on a corpus of 58 webTV sites from May 1, 2010 to May 1, 2015, we developed a pipeline for extracting information from the files collected by the INA. The aim was to explore the feasibility of using HTML files of the corpus at two given dates to enable data mining as a means of identifying the main characteristics of the corpus and the changes observed during the five-year period. A second aim was to provide future historians or students in Humanities and Social Sciences, who might not be experts in programming or web development, a best practice on how to perform web data scraping based on a corpus from the archived web. It meant we had to design an interface capable of extracting information from JSON files, so that users could then apply qualitative or quantitative processing methods. When working together, it very quickly became apparent that this experiment on a reduced number of websites could create a pipeline starting from a web archived corpus and transforming it into readable and actionable data for any researcher in the Humanities and Social Sciences.

In this chapter, we focus on methodology and do not delve into the history of webTV and the mediatization of activist memory on the web. Instead, our focus is on the approach developed to transform the archived data from the INA into usable data for historians and other researchers in the Social Sciences and Humanities. In doing so, we will demonstrate

how a highly localized case study, led by a small team and based on a limited corpus, can contribute to the discussion on research methods in web history, including from a transnational perspective.

Firstly, we will introduce the project framework, explain how the corpus was constructed, and outline the primary methodological challenges we faced at the outset of the project. Secondly, we will outline the methodological trajectory that enabled us to establish a processing pipeline, starting with the decompression of the archived files into HTML format and culminating in the extraction of targeted research data, particularly textual content, made accessible in CSV format. Finally, we will provide an overview of a potential application of the pipeline and discuss how this exploration highlights methodological challenges associated with the automated analysis of websites in a broader transnational perspective, by discussing the interoperability and reproducibility of the approach for transnational studies.

Screens in struggle (2019–2021): Framework, corpus, and methodological challenges

Screens in struggle, a research project associated with the INA

This historical research project “Screens in struggle” concerns the mediatization of social movements within webTV in the 2000s. It falls within the field of studying representations of the past on screen, including the web (Gebeil, 2021). Originating in the United States in the early 1990s, webTV relies on streaming or continuous downloading to distribute online content. Over time, both in France and abroad, there has been a proliferation of generalist and mainstream audio–visual media online, as well as alternative audio–visual media that have disseminated reports produced by citizens, independent journalists, and activists (e.g., Indy Media, Open TV). The term “webTV” is employed here in a broader sense to encompass websites and/or audio–visual platform channels that predominantly disseminate information in the form of freely accessible short videos (Christian, 2015). Beyond purely commercial webTV, this evolving genre also encompasses platforms that emerged in the early 2000s in opposition to established media. During this period, webTV has become a platform for a wide range of civic, social, and political expression, as shown by the proliferation of descriptors applied to them, including “minority” (Rigoni, 2010), “sub-urban,” “civic,” “militant,” “participatory,” “confessional,” “local,” and “media activist.” These “alternative” broadcasters are characterized by a predominantly “left-leaning” political orientation (Ferron, 2016) and a commitment to providing a voice for those typically marginalized and overlooked by traditional media and historiography on alternative media (Carpentier, 2011).

“Screens in struggle” examines how French webTV constructs the memory of social movements in the online sphere (Pigenet & Tartakowsky, 2014, p. 4). The primary objective includes documenting the history of webTV in the 2000s, exploring which historical events have been portrayed (such as the strikes of 1936 and May 1968), and analyzing the narratives crafted by webTV creators. Another objective is to contextualize the content archived by the INA, using an approach that combines corpus analysis from the legal deposit of the web with the collection of supplementary sources related to a select number of webTV sites. This

Screens in struggle

approach aims to shed light on the production conditions and motivations of the individuals behind this content, including oral interviews.

The third objective is part of a broader methodological examination of how web archives can be used by researchers who are not experts in IT development or data science, in collaboration with the INA's web legal deposit team. This methodological component is the focus of our discussion. The research is deliberately data-driven, with the intention of developing tools to identify the primary characteristics of a corpus comprising web pages and videos from the INA's web archives.

Corpus of archived webTV sites

To build a corpus of archived webTV sites, the INA web legal deposit (DL web) team provided a spreadsheet containing a collection from the database of collected sites, including the following information: Site name, URL, coverage (local/national), function (webTV), media (web, TV, radio), various notes entered by legal deposit documentalists and engineers, personality (e.g., composer, director), program (e.g., documentary), audience (youth, adult, general public, etc.), sector of activity (broadcaster, presenter, media group, etc.), service (free, paid, registration, etc.), and theme. From this table, we added several columns: Date of creation, URL address online in 2019, location, identifier of archived videos, URL of associated social networking accounts (Facebook, Twitter, YouTube, Dailymotion, Vimeo), publisher (local authority, company, trade union, association, employee, informal collective), quality of video archiving, inclusion in the study corpus (in/out), and memorial register (heritage, activist, etc.).

Based on the list of 623 webTV sites selected, we consulted the INA's web archives to fill in the missing fields and assess the accessibility of the videos from these webTV sites, the majority of which were created in the mid-2000s, before the INA organized its own systematic collection in 2009.

This consultation phase led to the discovery of some real treasures in the web archives, in particular webTV sites run by prolific and particularly creative activist groups with very marginal online visibility, such as Ali Bernard TV, created in 2008 and devoted to news from the grassroots Arnaud Bernard neighborhood in Toulouse. The entire corpus therefore comprises 58 sites, or 12% of the initial selection. After data cleansing, the corpus is defined based on URLs and thus includes all the pages of the 77 activist webTV sites identified beforehand and collected between April 1 and June 1, 2010 and 2015. The very local coverage of events is one well-identified characteristic of these community media, exemplifying the concept of glocalization. These entities, operating on a smaller scale, relate to national and transnational organizations, sharing common goals reminiscent of the alter-globalist online media of the 2000s (Carion, 2009). Consequently, alternative or minority webTV sites serve as valuable subjects for investigating cultural and activist circulations, contributing significantly to the understanding of a transnational web history. The construction of case studies is anchored in reduced corpora, fostering collaborative and interdisciplinary endeavors, and thereby advancing the development of appropriate methodologies.

While our desire was to use this relatively small corpus to design an approach that would allow for an initial, distant reading of the data, the characteristics of the sites in the corpus (small-scale, relatively marginal, structured randomly, short lifespan, uneven collection of videos) also reinforced the need to steer the project toward the study of the HTML pages from a diachronic perspective.

***Initial methodological challenges: Interdisciplinary collaboration
addressing the challenge of alternative archiving formats***

After more than a quarter of a century, web archives are no longer “archives without historians” (Rygiel, 2005). At various scales, institutions, researchers, organizations, and transnational collectives have worked tirelessly to develop uses, raise public awareness, and deepen our understanding of this incredibly complex heritage. Today, its exploration relies on a variety of methods, some more automated than others, based on datasets ranging from single website monographs to massive corpora. Exchanges between research communities and professionals have underscored the need to develop specific tools (Gebeil, 2021) and, above all, true “research instruments” (Ruest et al. 2022).

It should be noted that this initiative is not new to the field of web archives, but rather follows on from the use of dedicated analysis tools such as The Archived Unleashed Toolkit, Solr Wayback, and GLAM Workbench. These tools aim to retrieve information from web archives, such as TimeFind, which can be used to find the modification date of a web page using the Internet Archive, but also to explore collections of web archives. Very often these tools apply to live web formats or Web ARChive (WARC) files. The challenge here is to design a processing chain adapted to the specific archive format used by the INA (DAFF — Digital Archiving File Format).

Regarding the exploration of files derived from web archives with the goal of studying a collection while reducing technical barriers for users, the two primary projects that serve as benchmarks are Solr Wayback¹ and The Archives Unleashed (AU) Toolkit,² both of which were established in 2017. The first project, developed by the Royal Danish Library, provides users with the ability to explore data extracted from WARC files by combining HTML text and content from web pages in the corpus (including images, videos, and various documents). The AU Toolkit offers a set of tools for processing massive data from web archives, conducting textual and network analysis, and visualization. Like most web archive exploration tools, they are geared toward the WARC format, which has emerged as an ISO standard. However it is clear that a personal project, funded with €2,000 in 2019–2020 by the Temps, espaces, langages, Europe méridionale, Méditerranée (TELEMMMe) laboratory under a one-year associate researcher contract, does not aim to produce a tool as extensive in scope as Solr Wayback or AU.

While this research is a follow-up of these initiatives aiming at analyzing WARC files for research purposes, it has several specificities. Firstly, the INA does not use the WARC format but the DAFF format to store collected files. Secondly, this project was led by a small team. Thirdly, it aimed to make archived web data readable by historians, who are not experts in programming or data science, so that they can browse a massive corpus with basic digital skills. The starting point was to establish mutual needs. With regard to historians, despite easy access to the INA web archives, the diachronic research potential increases the amount of usable data even with a limited number of sites. As for the legal deposit of web media, it is a question of reconciling heritage missions with the desire to meet the needs of researchers and to increase the use of web archives in the scientific community.

The final aim was to find a method for exploring a heterogeneous corpus that is too extensive to be examined manually. This challenge often arises with born-digital materials. Therefore, the argument put forth was that to devise suitable tools and methods, it was necessary to get hands-on, in other words, to experiment together with what could be accomplished within a short timeframe and a limited budget, using DAFF files as a starting

Screens in struggle

point. In this case, it implies that the historian adopts an approach in which the expected outcome goes beyond merely unearthing past facts; it involves the development of a specific approach and toolset, complete with its contributions and limitations. Learning by doing (and therefore making mistakes) is also necessary for a better understanding of institutional specifics, refining research questions, and gaining technical proficiency to facilitate historians' access to usable data from archived web content.

Under these circumstances, this experience contributes to the broader discussion surrounding the development of interfaces that facilitate historians' engagement with web archives. It also addresses issues related to the format. The DAFF format employed by the INA represents an alternative to WARC among others. According to Wang and Xie (2020), WARC exhibits "significant performance penalties," leading several web archivists to opt for different formats. Furthermore, WARC is challenging to handle because, as noted by Ruest et al. (2022), "WARC can contain all the data of a web page and technical metadata about its collection, which is stored in record headers or WARC headers." The format used by the INA differs in this regard. The DAFF format was chosen in 2001, when web archiving was still in its infancy and the Internet Archive was the only institution practicing it on a large scale. Collection tools and storage formats were virtually non-existent. It was in this context that the INA began its preparatory work for web archiving, using a technical and scientific approach inherited from the introduction of legal deposit for radio and television a few years earlier. Between 2001 and 2007, the INA developed a web crawling robot, Phagosite, a web archive consultation system, Vortex, and a storage format, DAFF. At the same time, the International Internet Preservation Consortium (IIPC) was founded in 2003 by the Bibliothèque nationale de France and a dozen other institutions, including the Library of Congress and the Internet Archive. The INA joined the IIPC in 2004. The IIPC then decided to promote the tools and formats developed by the Internet Archive to its community. Unfortunately, these tools are not suitable for the very frequent collections carried out by the INA. The INA's collection system and its DAFF archive format allow for full de-duplication of archived content, which is essential when collections are frequent and the duplication rate of collected data is high.

With these objectives and specificities in mind, we have developed a pipeline for extracting, making available, and exploring corpus data.

Pipeline from web archive DAFF files to readable data

To facilitate a satisfactory processing chain from the extraction of the corpus to the researcher, we selected a sample of 15 sites whose files were extracted over two periods: April to June 2010 inclusive and April to May 2015 inclusive. The total size of data is 419 GB.

*Design of a processing chain from DAFF to the extraction
of archived HTML pages*

Once the corpus was selected, the INA developed a procedure for extracting data from the archived DAFF files into Newline Delimited JSON³ (NDJSON). A versatile data format, an NDJSON file is a text file encoded in UTF-8 (Unicode Transformation Format — 8-bit), where each line constitutes a JSON object. Accordingly, each of the pages exported for this project is depicted by a single line of JSON, and the structure is detailed in Appendix 1. The

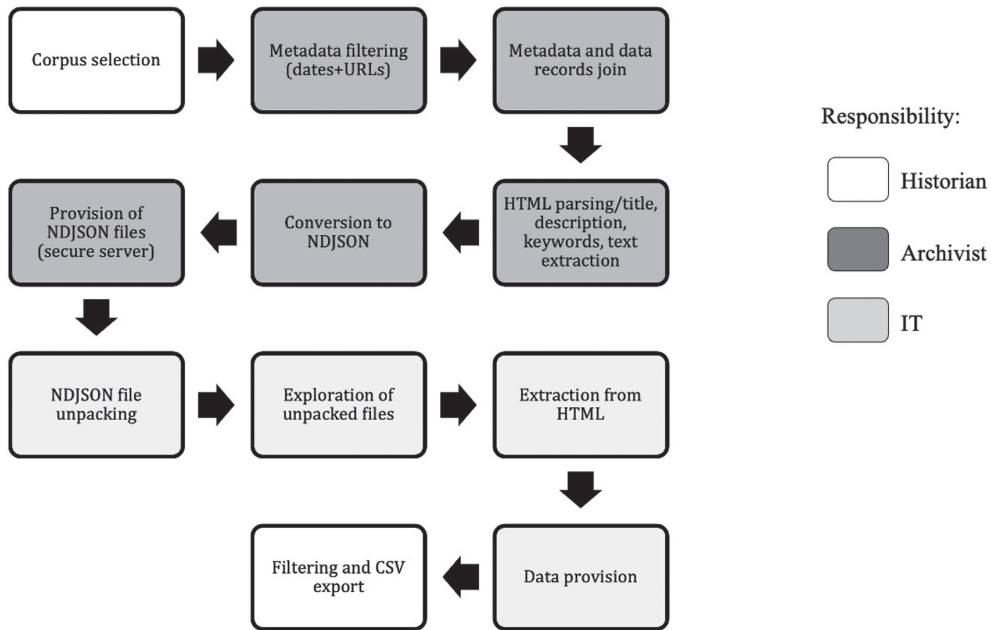


Figure 24.1 Pipeline: from corpus creation in web archives to readable data export.

benefit of this format is its compatibility with most contemporary programming languages. This involved several steps, as outlined in the pipeline diagram (see Figure 24.1):

- Filtering metadata corresponding to identified URLs and dates
- Joining the filtered metadata records with the data records
- For each data record, parsing the HTML and extracting the title, description, keywords, and text
- Converting to NDJSON format
- Making the NDJSON files available on a secure server

The DAFF format imposes a strict separation between the metadata of archived resources and their content. A metadata record is created for each collected resource, containing descriptive, spatial, and temporal information:

- URL
- Collection date
- Request and response headers
- SHA-256 signature of resource content
- A record is created when the content of the resource is new.
- The bytes of the resource content
- The SHA-256 signature of that content.

The identification of new content is determined by the uniqueness of its SHA-256 signature among those of all previously collected content. Each record is associated with one

Screens in struggle

or more metadata records. The INA has a Hadoop cluster, which allows it to efficiently perform massive processing on archived data by distributing data reading and processing across multiple servers. Hadoop is a technological building block enabling multiple servers to work together to process large quantities of data in reduced timeframes. Both the processing tasks and the data are divided and distributed across a cluster of servers for parallelization. In the case of the INA, a cluster of 10 servers, each equipped with 32 cores, can reduce processing times by a factor close to 300 (approximately 10×32). A task that would take several months with a traditional approach can thus be completed in just a few hours or days.

One of these processes is the full-text indexing of web pages. To this end, the INA stores all DAFF metadata records and the part of the data records corresponding to web pages on this infrastructure. By 2021, this will amount to 20 billion metadata records (16 TB compressed) and 9 billion data records (200 TB compressed) for web resources such as HTML pages. These data should be compared to the 120 billion metadata entries and 14 billion unique contents collected in total through the INA's legal deposit web collection. The storage of these data in DAFF format amounts to 3 petabytes but corresponds to 11 petabytes of collected data. Content deduplication facilitated by the DAFF archive format offers a valuable storage advantage, enabling frequent web collection activities to be maintained.

In the past, the INA has already exported web archives, directly in DAFF format as part of the ANR CARTEC project⁴ or converted to ARC format as part of a joint collection by IIPC members for the 2012 and 2016 Olympic Games. Today, the most widely used format for web archives is the WARC format developed by the IIPC, and many libraries allow it to be processed in different languages. A tool for converting DAFF to WARC is currently being developed but is not yet mature enough to be used in this project.

Furthermore, web archiving formats are not always the most suitable for automated processing (Wang and Xie, 2020; Ruest et al., 2020). Whether DAFF or WARC, these formats were designed to create archive files during data collection by web crawling robots. Consequently, tens of thousands of web resources of various types end up together in a single archive file, often in an undetermined order. It is worth noting that in the case of DAFF, the collection metadata, which also includes information from HyperText Transfer Protocol (HTTP) request and response headers, are completely separated from the collected content to enable content deduplication. The storage of archive files is accompanied by indexing to facilitate access and browsing of web pages in a Wayback Machine. However, when exporting data for the scientific community, it appears more appropriate to perform pre-processing to filter relevant data, such as HyperText Markup Language (HTML) pages, and assemble them into a structure and format directly usable by researchers.

In the WARC file format archiving ecosystem, various tools and formats facilitate the export and extract of data derived from web archives. Among these, the Web Archive Transformation (WAT) and WARC Encapsulated Text (WET) formats are the most referenced and most widely used in the community, notably by Common Crawl and Archive-It. The WAT format is specifically designed for storing metadata such as HTTP request and response headers. In cases the data represents an HTML resource, it stores titles, hyperlinks, and other standard HTML metadata. The WET format, on the other hand, is geared toward storing text extractions from textual resources, such as HTML content. In alignment with the practices of Common Crawl and Archive-It (Lohndorf, 2023), our project adopts a similar approach and extraction type. We exported the same metadata as the WAT and WET

formats, apart from hyperlinks. In our case we opted to export the entire HTML content. The corpus was defined by a list of 58 websites over two two-month intervals between May 1, 2010 and May 1, 2015. Each site is defined by a domain name, which makes it easy to identify the pages to export based on their URL and collection date. The export process is implemented as a set of Map-Reduce jobs that are executed against INA's Hadoop cluster and can be divided into three main steps:

Step 1: Go through all the metadata to extract those that correspond to the 58 sites in the two targeted date ranges (7,657,744 metadata matches). This metadata is deduplicated so that only one piece of metadata per URL is retained for each period (610,000 after de-duplication per URL). The metadata retained is the one with the collection date closest to the middle of each interval (May 1, 2010 and May 1, 2015).

Step 2: Browse the records and extract those that match the metadata records extracted in the previous step. This match is based on the SHA-256 signature of the content present in the metadata and data records.

Step 3: Perform the join (as in SQL) between the data and metadata records and write them in NDJSON format.

During the export process, the INA extracted the main text from each page using the Boilerpipe algorithm (Kohlschütter et al., 2010). An HTML page generally contains different types of textual information, which can be divided into the following categories:

- Main textual content, e.g., the text of an article
- Information related to the main content, e.g., links to similar content
- Information common to pages on the same site, e.g., navigation links to the main sections of the site
- Third-party information, e.g., advertising, social networking elements.

The task of identifying the main content of a web page is known as “boilerplate removal” and is a prerequisite for any textual analysis of data from web pages. It should be noted that for some sites, exporting the HTML of the pages does not always make it possible to obtain the textual content as it was presented on the web. This is particularly true for sites whose pages are generated by JavaScript on the client side (in the browser). With HTML alone, we also miss out on the visual aspect of web pages, as layout elements and visuals are referenced by HTML but not included.

From NDJSON to readable data

The IT specialist was responsible for providing the researcher with an interface (Corpus.Gamuzá) for accessing and exporting data from files originating from the INA's web archives.⁵ He codes in PHP to process and decompress the files. The extraction into NDJSON files is achieved using data loops⁶ derived from the SPIP Content Management System for the Internet, also generated in PHP. He also employs VueJS JavaScript.⁷ This task was conducted in several steps, as visible in the Corpus.Gamuzá interface (see Figure 24.1):

- Decompression of NDJSON files
- Exploration of the decompressed files
- Extraction and display by sections
- An online summary spreadsheet accessible to the researcher

Screens in struggle

Decompression of NDJSON files

For each line in the NDJSON file containing an HTML file, the IT specialist initially decompressed the NDJSON files to enable the interface to display the HTML content and the metadata from the first and last lines. His aim was to retrieve the metadata (date of collection, URL – Uniform Resource Locator, and http response) together with the data from the HTML (main title, subtitle, text) to obtain a unique identifier to which the text extracted from the page is linked, together with the date of collection and the URL. The URL can then be used as an identifier, and if there are multiple versions of the same page, the page is associated with the date to create a unique identifier for our corpus (not included). This decompressing operation encodes the NDJSON into HTML. If we take the example of a page from the La Tele Libre website (see Appendix 1), the .JSON files version groups together the metadata and the HTML text (accessible by clicking on “contenu brut”). From this information, the metadata and the HTML content was extracted from the first line to the last.

Exploration of decompressed files in the interface

As a next step the data scientist created a PHP program that allows the decompressed files to be queried as if there were a database. JSON data is formatted with the same fields because of Jérôme, resulting in standardized and structured datasets. At this stage, the objective is to process the files to extract content. However, the challenge arises when transitioning from many files, as the required processing time can strain machine resources. To facilitate the examination of HTML files in the corpus, he loaded the HTML files into the online interface, selecting one page out of every ten to avoid overwhelming the workspace. Consider the example of the O2zone webTV. For the 450 web pages collected on May 1, 2010, in the INA archive, we can access the HTML file and metadata in addition to the raw content still present at this stage.

This provides a substantial body of material that can be navigated to explore the data while maintaining a view of the raw content. This data examination enables the researcher and the IT specialist to identify information (both data and metadata) suitable for targeted extraction while cleaning out extraneous structural elements (navigation menus, advertisements, footers). Each page is probed through extractions within the HTML tags to gain insight into the file’s content. For instance, it becomes evident here that text enclosed in tags might be of interest for studying the site’s structure but not its editorial content. Moreover, searching for titles in the style of h2/h3 search engines does not work due to a specific HTML structure chosen by O2zone’s developers. Conversely, metadata in the HTML of the page (not from the archiving robot) facilitates the extraction of the title, site description, or keywords.

This step is invaluable as it allows the researcher to manually delve into the HTML of the corpus pages, gaining insight into their structure and available data. Not only does this vary depending on the website, but the HTML structure evolves over time, making this manual approach indispensable for data extraction preparation. For example, in the case of the O2zone website,⁸ we obtain three different periods with three different HTML structures. For instance, it was relevant to scrape the data through h1/h2 HTML title tags before 2011, but not after this year because of the use of a new Content Management System (CMS) to design the website’s structure. So, from the manual identification of the HTML structure

over the three periods of the site's history within the Gamuza.Corporus interface, the data scientist targeted the HTML tags and the Cascading Style Sheets (CSS) selector for each period to improve the quality of the extraction. We can see how he proceeded through the title function he created:

```
"preg_match_all('/ class=" (titre|field-content|nodeTitle  
nobreak)">(<[^>]*>)?(<[^>]*>)/si', $html, $match, PREG_SET_ORDER);"
```

By targeting "titre" and "field-content," the program can extract or exclude the title for the first and second period of the website's history, and "nodeTitle/nobreak" is more adapted to the third period of the website. Therefore, the program enables the selection of three data types, which are adapted to the three different periods and all the O2zone files we explored.

Targeted extraction

In the case of O2zone, the following data is included in the extraction: The entire HTML, titles, trackers, social media links, external and internal links, description, page text, page title, and keywords.

Making data accessible on the online user interface (from HTML to .CSV)

Developed using a dedicated JavaScript component (Vue.js) for creating dynamic tables, the spreadsheet enables researchers to filter, sort, and extract data in .CSV format for further analysis using other tools.

Discussion: Methodological challenges and prospects in the face of diachronicity

After a brief overview of the processing possibilities offered by the extraction, harvesting, and provisioning of textual data, we will see how this project feeds into discussions about the obstacles to automating archived web exploration, given the diversity of formats, but also changes in HTML structure.

Example of data handling

To illustrate the data processing possibilities offered by this first extraction, we will take the example of the webTV O2zone TV. This is an associative webTV site based in Aix-en-Provence (France) and was collected from 2004 until its closure in spring 2017. We have data available from 2010 to 2016. The initial processing I allows researchers to browse in the corpus by doing a remote reading.

For the historian, data export provides the opportunity for an initial exploration, with a view to refining or adding to their corpus. This not only eliminates the need for manual perusal of the web archive, but also facilitates the consideration of data and the identification of elements suitable for diachronic or nondiachronic comparisons. The historian can use the text extracted by data scientists to examine both the site's structural evolution and its content.

Screens in struggle

For example, with the extraction of text into CSV format from the web page on May 1, 2010, and that of May 1, 2015, following data cleansing, semantic dynamics can be seen using Voyant Tools, which is readily accessible online.⁹ Several transformations in the website's structure and content are observed using the word cloud creation tool, Cirrus.¹⁰

Firstly, we observe the evolution of the use of social media networks, which became more prominent in 2015 with the inclusion of hashtags, Twitter URLs, and team members' accounts that relayed information on the platform. This also coincided with the increased recognition of the "O2zone" brand in 2015 compared to the 2010 data, where the term "émission" (show) was prevalent.

However, the most intriguing aspect relates to the evolution of the themes addressed. In 2010, there was a strong emphasis on educational and family-related aspects, with terms like "apprendre" (learn), "family," "lycéen" (high school student), and "vérifié" (verified), suggesting actions related to media and digital literacy for young audiences. In 2015, we observe a diversification of themes while maintaining the educational purpose (represented by the term "éducation"). Additionally, there was coverage of local cultural events, such as cinema, possibly in line with the dynamics brought about by Marseille Provence 2013, the European Capital of Culture ("mp2013" term). Concurrently, terms like "plateaux" (platforms), "urbains" (urban), "forum," "groupe" (group), and "partageant" (sharing) indicate videos giving voice to citizens on urban planning or everyday life issues, aligning with the association's role as "citizen media."

The text also highlights the disappearance or at least a significant reduction in available conversational space on the website. In 2010, terms like "cordialement" (sincerely), "désolé" (sorry), and "enthousiasmé" (enthusiastic) indicate user comments to creators regarding the videos posted. There is no trace of these exchanges in 2015, possibly indicating that the conversation had been externalized to social media networks, relieving the team of website moderation tasks. This shift in the relationship between content creators and the audience is accompanied by two significant changes: The term "amateur," which was prevalent in 2010, disappears, and the target audience appears broader, with the shift from terms related to education to "populaire" (popular) and "citoyen" (citizen), suggesting a change in editorial direction.

These changes can be seen by analyzing semantic links around key terms. In 2010, the term "émission" (show) was linked to emotional and affective vocabulary (constater — observe, heureux — happy, fonctionne — works, etc.). In 2015, central terms like "O2zone" and "tv" are linked to video content and cultural events (reportage — report, mp2013, etc.), while the conversation is associated with team members' Twitter accounts.

This brief example underscores the significance of monitoring the evolution of archived websites and raises questions for further investigation. Data mining generates many questions, prompting a deeper inquiry to clarify the exact nature of transformations and their underlying causes in a qualitative perspective. Moreover, it may lead to an expansion of the corpus, focusing on semantic analysis surrounding the term "amateur" used in 2010 and comparing it with other webTVs.

Contributions and questions from a transnational perspective

In this discussion, we will emphasize the necessity of a collaborative and interdisciplinary approach when studying archived web data within a transnational perspective. We will then

argue for the value of a gradual approach based on a small corpus before discussing the limitations and prospects of automation due to HTML changes.

First and foremost, this experience reaffirms indispensable collaboration required among archivists, researchers, and computer science experts when delving into web archive data in history and the social sciences. This collaborative effort addresses one of the key challenges concerning methods and research based on web archives, particularly “the need for increased accessibility of Web archives and improved scalability of existing tools to leverage the full capacity and depth of Web archives” (Weber, 2020). The case study provides a secure framework for accessing web archive data, a crucial aspect of developing research based on these new materials. This individual initiative aligns with a broader trend of multidisciplinary research collaboration connecting scholars and archive curators. Additionally, examining web archives through three distinct approaches facilitates the conceptualization of tools and methods tailored to researchers. Each participant contributes information based on their area of expertise, thereby enhancing their skills across various disciplines. For instance, historians gain a better understanding of HTML, while archivists develop a finer grasp of the needs of social science researchers, which differ from those in IT research, promoting shared knowledge about web archives (Weber, 2020). From a technical perspective, this collaboration has allowed the INA to enhance its web page export capabilities by developing a new export mode that involves sites defined by URL selection rules and time periods.

Another significant contribution lies in the value of starting with the exploration of a small corpus (small data) as a prerequisite to building computational tools in a “slow” approach to adapt tools for the study of massive data. In this case, the corpus is relatively small, but the archived files still amount to 419 GB, not including visual and audio-visual content hosted on web pages. Therefore, manually navigating through this corpus would be extremely time-consuming. However, our approach systematically involved using a small sample for testing before considering full corpus-wide implementation. This has twofold benefits: It conserves computing power and contributes to the fight against digital pollution, and, additionally, it helps identify collective obstacles. This approach also holds pedagogical value for nonexpert researchers as it allows them to gradually acclimate to the data without being overwhelmed by its volume, reducing the technical entry barrier (Gebeil, 2021). Through this approach, we gained insight into the specificities of these small, sparsely texted, and mostly manually coded websites, often created by amateurs with little regard for HTML standards.

Beyond the challenge of formats, this experience highlights a salient point of discussion relevant to all corpora aiming to study web history from web archive data in an automated fashion: How to automate the extraction of textual data over time despite the instability of HTML structure on each site? In this case, the corpus from 2010 to 2015, comprising small amateur sites with non-standardized structures and code, presents an extreme challenge regarding HTML structure changes from unstructured information. Through the interface, all project participants quickly realized the extent of this challenge: the structure evolves over time and selecting information from web pages had to be continually readjusted for each site, whenever we noticed a structural change. This did not pose significant problems with a small corpus, but it constitutes a major obstacle to automating extraction from non-standard and older websites. In a recent study, Ertel et al. (2023) encountered the same difficulty, adapting their Python-based exploration tool for each change. However, their situation involved less variation and a more recent site, focusing on the year 2019, which reduces the challenges of selecting information judiciously from web page files. This was an institutional site developed

Screens in struggle

from a CMS, which is highly standardized, enabling the use of style sheets as reference points during extraction. As the nature of historical research involves capturing ruptures and continuities over time, this challenge presents an exciting methodological research perspective for web history.

Our targeting strategy, designed in 2020–2021, enabled us to deal with the HTML changes over time, and finally to scrape the data files from the 58 websites in our corpus. A similar method was also used by Ertel et al. in 2023 to analyze a corpus of websites over a three-month period in 2019. However, whether using HTML tags or CSS selectors to target specific text in pages with a consistent structure, this task limits full automation, and thus still requires human manual work to identify the important elements for text extraction for each version of a site over a given period. Ertel et al. (2023) came to the same conclusion: those responsible for this task must also have good knowledge of HTML and dedicate a substantial amount of time.

Based on this experience, two avenues for improvement will be explored. Firstly, to minimize the manual and time-consuming nature of identifying HTML structure changes over time, we envisage improving the tool by introducing a set of “type” filters, identified across the entire corpus at specific dates, which historians can test on their samples. We will provide a type of generic “toolbox” for initial data exploration, combining the use of the HTML tags and CSS selector extraction rules, which will identify specific elements within the web pages.

Secondly, another avenue is to combine the text from HTML files with transcriptions of video voices from these webTV sites. The INA has provided video metadata, along with transcriptions of video voices. Therefore, it is possible in the future to associate the text from HTML pages with video transcriptions for each version of the archived website. This would create an incredibly rich text corpus, enabling exploration of video content in relation to the text on the web page and enhancing the remote analysis of the site’s history. Similarly, the cataloguing of videos could also allow for the use of a visual exploration tool for videos enriched by HTML text data.

Conclusion

To explore marginalized websites archived by the INA between 2010 and 2015, the approach adopted in this experiment has resulted in the creation of a pipeline and an interface that provides extracted data to researchers. It aligns with the discussion on research methods based on web archives, contributing to the discourse on transnational web history and opening new avenues of study.

Web archives serve as a productive trading zone for the development of digital humanities. They allow for the creation of a shared practice space and interdisciplinary exchanges, as seen in the three-way pipeline involving archivists, IT specialists, and historians. This approach prompts researchers to consider the reproducibility needs of tools and processes developed by both the INA and Gamuza. Starting with a strong core on a small corpus has allowed for the evolution of the available tools, aligning with Weber’s observations (2020): “In order to continue developing Web archiving as a method there is a need to build intersections with existing domains of research to expand the applicability of the research method.” This approach also helps reduce the energy cost of experimentation. Furthermore, transnational studies based on web archives are affected by the presence of alternative formats alongside the predominant WARC format. By proposing a conversion process to NDJSON, our experiment contributes to considering the study of data from different archiving organizations. It

is now possible to insert DAFF files alongside WARC or other files into the database, facilitating the emergence of transnational corpora.

The approach and programs developed here, in support of a micro-historical approach based on web archaeology (Brügger, 2018, pp. 54–57) and the evolution of media discourses in a specific web sphere, are shared and can be reused for other inquiries. This includes studying the concepts of websites, considering the impact of evolving HTML specifications on how sites are constructed.

The proposed pipeline can facilitate a transnational approach by extracting data from web archives, encompassing both webTV channels and other sources. This example underscores inherent barriers to adopting transnational perspectives. Firstly, discussions revolve around tool interoperability. While the predominant format may be considered a standard, the diverse array of alternative formats also requires consideration. Consequently, we demonstrate that similar approaches used with WARC can be applied to alternative formats such as DAFF, highlighting that interoperability lies in the extraction process, which should be reproducible across different formats. This entails specific program development based on the existing work. Secondly, reproducing extractions from webTV corpora across various national domains, regardless of format, provides a dual perspective. On one hand, it enhances the identification of national or local specificities in the web of the 2000s, uncovering cultural circulations, such as those related to alter-globalist webTV. On the other hand, applying the pipeline to a corpus composed of other websites better addresses the challenge of mining evolving HTML, as demonstrated here. Tackling this challenge involves building a chronological directory of HTML structures used at a given moment by creators of each investigated site to improve the extraction of HTML data from archived web content. Therefore, this modest interdisciplinary and cross-category experience opens up avenues for a transnational perspective both historically and methodologically.

The convergence of these methodologies lays the groundwork for further exploration of validity indicators for the developed tools, marking a crucial area of research for the future of transnational studies.

Acknowledgements

This work was conducted as part of the associate researcher contract offered by the INA (French National Audiovisual Institute) for the “Screens in Struggle” project (2019–2021) proposed by Sophie Gebeil. This chapter focuses on a case study carried out between 2019 and 2021 by Sophie Gebeil (historian), Jérôme Thièvre, head of the web archive team at the National Audiovisual Institute (INA), and Cyrille Le Menestrel, an independent IT specialist. The selection of the sites to be included in the spreadsheet was made in collaboration with Boris Blanckemane, DL Web manager.

Notes

- 1 Solr Wayback, <https://github.com/netarchivesuite/solrwayback>.
- 2 The Archives Unleashed Toolkit, <https://github.com/archivesunleashed>.
- 3 NDJSON <https://web.archive.org/web/20230503180730/http://ndjson.org/>.
- 4 Captation, Analyse et Représentation Temporelle des Évolutions d’un Corpus — CARTEC project, French national Agency, Project ID: ANR-07-MDCO-0016.
- 5 Traitement Corpus Web, Gamuza, <http://gamuza.net>.

Screens in struggle

- 6 Data Loops: https://web.archive.org/web/20221005224357/www.spip.net/fr_article5444.html#BOUCLE-DATA-.
- 7 The code is available online: Gamuza, Outil Corpus, Web function, Github, line 174, https://git.gamuza.fr/Outils_archives/corpus_web/src/branch/master/corpus_web_fonctions.php#L174.
- 8 O2zone.tv, for instance: <https://web.archive.org/web/20100715022143/www.o2zone.tv/>.
- 9 Text from archived O2zone web page of May 1, 2010 inserted in Voyant Tools: <https://voyant-tools.org/?corpus=bec8b07467e544d3ec2bac2574cd8247>; Text from archived O2zone web page of May 1, 2015 inserted in Voyant Tools: <https://voyant-tools.org/?corpus=7417664489491cc5723c86094116eba6>.
- 10 Most frequent word cloud from the text of the 01/05/2010 version: <http://web.archive.org/screenshot/https://voyant-tools.org/?corpus=bec8b07467e544d3ec2bac2574cd8247&stopList=keywords-1d60d8f8eb8b485c27f7772913788488&visible=85&view=Cirrus>; Most frequent word cloud from the text of the 01/05/2015 version: <https://web.archive.org/web/20240320105709/http://web.archive.org/screenshot/https://voyant-tools.org/?corpus=7417664489491cc5723c86094116eba6&stopList=keywords-c68a17c16a0b010b23e92e0a5ac16ce5&visible=85&view=Cirrus>.

References

- Brügger, N. (2018). *The Archived Web: Doing History in the Digital Age*. MIT Press.
- Carion, F. (2009). Nouvelles technologies et altermondialisme. *Terminal*, 103–104. <https://doi.org/10.4000/terminal.2849>.
- Carpentier, N. (2011). *Media and participation: A site of ideological-democratic struggle*. Intellect. The University of Chicago Press.
- Christian, A. J. (2015). *Web TV Networks Challenge Linear Business Models*. Carsey-Wolf Center Media Industries Project.
- Ertel, F., Donig, S., Eckl, M., Gassner, S., Göler, D., & Rehbein, M. (2023). Using web archives for an explorative study of the web presence of German parties during the European election 2019. *Quality & Quantity*, 58, 603–625. <https://doi.org/10.1007/s11335-023-01654-3>.
- Ferron, B. (2016). Professionnaliser les “médias alternatifs”? Enjeux sociaux et politiques d’une mobilisation (1999–2016). *Savoir/Agir*, 38, 21–28.
- Gebeil, S. (2021). *Website Story: Histoire, mémoires et archives du Web*. INA.
- Kohlschütter, C., Fankhauser, P., & Nejdil, W. (2010). Boilerplate detection using shallow text features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM ’10* (pp. 441–450).
- Lohndorf, J. (Updated 1 year ago). Derivative data sets. Retrieved from <https://web.archive.org/web/20231218141441/https://support.archive-it.org/hc/en-us/articles/360013691371-Derivative-Data-Sets>.
- Pigenet, M., & Tartakowsky, D. (2014). *Histoire des mouvements sociaux en France: De 1814 à nos jours*. La Découverte.
- Rigoni, I. (2010). Éditorial. Les médias des minorités ethniques. Représenter l’identité collective sur la scène publique. *Revue européenne des migrations internationales*, 26(1), 7–16.
- Ruest, N., Fritz, S., & Milligan, I. (2022). Creating order from the mess: web archive derivative datasets and notebooks. *Archives and Records*, 43(3), 316–331. <https://doi.org/10.1080/23257962.2022.2100336>.
- Ruest, N., Lin, J., Milligan, I., & Fritz, S. (2020). The archives unleashed project: Technology, process, and community to improve scholarly access to web archives. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL ’20)* (pp. 157–166). <https://doi.org/10.1145/3383583.3398513>.
- Rygiel, P. (2005). Des archives numériques sans historiens? Un point de vue. *Matériaux pour l’histoire de notre temps*, 79, 11–13. <https://doi.org/10.3406/mat.2005.1033>.
- Wang, X., & Xie, Z. (2020). The case for alternative web archival formats to expedite the data-to-insight cycle. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (pp. 177–186). <https://doi.org/10.1145/3383583.3398542>.
- Weber, M. S. (2020). Web archives: A critical method for the future of digital research. *WARCnet Papers* ISSN 2597-0615. Aarhus.

Tools

Commoncrawl, <https://web.archive.org/web/20240104223757/https://commoncrawl.org/get-started>.
IETF, RFC editor, <https://web.archive.org/web/20220308004747/https://datatracker.ietf.org/doc/html/rfc7089>.
GLAM Workbench, <https://glam-workbench.github.io>: Tim Sherratt. (2021). GLAM Workbench (version v1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.5603060>.
Cykelero, <https://github.com/Cykelero/timefind>

Appendix 1: Latelelibre.fr NDJSON

Appendix1: Latelelibre.fr NDJSON

Two examples:

- latelelibre.fr_example.jsons: an extract of the first 50 json elements of the 2015 version of the Latelelibre website in "jsons" format.
- latelelibre.fr_example.json: an element of the latelelibre 2015 website in json format, with indentation and line breaks to make it easier to read.

```
/* emplacement et identifiant du record de donnée dans notre stockage */
"dataId": {
/* sha256 de la donnée, sert d'identifiant unique pour chaque donnée */
"contentId":
"ee601d865509c75dc911378d6e9fcf1fd5bbf0a1cdd466bc3e3ded306da6b153",
/* identifiant du fichier dans lequel est stockée la donnée */
"fileId": "C22394D0-E9A8-11E4-BA46-755A51147092",
/* position de la donnée dans le fichier de stockage */
"fileOffset": 6629885481
},
/* metadonnées de collecte */
"metadata": {
"date": "2015-04-23T13:49:08Z",
"charset": "UTF-8",
"referer_url": "http://latelelibre.fr/archives/rubrique/journal/tribune/page/2/",
"agent": "Mozilla/5.0 (compatible; INA dlweb; +http://www.institut-national-audiovisuel.fr/collecte-depot-legal-web.html; robot.dlweb@ina.fr)",
"client_lang": "fr, en",
"level": "4",
"sgebeil.site_id": "latelelibre.fr_2015-05-01T00:00:00Z",
"ip": "92.243.0.201",
"length": "45442",
"active": "1",
```

Screens in struggle

```
"type": "text/html",
"url": "http://latelelibre.fr/2007/02/01/des-bourdes-gaffes-et-
autres-boulettes-en-politique/",
"content": "ee601d865509c75dc911378d6e9fcf1fd5bbf0a1cdd466bc3e3
ded306da6b153",
"crawl_session": "latelelibre.fr@20150422T232302Z",
"original_url": "http://latelelibre.fr/2007/02/01/des-bourdes-
gaffes-et-autres-boulettes-en-politique/",
"client_country": "fr",
"client_ip": "10.3.209.54",
"page": "1",
"inline_level": "0",
"status": "ok"
},
"extractionMetadata": {
  "charset:html": "none",
  "text:size": "45019",
  "html:parse:duration": "1",
  "charset": "UTF-8",
  "charset:origin": "detected",
  "html:extract:duration": "1",
  "content:size": "45442",
  "contentId": "ee601d865509c75dc911378d6e9fcf1fd5bbf0a1cdd466bc3
e3ded306da6b153",
  "charset:detected": "UTF-8",
  "html:text:duration": "3"
},
"extractionContent": {
  "htmlmeta:og:type": "article",
  "htmlmeta:article:tag": "Chirac,gaffe,Iran",
  "htmlmeta:twitter:site": "@la_tele_libre",
  "htmlmeta:og:locale": "fr_FR",
  "htmlmeta:html:lang": "fr-FR",
  "htmlmeta:article:section": "Le Journal",
  "htmlmeta:twitter:description": "La nouvelle « Bécassine » ...",
  "htmlmeta:og:description": "La nouvelle « Bécassine » ...",
  "htmlmeta:link:canonical": "http://latelelibre.fr/2007/02/01/
des-bourdes-gaffes-et-autres-boulettes-en-politique/",
  "htmlmeta:og:url": "http://latelelibre.fr/2007/02/01/des-bour
des-gaffes-et-autres-boulettes-en-politique/",
  "boilerpipe:text": "Publié le\nLa nouvelle « Bécassine » ...",
  "htmlmeta:article:published_time": "2007-02-01T15:49:28+00:00",
  "htmlmeta:twitter:title": "DES BOURDES, GAFFES ET AUTRES
BOULETTES EN POLITIQUE - LaTeleLibre.fr",
  "htmlmeta:twitter:creator": "@la_tele_libre",
```


The Routledge Companion to Transnational Web Archive Studies

```
"htmlmeta:og:site_name": "LaTeleLibre.fr",  
"htmlmeta:og:title": "DES BOURDES, GAFFES ET AUTRES BOULETTES EN  
POLITIQUE - LaTeleLibre.fr"  
}
```

25

TOWARDS TRANSNATIONAL RESEARCH DATA MANAGEMENT PRACTICES FOR WEB ARCHIVES

Challenges and possibilities

*Eld Zierau, Beatrice Cannelli, Sally Chambers, Olga Holownia,
Sharon Healy, Ditte Laursen, Susan Aasman, and
Ulrich Karstoft Have*

Introduction

Websites and social media are increasingly being archived as part of the public record as a method for documenting contemporary society and as historical sources. Consequently, such born-digital materials are being analysed in a wide range of research domains including the humanities and social sciences, computer science, library, and information sciences, etc. (Vlassenroot et al., 2019; Healy et al., 2022). As born-digital resources are by definition data, good research data management practices are crucial. Nonetheless, research in this domain is difficult to find. That is why this chapter addresses the questions as to how researchers use data management practices when dealing with web-related sources and in addition, how archives can support such data management practices. By exploring these issues, this chapter will contribute to filling a void through an examination of current research practices for this type of data, including an exploration of the type of challenges faced by researchers for the application of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles to archived web data. The methodology includes a series of semi-structured interviews that were undertaken with both researchers who use web or social media archives for their research and cultural heritage institutions who would like to improve access to their born-digital archives for research. Along with findings from these interviews, this chapter is also based on a literature review and the authors' own experiences developed both in research and archival practices.

The chapter is divided into the following sections. After the introduction, the theme of *research data management for web archive studies* is introduced. Following this, there is an exploration of current *data management practices* related to web archives, from both

a researcher and archival perspective, using the FAIR Data Principles (Wilkinson et al., 2016) as a framework. The *discussion section* provides an overview of these investigations, examining progress and where the remaining challenges lie. In *conclusion*, the challenges of research data management for the *transnational* study of web and social media archives are highlighted, including an analysis of where the responsibilities of research data management lie, as well as possible avenues for future research. Finally, the ongoing importance of collaboration between international research and archival communities, such as WARCnet, European Open Access initiatives, International Internet Preservation Consortium (IIPC), and Archives Unleashed is emphasised.

Approach and methodology

To get a better understanding of the current research practices and data management challenges for archived web and social media resources, we adopted a three-fold approach. First, we conducted a review of the existing literature concerning data-centred research projects and research data management practices related specifically to web and social media archived data. Relevant references have been distributed throughout the chapter. Then, as some of the co-authors of this study had either previously worked, or were working with the archived web at the time this research was conducted, we decided to capture some of the considerations emerging from their direct experience with this type of resource. Finally, a series of three semi-structured interviews were carried out with a researcher working with historical home pages, a member of a research project related to archiving social media data, and a representative of a cultural heritage institution who provides access to such important resources.

The interviews were carried out online via video conference between the end of 2022 and April 2023. Participants were provided with a set of semi-structured and open-ended questions before the scheduled interview to allow them time to reflect on all aspects involved in the research data management lifecycle of web-archived material. The set of questions compiled for this study was articulated in seven sections, corresponding to the phases identified in the web-archived data lifecycle. Each question was accompanied by a few prompts to facilitate responses and spark additional considerations. Participants were asked, first of all, to briefly provide an overview of their research project to set the context in which web archives were being used. This was followed by questions regarding access to and challenges encountered when accessing web archives; the second section revolved around the creation of datasets (e.g. type of data collected, organisation strategies, and any additional data that had been created in relation to the one collected); then, participants were prompted to describe management practices before, during and after the research process. Their experience using a data management plan and any challenge that emerged from this; the interview then focussed on data citation practices and challenges related to citing web archived material, sustainability, and referencing issues; next, participants were asked to reflect on ethical and legal challenges that may affect their research using web archives; the final two groups of questions focussed on tools used to manage web archived data, any technical issues that participants might have experienced, practices and problems related to long term preservation related to their research data, concluding with a reflection on preservation challenges they expect to arise in 20 years. Data from the interviews were then examined taking the FAIR principles into account.

Research Data Management for web archive studies

Since the last decade, Open Science, Research Data Management (RDM), and Data Management Plans (DMP) are increasingly becoming an integral part of research practice. For researchers working with digital sources or data, such as web and social media archives, a good understanding of the Research Data Management Lifecycle is essential. Data-centred researchers using web archival materials have already started addressing the need for a systematic workflow (see Brügger, 2018; Brügger and Laursen, 2019; Milligan, 2019). An example of a current RDM Lifecycle has been provided by the UK's Joint Information Systems Committee (JISC) as part of their Research Data Management Toolkit (2018). The toolkit aims to support researchers throughout the entire RDM lifecycle by ensuring that their research data is managed, stored, shared, and preserved safely. The Research Data Lifecycle (see Figure 25.1) is used as a framework for structuring the toolkit.

In JISC's research data lifecycle, six key steps of research data management have been identified: (i) plan and design, (ii) collect and capture, (iii) collaborate and analyse, (iv) manage, store, and preserve, (v) share and publish, and (vi) discover, reuse and cite. These are linked to four core stages: (a) data creation and deposit; (b) managing active data; (c) data repositories and archives, and (d) data catalogues and registries. The toolkit details

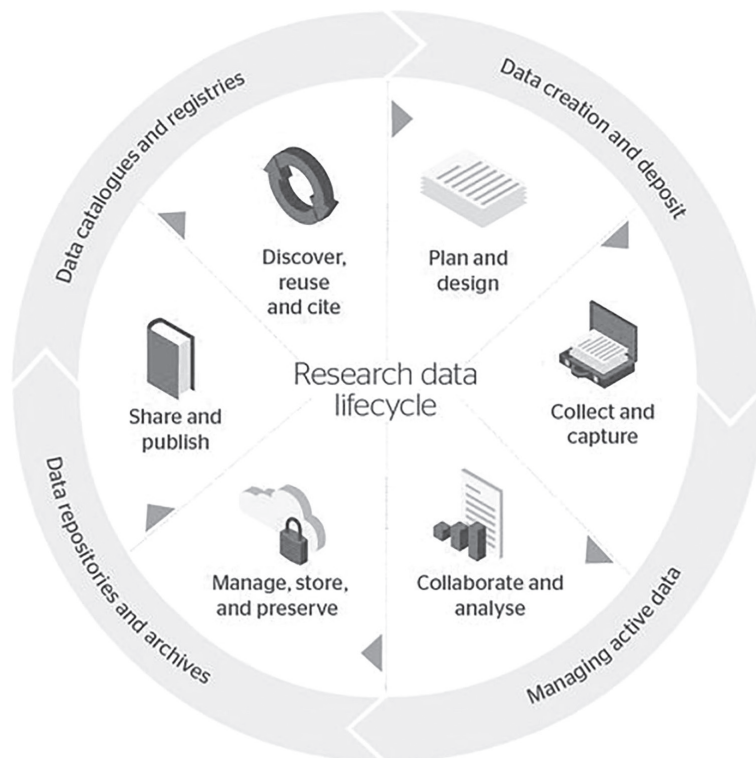


Figure 25.1 A visualisation of the Research Data Lifecycle from JISC's Research Data Management Toolkit.

the activities that need to be undertaken during each step of the process. For example, the “plan and design” phase further breaks down into data management planning, policy compliance, FAIR principles in research data management, and costing. Furthermore, the Research Data Management Toolkit (2018) provides a downloadable Research Data Management Checklist, to help researchers in their research data management planning.

Additionally, there is an increasing variety of guidelines and good practices for Research Data Management, such as OpenAIRE Guidelines on How to make your data FAIR (OpenAIRE, n.d.), for example, particular communities and the DARIAH Pathfinder to Data Management Best Practices in the Humanities (Tóth-Czifra, 2019). Many universities also provide institutional RDM support, such as Ghent University’s Research Data Management team (Ghent University, n.d.). Research Data Management is also further supported by initiatives in several European countries through tools such as DMP Online¹ provided by the UK’s Digital Curation Centre (DCC, n.d.), which assists researchers in the creation, reviewing, and sharing of their Data Management Plans (DMPs) in line with institutional and funder requirements.

However, when it comes to web and social media archives particularly, there is currently a lack of specific guidance on Research Data Management for researchers using web and social media archives in their research. It is this gap that this chapter is intending to address.

Findings: Data management practices

To understand in detail what the state of the art is in terms of digital management of web and social media archives, the authors undertook a series of semi-structured interviews that were carried out with both researchers working with web archive data and cultural heritage institutions providing access to such data, supplemented with insights from the authors as required. These findings cover data management practices from both a researcher and an archival perspective. From the researchers’ perspective, it is a question about how data management practices are used for research, while the archival perspective is a question about how archives can support data management practices.

The findings are structured following the FAIR Data Principles. Originally introduced in 2016, the FAIR Data Principles were established to advocate for the improvement of the infrastructure to support the reuse of scholarly data (Wilkinson et al., 2016). Since then, they have become a core component of Research Data Management best practices and are becoming increasingly used in a variety of domains.

To date, there is little understanding as to how these principles relate specifically to research using web and social media archives. As a first step towards addressing this challenge, the question of “How FAIR are our web archives?” was explored within the WARCnet Community (Chambers, 2020). Each FAIR data principle is taken in turn and then applied to web archives which serves as an examination of how FAIR the found data management practices are, as outlined below.

Findable

Findable: *For data to be findable there must be sufficient metadata; there must be a unique and persistent identifier; and the data must be registered or indexed in a searchable resource.*

(Wilkinson et al., 2016)

Findability, in terms of web and social media archives, has several aspects. Firstly, regarding *sufficient metadata*, the WARC (Web ARChive) Standard (ISO 28500:2017, 2017) enables metadata about the URLs harvested for the web archive to be included in the WARC files and *indexed* into searchable CDX (Internet Archive, n.d.) or SOLR (Apache Software Foundation, 2023) indexes. The use of the WARC Standard by web archives is an important first step to ensuring that sufficient metadata is provided. Furthermore, there are metadata standards available specifically for web archives. For example, the recommendations for *Descriptive Metadata for Web Archiving* from the *OCLC Research Library Partnership Web Archiving Metadata Working Group* (Dooley and Bowers, 2018).

Discoverability is another key aspect of findability. This relates to the recommendation that the data must be registered or indexed in a *searchable resource*. For example, the Dutch National Web Archive Registry,² provides an overview of which websites in the Netherlands have been archived, where, since when, why, how, and how often. Other examples include the cataloguing of web archive datasets in a library catalogue, such as the case of the UK Web Archive's archived web collection about the Ebola Outbreak in West Africa in 2014 (UK Web Archive and Bodleian Library, 2014) which has been catalogued in the British Library catalogue.

However, the use of *unique and persistent identifiers* is more challenging. In the case of one interviewee, the National Web Archive did not include the web sources the researcher needed. It was also not possible for the researcher to ask for the resources to be archived either, as they were outside the scope of the web archive, e.g. the web archive in question did not archive websites of private individuals due to legal and ethical reasons. In this case, the researcher needed to self-archive the required websites. The archived versions were then stored in an ordinary university storage system, and therefore, not made available as a searchable resource to be accessed by other researchers. Furthermore, the archived websites were also not assigned unique and persistent identifiers. The challenge with persistent identifiers is described in further detail in the FAIR Reusable section below.

In another case, a researcher experienced that the web archive URL references that had been included in a research paper became invalid overnight. This was the case for research referencing web materials in the National Library of Ireland. This happened because the Library migrated its holdings from an Internet Memory Foundation (IMF) platform³ to an Archive-IT web archive service⁴ (Aturban, 2019a, 2019b; Aturban et al., 2021). All the old archive URLs broke, which is an example of why archived URLs cannot be regarded as persistent references.

Using web archives, particularly for research purposes, requires quite a high level of digital literacy. For example, initially, many web archives did not provide a full-text search option. It was, therefore, only possible to search the web archive by URL. As a result, providing training for researchers on how to use the web archive becomes key. In a third case, the interviewee was offering services to researchers, where they taught researchers, e.g. how they could find web materials in their web archive by having the URL and date. However, more recently, new tools to make web materials in web archives searchable are now available, e.g. by full-text search and filtering functions, as it, for instance, is the case for many of the web archives using SolrWayback.⁵ This makes archived URLs findable if you know the archive, the URL, and the archival timestamp.

Within WARCnet, some researchers have used the Persistent Web IDentifier (PWID) (Zierau, 2022) as a Uniform Resource Name (URN). The PWID consists of the information that researchers are taught to use when finding an archived URL, with extra information

added about whether it is a page or just the file that is referenced. One of the researchers used PWIDs for referencing archived web pages that have been part of the research (Meyer and Zierau, 2019), while the other researcher undertook research based on statistics on large web archive corpora, which were then documented by listing all web archive files in the corpus by PWIDs (Zierau and Møldrup-Dalum, 2021). Further details about PWIDs are available in the FAIR Reuse section.

Accessible

Accessible: *To be accessible, metadata and data should be readable by humans and by machines, and it must reside in a trusted repository.*

(Wilkinson et al., 2016)

When it comes to the accessibility of web archives, two aspects need to be considered based on FAIR. The first one concerns the matter of making archived web information accessible through the use of metadata and data which should be “readable by humans and by machines.” Interviewed web archives observed that due to the large amount of data collected, time constraints, and often lack of resources to adequately support metadata enrichment activities, they tend to retain the metadata automatically extracted as “some metadata is better than none.” However, a few studies have highlighted how web archive users would benefit from more documentation and enrichment of metadata, as it is essential to facilitate access and research using web-archived material (Costea, 2018; Healy et al., 2022).

The second FAIR element concerning access involves archived web data residing in a trusted repository, looking at ensuring long-term access that can only be achieved through proper digital preservation. Concerning bit preservation which ensures that the bits forming the web material remain intact and readable (Lavoie and Dempsey, 2004), it is worth noting how there seem to be various levels of long-term bit preservation in web archives, so much so that the web material is only as safe as the level of bit safety existing in the repository that it is placed in. Furthermore, there is also an aspect of the archives’ sustainability over a long period. Web archives differ in both regarding the mandate and sustainable funding, where the sustainability of national web archives with a mandate based on a legal deposit act may be more sustainable than web archives based on project funding (ISO 690:2021, 2021) Appendix B. For some web collections, especially those created for pilot preservation projects, often there are no specific long-term preservation strategies to ensure access to the archived material: due to the legal framework within which these datasets were created, data cannot be reused after the end of the research project and, consequently, no specific provisions have been made to make that content available to the public neither in the short nor in the long term. Nevertheless, where proper long-term strategies have been put in place to ensure bit preservation, access may be provided as long as there are also plans for logical preservation, which must ensure that the bits remain interpretable. On this subject, there are large international communities that have been working on creating emulators for old web browsers that can be connected to the existing web archives. This is the case of the Oldweb. Today’s project,⁶ an emulator that allows users to browse live or archived web pages visualising them as they would look like in the old web, on specific browsers at a chosen point in time. Furthermore, there are ongoing efforts at the international level to address the challenges of using and getting emulators for browsers and deprecated formats (Cochrane et al., 2022). The strategy for emulation will depend on the continuity of this work.

One of the primary challenges that emerged from the interviews in terms of access is the fact that to use web archives, researchers have to physically visit the buildings of the archiving institutions. Due to national legal frameworks, many web archives are only accessible onsite through designated terminals in designated reading rooms, with some institutions only providing access to accredited researchers. This is due to restrictions put in place to protect individuals' data protection, copyright, and third-party interests, representing, however, a great challenge for end-user researchers in terms of access and use of web-archived material (Truter, 2021). Recent studies have indeed highlighted how these restrictions constitute important geographical and socio-economic barriers not only to researchers who have to travel to designated locations to access the required data (Healy et al., 2022; Maurer, 2022), but also present challenges for the use of web archives as teaching resources (Winters, 2020). Moreover, the diversity of national legal frameworks concerning web archives that preserve the web under national electronic legal deposit legislation presents additional challenges to the sharing of data across borders and its reusability. However, the Royal Danish Library in Denmark (Laursen, 2020) has developed a legal agreement for providing research access to archived web materials, which could be an important step in this direction.

Among the issues surrounding the accessibility of web-archived content, there is also a lack of understanding and, consequently, a need for ad hoc training on how to approach such digital resources. This also includes understanding the difference between archived web pages where there can be missing information or temporalities and archived snapshots of a web page which give a more precise look but can miss information hidden in the HTML as link addresses. Findings from the WARST survey (Web Archives — Researcher Skills & Tools Survey) study showed that users of the web archives would find valuable introductory training courses on web archiving or additional guidance on methods and approaches to carry out research using web archives (Healy et al., 2022). However, only some of the web archives participating in this study offer training to some degree. For example, the Portuguese web archive, which is also openly accessible online, has made available on their website a series of modules and video courses to support research of web archived resources.

Creating corpora and datasets from web archives is an ongoing effort to increase (transnational) access, especially for research purposes. There are notable examples of transnational collections such as the Novel Coronavirus IIPC Collection which is freely accessible via Archive-It (see Geeraert et al., What can we learn from URLs?). With a growing number of derived datasets and information about those datasets, the findability and discoverability of existing datasets remain a challenge. Zenodo⁷ a popular open repository has emerged as a valuable platform for hosting and sharing information about web archive datasets. The Archives Unleashed project, adopted Zenodo as a platform to document almost 40 derived datasets that had been prepared for the datathons organised for researchers.⁸ Alongside this, for computational access to web archive data, the Archives Unleashed Notebooks (n.d.)⁹ and the Web Archives Notebooks¹⁰ as part of the GLAM Workbench, are particularly intended to encourage reuse.

Some libraries now publish citable datasets — either created by web archivists or results of collaborative research projects — through their online repositories. These datasets range from a selected period of a domain crawl (e.g. UK Web Archive) to data extracted from specific collections to metadata about archived web objects (e.g. Library of Congress¹¹) to thematic datasets, such as the National Library of the Netherlands' special Web Collection Chinese Netherlands (2019)¹² and Lin et al. (2019) or the 2019 European Elections¹³ from the Luxembourg Web Archive at the National Library of Luxembourg. The creation of

such web archive datasets often takes place in the context of GLAM (Galleries, Libraries, Archives, and Museum) Labs,¹⁴ such as the Web Archive Austria,¹⁵ part of the Austrian National Library's ONB Labs¹⁶ initiative, which has been archiving the "Austrian webspace" based on domain, thematic and event-based crawls since March 2009. ONB Labs primarily provides access to the metadata about which websites are archived as part of the Austrian National Library's "ongoing crawls." These "seed lists"¹⁷ are available in JSON format from the ONB Labs website.

It is important to mention, as revealed in the interviews, that most web archives do not currently offer guidelines for publishing datasets, and the large corpora created, e.g. language analysis research projects are not available at the moment. Due to the complexity of the content in web archives and restrictions imposed by the legal deposit legislation, collaborative research projects typically take a long-term approach involving technical experts working closely with researchers to extract meaningful data from domain crawls or selected web archive collections. Two notable examples of such projects include Néonaute (Aubry et al., 2018), which focuses on seven years of news sites from the National Library of France legal deposit collections, and the analysis of word changes over time using the JISC UK Web Domain Dataset (1996–2013) from the UK Web Archives (Tsakalidis et al., 2021). In addition, the ARCH (Archives Research Compute Hub)¹⁸ — a platform developed by the Archives Unleashed Project and the Internet Archive that offers 12 different types of derivatives from collections hosted by Archive-It, could be an important move towards FAIR-fication of web archive datasets.

An important development concerning web archive datasets is the development of models and checklists that can be implemented to make these datasets more FAIR. The Archives Unleashed Project has developed the "Web ARCHival Collection Description Model" with three important stages leading "from the unprocessed mass of WARC files, to the more orderly and usable derivative file formats, and finally to these interactive computational notebooks" (Ruest et al., 2022). "A Checklist to Publish Collections as Data in GLAM Institutions" (Candela et al., 2023) could also be adopted by web archives that create web archive datasets.

Interoperable

Interoperable: *Data must share a common structure, and metadata must use recognised, formal terminologies for description.*

(Wilkinson et al., 2016)

Interoperability relates to structuring data in such a way that it can be easily exchanged and interpreted by others. It is about developing a "common understanding" between similar types of data sources to ensure their comparability. This includes using a *common structure* for the data, for example, as formal standards for knowledge representation and use of controlled vocabularies. Interoperability for web archives can concern various aspects, such as search, analysis, and the possibility to reference persistently across archives, which will make it possible to create transnational collections.

The basis for most interoperability among web archives is the use of the ISO (International Organization for Standardization) Standard for the WARC File Format (ISO 28500:2017, 2017) plus the related WARC Implementation Guidelines (Oury, 2009). The WARC standard is for archiving web information along with metadata collected by the crawlers.

However, as pointed out by Maurer (2022): “handling the raw WARC files [is] difficult for all parties, ...” Reasons for this is the varied use of WARC, which can be caused by e.g. curatorial and/or technical constraints (Brügger, 2021b, p. 217).

In addition to WARC, there are other technical standards such as CDX, for providing a list of URLs indexed in a particular web archive or SOLR indexes. Such indexes can be used for access interfaces, search facilities, and offering of API services, e.g. CD XML API / CDX Server API (Truman, 2016). There are also several tools based on web archive meta-data that can support different ways to support interoperability. A well-known example is the Memento protocol (Jones et al., 2021), which provides time-based versioning of references to web archive web pages. However, it works best for web archives accessible on the live web. Furthermore, an example of a method to summarise provenance information for comparison between open as well as closed web archives is the initiative to extract general information from web archives. This study is described in detail in the chapter “Comparing the holdings of closed national web archives through summaries” (Nielsen et al., 2024) of this book.

Researchers often have specific research questions they want to examine, which need to be based on corpora of web material from several archives (see Milligan, 2019; Brügger and Schroeder, 2017; Brügger and Laursen, 2019). There are examples of workflows for extracts, where tools to examine a corpus can be provided for researchers. This was the case in the Archives Unleashed project using the Archive-It platform¹⁹ (Schafer et al., 2024), and is also the case for a Danish researcher using the Danish Web archive (Have, 2020). However, interoperability across web archives is still a challenge, due to e.g. legal, ethical, and technical challenges (Maurer, 2022; Brügger, 2021b). To overcome these challenges, a research infrastructure should be created between the web archives with the data and the research teams wishing to use the data (Brügger, 2021b).

Reusable

Reusable: *Data and collections must have clear usage licenses and clear provenance, and meet relevant community standards for the domain.*

(Wilkinson et al., 2016)

As outlined in the FAIR Principles, the goal of FAIR is to optimise the reuse of data. This includes clear licenses for the data; detailed provenance information and the use of domain-relevant community standards. Additionally, the original resource must be preserved and persistently referenceable to ensure reusability. Although it can be argued that preservation is part of FAIR Accessible and that persistent identification is part of FAIR Findable, the long-term perspective is that both are essential as part of FAIR reusable. To date, only a few examples of the reuse of web archive data for research exist, including Sherratt and Jackson (2021), Brügger (2021a, 2021b), and Eldakar and Holownia (2022).

Concerning clear usage and provenance, this is a general challenge for web researchers, concerning the lack of guidance and training provided for those undertaking research in web archive studies (Truter, 2021). Furthermore, lack of RDM training for researchers about managing web materials through the research data lifecycle, including reusability (Jackson, 2015). This is also one of the consequences of the challenges with the volume of data as well as the complexities of different media types and formats.

Regarding sharing archived web materials for reuse, one of the main challenges is legal restrictions, which include copyright and third-party ownership, privacy policies, and GDPR

(Truter, 2021; Winters, 2020). For some national web archives, it means that they are not open to researchers who are not citizens of the country. For creation as well as analysis of transnational datasets, this is, therefore, even more challenging (Truter, 2021).

Researchers also face difficulties arising from ethical concerns. According to Maemura (2018), ethical challenges stem from the “ethical ramifications of material utilization,” as well as concerns regarding “consent” and the researcher’s obligation to the individuals portrayed in the data. Ogden et al. (2022) propose that researchers must exercise caution when utilising web archives for studying socially marginalised groups. They emphasise the significance of contemplating “how specific vulnerabilities might worsen over time when linked to individuals — such as in the case of studying minors, communities dealing with self-harm, or identifying past instances of illicit activities.” The CARE Principles for Indigenous Data Governance (Carroll et al. 2021), which address key ethical issues such as data sovereignty and data governance, can provide useful guidance. An example from the interviews was that the researcher found it unethical to harvest personal websites into the Internet Archive,²⁰ as the personal information would be exposed in a much larger context and over a much longer period than was originally intended. However, at the same time, the lack of archival copy would mean that this historical source would disappear as a result of the closure of the internet service provider where the websites were hosted.

The basic requirement for reuse is the long-term preservation of archived material, which is true for any studies involving references to web materials. Many studies have shown that the web is transient, where links to live web pages stop resolving or resolve to content that has changed since referencing (Liebler and Liebert, 2012; Nyvang et al., 2017), which is why web archives with preservation are essential for providing reusable web research data.

First of all, researchers need to make sure that they can reference their archived web materials. In some cases, we saw that there were researchers who only had the possibility to store the web material in open web archives, struggling with ethical issues, as the web pages were in some cases highly personal. Other researchers had the opportunity to have web material included in closed archives (e.g. Meyer and Zierau, 2019). This does not necessarily need to be in a web archive but could be in other research data management systems with preservation functionalities. Large-scale initiatives such as the European Union’s Open Science policy (European Union, n.d.) and the European Open Science Cloud (EOSC) (European Commission, n.d.) for sharing publications, data, and analysis tools across borders and scientific disciplines can be useful here and are intended to stimulate reuse.

Secondly, the researcher must consider the possibility of long-term access to the archived version. For web archive material, this depends on how well a given web archive performs digital preservation of the web material as described under FAIR Access. For extracts, it will likewise depend on the preservation of, for example, a research data management system.

Persistent referencing of preserved web material is also essential for reuse. The WARST study also examined citation practices and found that researchers have challenges citing archived web due to a lack of guidelines, standards, and best practices, and also problems with finding the dates or creators for the websites in a web archive. They further found challenges with citing archived web content from web archives with restrictive access due to legal deposit legislation (Healy et al., 2022).

We also saw that there were challenges in referencing the web material in web archives with archive URLs, but also an alternative to using the PWID (Zierau, 2023). There are

other referencing techniques like DOI, Handle, and ARK which are used in many research data management systems, and therefore can be good choices for extracts, where there are also examples like the IIPC Collaborative Collections (n.d.)²¹ and Archives Unleashed datasets.²² However, they are not suitable for references to web pages in web archives, as they “require the active updating of URL redirects” (ARK, n.d.), while the PWID is based on the archived metadata.

Discussion

In this chapter, we have examined the increasing use of Research Data Management (RDM) focusing on the specific RDM challenges in relation to research using web and social media archives. In this section, we will provide an overview of the results of these investigations, and examine where progress has been made and where the remaining challenges lie. In the final section, we will give recommendations for ensuring good practices for research data management, both for web archives who wish to stimulate research use of their collections and for researchers who study web and social media resources. Furthermore, we will provide suggestions of where further work and research is needed.

As described in the chapter, there is a gap between general RDM best practices and what is required when studying archived web and social media materials. An overview of key challenges and next steps is provided below, where the FAIR Data Principles have been used as a framework for structuring these reflections.

Findable: When it comes to the findability of archived web materials, standards such as the WARC format and the OCLC recommendations for the Descriptive Metadata for Web Archiving (Dooley and Bowers, 2018) have provided a good basis for ensuring *sufficient metadata* for web materials. However, it was noted that more documentation and enrichment of metadata are essential to facilitate access and research using web-archived material because of the nature and volume of such data. A more challenging area is the indexing of such information in *searchable platforms*, especially as a large proportion of web archives are not openly accessible. There are some examples of large-scale discovery platforms for archives web materials; collection-level descriptions of thematic web archive collections, as well as examples where archived web materials have been described integrated alongside other types of resources, for example, in library catalogues. Increasingly, there are examples of datasets that have been derived from larger-scale collections and deposited in open research data repositories such as Zenodo. While this enables transnational use and makes such datasets citable, for example in journal articles reusing such datasets, there is a risk that such datasets become scattered and not easily findable by researchers.

Accessible: Due to the legal and sometimes ethical constraints on web archives, accessibility is one of the key challenges when it comes to the study of the archived web. Access to web archives remains, in many cases, onsite access in the reading rooms of the archival institutions. For researchers who use archived web and social media regularly, this presents a practical access challenge, which was indeed exacerbated during the COVID-19 pandemic. However, this lack of open access to web and social media archives is a barrier to their more general uptake and use by the wider international research community. The creation and publication of derived datasets, such as seed lists and CDX files, does help in increasing the accessibility of web archives. In some countries, a proactive approach to the legal challenges has been taken. For example, the Royal Danish Library in Denmark has developed a legal

agreement for providing research access to archived web materials (Laursen, 2020), which could be inspirational for other web archives.

Interoperable: Due to the technical nature of web and social media archives, much thought has gone into the creation of standards to facilitate interoperability between web archives. For example, the WARC Standard is widely used in web archives internationally, even if the situation is more complex when it comes to social media archives. Additionally, there are a range of other protocols in the web archiving sphere that have been developed to foster interoperability, such as CDX files which are used to publish the list of URLs that have been indexed in a particular web archive. Furthermore, the Memento protocol has been successfully taken up in a range of web archives, and initiatives like the one described in the chapter “Comparing the holdings of closed national web archives through summaries” further showcase its use. The IIPC has been instrumental in fostering interoperability in the web archiving community by offering a platform for technical experts to communicate on specific issues and challenges they face, which has led to the shared development of web archiving tools and related standards. The use of internationally recognised standards and a common understanding of how to share archived web resources is an essential first step towards enabling transnational research across web archives. While the creation of transnational archived web collections is still hampered by nationally based legislation, their potential to be compared transnationally increases significantly if at least national datasets can be made available based on agreed standards.

Reusable: As the ultimate goal of the FAIR Data Principles is to encourage reuse, this is even more important when it comes to research use of web and social media archives. In addition to the richness of the metadata, the findability of archives, web materials, and the legal and ethical constraints highlighted above, there are additional challenges when it comes to reuse. Due to the inherent ephemeral and dynamic nature of the web and social media, the archiving of web materials is essential for its study. While web archives have made significant advances in this area, two particular challenges remain. Firstly, the persistent identification of archived web material is essential for referencing and citation of this data for research. The development of the PWID (Zierau, 2022) has made a major contribution here. However, the challenge is now to ensure the wide-scale uptake of the PWID across a large number of web archives. Secondly, long-term preservation is essential for archived web materials if they are going to be both persistently cited and used for future research. Due to the complexity of digital preservation in general, and in particular, for future interoperability of complex unstructured materials like web data, not all web archives have undertaken all the necessary steps in digital preservation to ensure future access.

In addition to the issues discussed about FAIR, it is important to note that a fairly advanced level of digital and data literacy is required for researchers who wish to study the archived web. This data literacy is required at all stages of the Research Data Management Lifecycle from the planning and design phase, through the collection, capture, and analysis phases, to the management, storage, and preservation of the resulting research data. Here the IIPC’s Training Materials²³ can offer a good introduction to using web archives for research, alongside the more advanced use of web archives for research including electronic lab notebooks, such as those provided through the GLAM Workbench.²⁴ Due to the specific nature of archived web and social media resources, the importance of domain-specific training materials is essential and an area which requires more attention in the future.

Conclusions

Research data management practices are important for researchers who use digital sources, but even more important for researchers who use web and social media archives for their research, both due to their digital complexity, as well as the legal and ethical issues. While information about good practices in research data management is becoming more prevalent in general, less specific information for web and social media archives is available. The contribution of this chapter has been to increase the level of understanding about transnational research practices for the study of web and social media archives, as well as the particular challenges related to the research data management of such resources. Finally, the discussion section identified which aspects of the FAIR Data Principles were most challenging for web and social media archives, and provided recommendations on the focus for future efforts. As a result of these investigations, it would seem that the responsibilities that lie with the libraries and other cultural heritage institutions that undertake web and social media archiving activities and which role the researchers should play in this process remain a little blurred. This calls for continued collaboration between researchers and cultural heritage professionals when it comes to the study of archives web and social media resources, and the development of a shared responsibility for ensuring good research data management practices.

The original intention of this chapter was to address research data management for the transnational study of web and social media archives. What we have seen is that many ongoing national and transnational research projects share common needs and challenges. Nationally based archiving activities and legislative practices are the most important of those challenges. At the same time, there is a growing awareness of addressing these needs by developing supportive systems with regard to tools, workflows, collaborative networks, and support institutions. Some are on local or national level, but others are truly internationally orientated. It is through international research networks such as WARCnet that the interoperability between national initiatives can be fostered.

Notes

- 1 DMPOnline (n.d.). Plan to make data work for you. Web archive. <https://web.archive.org/web/20230610114804/https://dmponline.dcc.ac.uk/>.
- 2 Nationaal register webarchieven (n.d.). Register. Web archive. <https://web.archive.org/web/20231005030219/https://registerwebarchieven.nl/register>.
- 3 Internet Memory Foundation (2018, January 19). NLI WEB ARCHIVE: A record of the online life of Ireland. Web archive. <https://web.archive.org/web/20180119123712/http://collection.europarchive.org/nli/>.
- 4 Archive-It (2019, March 1). National Library of Ireland. Web archive. <https://web.archive.org/web/20190301215300/https://archive-it.org/home/nli>.
- 5 Github (n.d.). netarchivesuite / solrwayback. Web archive. <https://web.archive.org/web/20240305025759/https://github.com/netarchivesuite/solrwayback>.
- 6 Github (n.d.). oldweb-today. Web archive. <https://web.archive.org/web/20231029094007/https://github.com/oldweb-today/oldweb-today>.
- 7 Zenodo (n.d.). Zenodo. Web archive. <https://web.archive.org/web/20240229232320/https://zenodo.org>.
- 8 The Archives Unleashed Project (n.d.). Datasets. Web archive. <https://web.archive.org/web/20231220215859/https://archivesunleashed.org/publications/#datasets>.
- 9 The Archives Unleashed Project (n.d.). Archives Unleashed Notebooks. Web archive. <https://web.archive.org/web/20231220215859/https://archivesunleashed.org/notebooks/>.

- 10 GLAM Workbench (n.d.). Web Archives. Web archive. <https://web.archive.org/web/20240303045125/https://glam-workbench.net/web-archives/>.
- 11 Library of Congress (n.d.). Web Archive Datasets. Web archive. <https://web.archive.org/web/20230929035544/https://labs.loc.gov/work/experiments/webarchive-datasets>.
- 12 KB Lab (n.d.). Web Collection Chinese Netherlands. Web archive. <https://web.archive.org/web/20240301104029/https://lab.kb.nl/dataset/web-collection-chinese-netherlands>.
- 13 Luxembourg Web Archive (n.d.). 2019 European elections. Web archive. <https://web.archive.org/web/20231219030056/www.webarchive.lu/what-we-have/2019-european-elections/>.
- 14 International GLAM Labs Community (n.d.). International GLAM Labs Community. Web archive. <https://web.archive.org/web/20240304174441/https://glamlabs.io/>.
- 15 Webarchive Austria (n.d.). Webarchive Austria. Web archive. <https://web.archive.org/web/20240301105245/https://labs.onb.ac.at/en/datasets/webarchive/>.
- 16 Webarchive Austria (n.d.). ONB Labs. Web archive. <https://web.archive.org/web/20231003101005/https://labs.onb.ac.at/en/>.
- 17 Webarchive Austria (n.d.). Selective json. Web archive. <https://web.archive.org/web/20240301105415/https://webarchiv.onb.ac.at/data/selective.json>.
- 18 The Archives Unleashed Project (n.d.). Archives Research Compute Hub (ARCH). Web archive. <https://web.archive.org/web/20231220215900/https://archivesunleashed.org/arch/>.
- 19 Internet Archive (n.d.). Welcome to Archive-It!. Web archive. <https://web.archive.org/web/20240229124717/https://archive-it.org/>.
- 20 Internet Archive (n.d.). Wayback machine. Web archive. <https://web.archive.org/web/20240229013956/https://archive.org/>.
- 21 International Internet Preservation Consortium (n.d.). Collaborative collections. Web archive. <https://web.archive.org/web/20231224113451/https://netpreserve.org/projects/collaborative-collections/>.
- 22 The Archives Unleashed Project (n.d.). Datasets. Web archive. <https://web.archive.org/web/20231220215859/https://archivesunleashed.org/publications/#datasets>.
- 23 International Internet Preservation Consortium (n.d.). Training materials. Web archive. <https://web.archive.org/web/20231123175400/https://netpreserve.org/web-archiving/training-materials/>.
- 24 GLAM Workbench (n.d.). Web Archives. Web archive. <https://web.archive.org/web/20240303045125/https://glam-workbench.net/web-archives/>.

References

- Note: Archived web references are provided as PWID URNs registered at IANA, archived version at (www.iana.org/assignments/urn-formal/pwid, <https://web.archive.org/web/20231012093849/www.iana.org/assignments/urn-formal/pwid>, urn:pwid:archive.org:2023-10-12T09:38:49Z:page:www.iana.org/assignments/urn-formal/pwid) which includes description of resolving a PWID.
- 2019 *European Elections*. Luxembourg Web Archive. urn:pwid:archive.org:2023-12-19T03:00:56Z:page:www.webarchive.lu/what-we-have/2019-european-elections/.
- Apache Software Foundation*. (2023). Solr. urn:pwid:archive.org:2024-02-29T12:28:02Z:page:https://solr.apache.org/.
- Archives Unleashed Notebooks*. (n.d.). The Archives Unleashed Project. urn:pwid:archive.org:2023-08-24T21:38:45Z:page:https://archivesunleashed.org/publications/#datasets.
- Home of the Archival Resource Key (ARK)*. (n.d.)
- Aturban, M. (2019a, September 10). Where did the archive go? Part 2: National Library of Ireland [Blog post]. Web Science and Digital Libraries Research Group. urn:pwid:archive.org:2023-12-25T00:15:58Z:page:https://ws-dl.blogspot.com/2019/09/2019-09-10-where-did-archive-go-part-2.html.
- Aturban, M. (2019b, September 25). Where did the archive go? Part 3: Public Record Office of Northern Ireland [Blog post]. Web Science and Digital Libraries Research Group. urn:pwid:archive.org:2024-02-26T22:43:53Z:page:https://ws-dl.blogspot.com/2019/09/2019-09-25-where-did-archive-go-part-3.html.

Towards transnational research data management practices

- Aturban, M., Nelson, M. L., & Weigle, M. C. (2021). *Where Did the Web Archive Go?* (arXiv:2108.05939). arXiv. urn:pwid:archive.org:2024-02-28T05:58:40Z:page:http://arxiv.org/abs/2108.05939.
- Aubry, S., Cartier, E., & Stirling, P. (2018, November 15). Néonaute: Mining web archives for linguistic analysis. In *IIPC Web Archiving Conference*, Wellington. urn:pwid:archive.org:2023-12-02T13:52:06Z:page:https://digital.library.unt.edu/ark:/67531.
- Brügger, N. (2018, December 14). *How to Dig into the History of a Nation's Web? The Development of the Danish Web 2006–2015*. IIPC RSS Webinar, Online.
- Brügger, N. (2021a). Digital humanities and web archives: Possible new paths for combining datasets. *International Journal of Digital Humanities*, 2(1–3), 145–168. https://doi.org/10.1007/s42803-021-00038-z.
- Brügger, N. (2021b). The need for research infrastructures for the study of web archives. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web* (pp. 217–224). Springer International Publishing. https://doi.org/10.1007/978-3-030-63291-5_17.
- Brügger, N., & Laursen, D. (Eds.). (2019). *The Historical Web and Digital Humanities: The Case of National Web Domains*. Routledge.
- Brügger, N., & Schroeder, R. (Eds.). (2017). *The Web as History: Using Web Archives to Understand the Past and the Present*. UCL Press.
- Candela, G., Gabriëls, N., Chambers, S., Pham, T.-A., Ames, S., Fitzgerald, N., Hofmann, K., Harbo, V., Potter, A., Ferriter, M., Manchester, E., Irollo, A., Van Keer, E., Mahey, M., Holownia, O., & Dobrev, M. (2023). *A Checklist to Publish Collections as Data in GLAM Institutions*. https://doi.org/10.48550/ARXIV.2304.02603.
- Carroll, S. R., Herczog, E., Hudson, M. et al. (2021). Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Sci Data*, 8, 108. https://doi.org/10.1038/s41597-021-00892-0.
- Chambers, S. (2020, May 6). *Web Archives for Open Science: How FAIR Can We Go?* WARCnet Kickoff Meeting. urn:pwid:archive.org:2024-02-29T16:52:04:page:https://cc.au.dk/en/warcnet/presentations/kickoff-meeting-2020.
- Cochrane, E., Rechert, K., Oberhauser, J., Anderson, S., Fox, C., & Gates, E. (2022). Useable software forever: The Emulator as a Service Infrastructure (EaaS) program of work. In *Proceedings IPres 2022 Glasgow 12–16 September 2022. iPres 2022, the 18th International Conference on Digital Preservation*, Glasgow. https://doi.org/10.17605/OSF.IO/8BCZF.
- Costea, M.-D. (2018). *Report on the Scholarly Use of Web Archives* [Report]. NetLab. urn:pwid:archive.org:2023-10-28T14:39:32Z:page:http://netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf.
- Digital Curation Centre (DCC). (n.d.). *DMPOnline* [Page]. urn:pwid:archive.org:2023-12-26T21:10:51Z:page:https://dmponline.dcc.ac.uk/.
- Dooley, J., & Bowers, K. (2018). *Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. OCLC Research Dublin, Ohio. https://doi.org/10.25333/C3005C.
- Eldakar, Y., & Holownia, O. (2022, June 29). *Republishing IIPC Collections Through Alternative Interfaces*. IIPC Web Archiving Conference, Library of Congress, May 2022. urn:pwid:archive.org:2023-01-01T05:23:11Z:page:www.youtube.com/watch?v=15AGoOJBM6E.
- European Commission. (n.d.). *European Open Science Cloud (EOSC)*. Research and Innovation. urn:pwid:archive.org:2023-09-08T04:09:52Z:https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en.
- European Union. (n.d.). *EU's Open Science Policy*. Research and Innovation. urn:pwid:archive.org:2023-08-29T19:20:31Z:page:https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science_en#the-eus-open-science-policy.
- Geeraert, F., De Wild, K., & Aasman, S. (2024). What can we learn from URLs? Understanding the scope of COVID-19 web archive collections for transnational analyses. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge Companion to Transnational Web Archive Studies*. Routledge.
- Ghent University. (n.d.). *Research Data Management (RDM)*. urn:pwid:archive.org:2023-12-07T16:07:09Z:page:www.ugent.be/en/research/datamanagement.
- Guides for Researchers: How to Make Your Data FAIR*. (n.d.). OpenAIRE. urn:pwid:archive.org:2023-10-25T13:41:12Z:page:www.openaire.eu/how-to-make-your-data-fair.

- Have, U. (2020, May 6). *Hand Me the Data! What You Should Know as a Humanities Researcher Before Asking for Data from a Web Archive*. WARCnet Kickoff Meeting. urn:pwid:archive.org:2023-09-01T08:24:58Z:page:https://cc.au.dk/en/warcnet/presentations/kickoff-meeting-2020.
- Healy, S., Byrne, H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., & Jansma, R. (2022). *Skills, Tools, and Knowledge Ecologies in Web Archive Research*. WARCnet Special Report. https://doi.org/10.23636/9SEX-KV63.
- Home of the Archival Resource Key (ARK)*. (n.d.). ARK Alliance. urn:pwid:archive.org:2023-08-22T15:08:01Z:page:https://arks.org/.
- How to Make Your Data FAIR*. (n.d.). OpenAIRE. urn:pwid:archive.org:2023-10-25T13:41:12Z:page:www.openaire.eu/how-to-make-your-data-fair.
- IIPC Collaborative Collections*. (n.d.). Netpreserve. urn:pwid:archive.org:2023-09-22T23:36:58Z:page:https://netpreserve.org/projects/collaborative-collections/.
- Internet Archive. (n.d.). CDX file format. urn:pwid:archive.org:2024-01-14T23:14:28Z:page:https://archive.org/web/researcher/cdx_file_format.php.
- ISO 28500:2017. (2017). *Information and Documentation — WARC File Format*. [Standard]. International Organization for Standardization.
- ISO 690:2021. (2021). *Information and Documentation — Guidelines for Bibliographic References and Citations to Information Resources*. [Standard]. International Organization for Standardization.
- Jackson, A. (2015, November 20). The provenance of web archives [Blog post]. UK Web Archive Blog. urn:pwid:archive.org:2022-07-05T05:47:00Z:page:https://britishlibrary.typepad.co.uk/webarchive/2015/11/.
- Jones, S. M., Klein, M., Sompel, H. v.d., Nelson, M. L., & Weigle, M. C. (2021). Interoperability for accessing versions of web resources with the memento protocol. In D. Gomes, E. Demidova, J. Winters, & T. Risse (Eds.), *The Past Web* (pp. 101–126). Springer International Publishing. https://doi.org/10.1007/978-3-030-63291-5_9.
- Laursen, D. (2020, May 11). *Developing a Legal Agreement for Research-Access to a Web Archive*. WARCNet Meeting, Online. https://youtu.be/kvqKSgVPuXE?si=21HCbAPoq963UIiu.
- Lavoie, B., & Dempsey, L. (2004). Thirteen Ways of Looking at ... Digital Preservation. *D-Lib Magazine*, 10(7/8). https://doi.org/10.1045/july2004-lavoie.
- Lieber, R., & Liebert, J. (2012). Something rotten in the state of legal citation: The life span of a United States Supreme Court citation containing an internet link (1996–2010). *Yale Journal of Law & Technology*, 273(15). urn:pwid:archive.org:2023-09-24T11:31:35Z:page:https://yjolt.org/something-rotten-state-legal-citation-life-span-united-states-supreme-court-citation-containing.
- Lin, K., Teszelszky, K., & De Bode, P. (2019). *Collection Description Chinese Netherlands Web Collection*. Zenodo. https://doi.org/10.5281/ZENODO.5526728.
- Maemura, E. (2018). What's cached is prologue: Reviewing recent web archives research towards supporting scholarly use. In Proceedings of the Association for Information Science and Technology (vol. 55(1), pp. 327–336). https://doi.org/10.1002/pra2.2018.14505501036.
- Maurer, Y. (2022, August 10). *Investigate Holdings of Web Archives through Summaries: Cdx-Summarize*. urn:pwid:archive.org:2024-07-26T13:18:08Z:page:https://netpreserveblog.wordpress.com/2022/08/10/investigate-holdings-of-web-archives-through-summaries-cdx-summarize/.
- Meyer, M. K. K., & Zierau, E. (2019). Spitting Image — Press photographs and memes as digital cultural heritage in NetarkivetSpitting Image. *Fund Og Forskning, Det Kgl. Bibliotek, Copenhagen*, 58. urn:pwid:archive.org:2021-03-10T23:09:44Z:page:https://tidsskrift.dk/fundogforskning/article/view/125306.
- Milligan, I. (2019). *History in the Age of Abundance? How the Web Is Transforming Historical Research*. McGill-Queen's University Press.
- Nielsen, J., Maurer, Y., & Zierau, E. (2024). Comparing the holdings of closed national web archives through summaries. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge Companion to Transnational Web Archive Studies*. Routledge.
- Nyvang, C., Kromann, T. H., & Zierau, E. (2017). Capturing the web at large: A critique of current web referencing practices. *Researchers, Practitioners and Their Use of the Archived Web*, 1–9. https://doi.org/10.14296/resaw.0004.
- Ogden, J., Kurzmeier, M., Clavert, F., & Currie, M. (2022). How to design web archives research. *SAGE Research Methods: Doing Research Online*. https://doi.org/10.4135/9781529610147.

- OpenAIRE. (n.d.). Guides for researchers - How to make your data FAIR. urn:pwid:archive.org:2024-04-04T06:27:34Z:page:www.openaire.eu/how-to-make-your-data-fair.
- Oury, C. (2009, January 27). WARC Implementation Guidelines 1.0. urn:pwid:archive.org:2024-05-13T20:13:59Z:page:https://iipc.github.io/warc-specifications/guidelines/warc-implementation-guidelines/.
- Research Data Management Toolkit*. (2018, March 30). JISC. urn:pwid:archive.org:2023-11-09T15:12:55Z:page:https://beta.jisc.ac.uk/guides/research-data-management-toolkit.
- Ruest, N., Fritz, S., & Milligan, I. (2022). Creating order from the mess: Web archive derivative datasets and notebooks. *Archives and Records*, 43(3), 316–331. https://doi.org/10.1080/23257962.2022.2100336.
- Schafer, V., Aasman, S., Clavert, F., de Wild, K., & Sirajzade, J. (2024). The challenges of searching for women in the COVID-19 web archive collections: Promises, achievements, and pitfalls. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge Companion to Transnational Web Archive Studies*. Routledge.
- Sherratt, T., & Jackson, A. (2021). *GLAM Workbench — Web Archives* [Wiki]. GLAM Workbench/GitHub.Io. urn:pwid:archive.org:2021-01-20T18:33:00Z:page:https://glam-workbench.github.io/web-archives/.
- Tóth-Czifra, E. (2019). DARIAH pathfinder to data management best practices in the humanities. *DARIAH-Campus*. urn:pwid:archive.org:2022-11-02T15:49:26Z:https://campus.dariah.eu/en/resource/posts/dariah-pathfinder-to-data-management-best-practices-in-the-humanities.
- Truman, G. (2016). *Web Archiving Environmental Scan* (Library Report Harvard Library Report, 2016; Harvard Library Report). Harvard Library. urn:pwid:archive.org:2023-07-10T00:10:20Z:page:https://dash.harvard.edu/handle/1/25658314.
- Truter, V. (2021, September 1). Research data management and sharing practices of researchers in web archive studies [Conference presentation & abstract]. In *Conference: Engaging with Web Archives 4 Digital Humanities (#EWA4DH)*, Maynooth University Arts and Humanities Institute, Co., Kildare, Ireland. urn:pwid:archive.org:2023-09-03T23:52:52Z:page:https://ewaconference.com/ewa4dh-2021/ewa4dh-programme/.
- Tsakalidis, A., Basile, P., Bazzi, M., Cucuringu, M., & McGillivray, B. (2021). DUKweb, diachronic word representations from the UK Web Archive corpus. *Scientific Data*, 8(1), 269. https://doi.org/10.1038/s41597-021-01047-x.
- UK Web Archive, & Bodleian Library. (2014). *Ebola Outbreak, West Africa 2014*. UK Web Archive.
- Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, 1(1), 85–111. https://doi.org/10.1007/s42803-019-00007-7.
- Web Collection Chinese Netherlands*. (2019). KB Lab. urn:pwid:archive.org:2024-03-01T10:40:29Z:page:https://lab.kb.nl/dataset/web-collection-chinese-netherlands.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18.
- Winters, J. (2020). Giving with one click, taking with the other: Electronic legal deposit, web archives and researcher access. In P. Gooding & M. Terras (Eds.), *Electronic Legal Deposit: Shaping the Library Collections of the Future* (pp. 159–178). Facet Publishing.
- Zierau, E. (2022). *URN Namespace Registration for Persistent Web IDentifiers (PWID)*. Internet Assigned Numbers Authority (IANA). urn:pwid:archive.org:2023-10-12T09:38:49Z:page:www.iana.org/assignments/urn-formal/pwid.
- Zierau, E., & Møldrup-Dalum, P. (2021, June 15). Making web collections for research sustainable & reusable — possibilities and challenges experienced. In *2021 International Internet Preservation Coalition (IIPC) General Assembly and Web Archiving Conference, June 14–16, 2021*. urn:pwid:archive.org:2023-09-18T16:40:48Z:page:https://digital.library.unt.edu/ark:/67531/metadcl827590/.

26

THE IMPORTANCE OF LEGAL REQUIREMENTS FOR WEB ARCHIVES STUDIES IN BELGIAN AND FRENCH LAWS

Lise-Anne Denis

Introduction

The objective of this chapter is to expose and analyse in a synthetic and accessible way the various obligations imposed on archiving institutions and researchers when it comes to the personal data contained in the web pages they intend to archive or use in their research, as well as the impact of copyright protection on some content aimed for the archives. It focuses on European law, and on the particularities of Belgian and French laws. This will also show how researchers, when considering to study archives collected in other countries, have to take into account the particular rules put in place by said countries.

Among the legal requirements that researchers and archiving institutions tackling web archiving activities must keep in mind, we can first mention personal data protection, which is regulated at the European level by the General Data Protection Regulation (or “GDPR”). This regulation directly imposes several obligations to all researchers processing personal data. A derogatory regime is provided for scientific or historical research purposes and lightens the researchers’ obligations. As some freedom is left to Member States regarding this regime, it has been implemented slightly differently in Belgian and French laws.

A second important legal framework for web archives studies is copyright law, as a large portion of the content found on the internet, and therefore in the web archives, is protected by copyrights. In principle, a researcher has to obtain the authorization of the copyright holder before reproducing or communicating any protected content. Copyright is also governed by several European Directives. As some freedom is left to Member States for the implementation of these Directives, Belgian law and French law vary on this matter as well.

The impact of personal data protection on web archiving

Personal data protection has taken a turn with the adoption in 2016 of the European General Data Protection Regulation (2016/679), or GDPR. Indeed, this particular legal instrument

The importance of legal requirements for web archives studies

has had such an impact that it has become part of conversations among all types of people, even those unfamiliar with any legal matters and of no juridical background.

Although the principles established in this Regulation represent nothing new, the seriousness of the sanctions at stake was such that it compelled most companies to carefully stick with it. Hence, the incredible amount of emails every European citizen received in the spring of 2016, made sure they were aware of the processing of personal data accomplished by every single website they had once given their email address to.

The objective of personal data protection is intrinsically linked to privacy, which has taken a growing part of our preoccupations alongside the evolution of the digital world. Indeed, “big data” has become an unavoidable phenomenon, with many of those data being considered “personal”, which means they can lead to the identification of a person. It does not go unnoticed that we are brought to give our personal information in a gigantic amount of online situations. On top of the “public” and active cession of our data, we must add the “secretive” and passive collection of our personal information that results from our navigation on the web. The “Cookies” that we tend to accept without thinking, nor reading the information presented to us on the characteristic pop-up windows, is an example of the latter. The collection of such important amounts of our private information can have serious and negative consequences, such as discrimination if they are used for malicious purposes (Tambou, 2020, p. 5). It was therefore necessary to regulate the use of personal data, especially in the digital world, which was part of the GDPR’s goal.

In this part of the chapter, we will first study the European framework of personal data protection, and continue with the particularities of its implementation in Belgian and French laws. We will end this section with an explanation of the impact of personal data protection on web archiving and research among web archives, starting with the preservation stage and ending with access to web archives as well as their reuse.

Scope of the European framework

As mentioned just before, the GDPR has had the effect of a tsunami on the activities of almost all companies dealing with a clientele, as it automatically led them to keep personal data on their customers. This Regulation has introduced many obligations to companies and private individuals as soon as they process data that can identify people. We will focus here on explaining the wide scope of the GDPR as well as on the obligations and principles that are the most relevant regarding web archiving.

This Regulation has the particularity to apply to a very large array of situations and its scope is wide in three ways. Firstly, materially. Indeed, it is specified that the obligations contained in the GDPR must be respected as soon as personal data are being processed. Both personal data and their processing are defined very widely: a “personal data” is any information relating to an identified or identifiable natural person (article 4, (1)), while “processing” can take the form of any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means (article 4, (2)). To have a better idea of the meaning of this vague definition, examples of processing are given: it is the collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction. Therefore, as soon as personal data are collected for a purpose by a company, it will generally have to comply with the GDPR.

Secondly, the scope is wide regarding the categories of people or entities who have to respect the regulation. The obligations contained in the GDPR have to be respected when personal data are being processed by a controller, which is defined as the natural or legal person, public authority, agency, or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data (article 4, (7)). Exceptions are however provided for natural people in the course of a purely personal or household activity as well as for competent authorities for criminal and judicial purposes, among other categories of activities which are also excluded and that we will not go to the length of explaining here.

Third, geographically. One of the most striking novelties introduced by the GDPR is its large territorial scope (Tambou, 2020, p. 95). As it is applicable to all processing of personal data belonging to European citizens (article 3), it must be respected even by all non-European companies collecting and using those data. This entailed that the big American digital platforms also had to comply with it, such as Google or Facebook, considering they collected massive amounts of personal data belonging to people living in Europe. However, the application of the GDPR to controllers operating outside of the EU is limited to processing activities related to the offering of goods or services (irrespective of whether a payment of the data subject is required) to data subjects in the Union, or to the monitoring of their behaviour as far as their behaviour takes place within the Union (Tambou, 2020, p. 98). It therefore will not necessarily entail that archiving institutions or researchers even outside Europe will have to respect the GDPR if they intend to archive or reuse web pages containing the personal data of European citizens. One must still be careful when envisaging such activities as it is an easy bridge to cross considering the accessibility of the World Wide Web.

Principle of purpose limitation, data minimization and storage limitation

As the aim of this chapter is not to present the content of the GDPR extensively and exhaustively, we can limit our analysis to three principles which, according to us, represent the core of the GDPR and are the most relevant principles regarding web archiving. They can be studied together as they are all somewhat connected to each other and derive from the purpose that must be at the centre of every personal data processing.

Firstly, every processing of personal data must be justified by a legitimate purpose (article 5, 1, (b)). Indeed, all personal data must be collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes.

Secondly, all personal data processed must be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed (article 5, 1, (c)). This is referred to as the data minimization principle.

Third, once the data have been collected, their storage must be, in principle, limited to the amount of time necessary for the purposes for which the personal data are processed (article 5, 1, (e)).

Finally, it is also important to pinpoint the fact that “data protection by design and by default” must be ensured when the processing of personal data is envisaged (article 25). On the one hand, it means that the risks involved, as well as the scope, nature, context, and purposes of the processing, must be considered at the time of the determination of the means for processing and at the time of the processing itself. On the other hand, it means that appropriate technical and organizational measures must be implemented to ensure that, by default, only personal data that are necessary for each specific purpose of the processing are processed. This particular rule entails that personal data protection cannot be quickly

The importance of legal requirements for web archives studies

solved later down the line of the establishment of a web archiving process or at the last stage of research: it must be taken care of right at the beginning of any discussion involving the archiving or reuse of personal data.

Obligations imposed by the GDPR

Among the numerous requirements entailed by the application of the GDPR, the most constraining is perhaps the obligation to base any processing of personal data on one of the lawfulness bases exhaustively listed in the Regulation, the obligation to inform the data subject (i.e., the person whose personal data are being processed) of the processing, and the obligation to allow the data subjects to exercise their rights.

As mentioned, any processing of personal data must be justified by one of the legal bases listed in the Regulation to remain lawful (article 6). These legal bases are the following: the consent given by the data subject for the processing of his or her personal data, the necessity of the processing for the performance of a contract to which the data subject is party, the necessity of the processing to comply with a legal obligation imposed to the controller, the necessity of the processing to protect the vital interests of the data subject, the necessity of the processing for a task carried out in the public interest or the necessity of the processing for the purposes of the legitimate interests pursued by the controller or a third party. This entails that, on top of a specific purpose that is required for any personal data processing, the controller must also find itself in at least one of these situations, making it quite hard to act spontaneously when it comes to any activity involving personal data.

The obligation of information imposed on the controller implies that a list of information will have to be communicated to the data subject when his or her personal data are processed (articles 12 to 14). Among those pieces of information, which vary slightly if the personal data are collected from the data subject or not, we find the identity and contact details of the controller, the purposes of the processing, the recipients of the personal data, or the existence of automated decision-making.

On top of that, a requirement to allow the data subjects to exercise their rights must also be respected. Several rights have been awarded to people regarding the use made of their personal data to protect their rights to privacy: they have the right to transparency regarding the processing of their personal data (which constitutes the other side of the obligation of information coin), the right to access their personal data (article 15), the right to rectification (article 16), the right to erasure (otherwise known as the “right to be forgotten”) (article 17), the right to restriction of the processing (article 18), the right to data portability (article 20), and the right to object automated decision-making (roughly meaning decisions made “by” an algorithm) (article 21). Any controller of a processing involving personal data will therefore have to make sure people whose personal data have been processed are able to exercise these rights. Information on these rights must also be provided to the data subjects.

Implementation in Belgian law

As mentioned, some elements related to the processing of personal data that entered the scope of the GDPR were not decided by the European legislator itself, but rather left at the will of European Member States. It therefore implies that some consequences of personal data processing vary from state to state.

Among the elements that were left to be regulated by national law, the implementation of the derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes (article 89) is perhaps the most interesting regarding web archiving and research. We will therefore limit our analysis to this particular aspect, for both Belgian and French laws.

Member States received some freedom for the implementation of several elements of article 89. First of all, they could choose which of the rights of data subjects the derogations could apply to: they could choose to authorize derogations from the right of access, the right to rectification, the right to restriction of processing, the right to receive notification of rectification or erasure of personal data, the right to data portability and the right to object to the processing. Those derogations can differ depending on the type of motive invoked: either the processing for archiving purposes in the public interest or the processing for scientific, historical research, or statistical purposes. Second, the Member States had to decide on the appropriate safeguards that would compensate for those derogations.

Regarding Belgian law, the law of the 30th of July 2018 implementing those aspects of the GDPR has awarded all the derogations aforementioned to controllers who process personal data for archiving purposes in the public interest as well as for researchers. As to the safeguards that had to be put in place if the derogations were implemented, the Belgian law provides that either a code of conduct approved by the Belgian data protection authority has to be complied with by the controller, or an array of specific rules have to be respected. Regarding those rules, they consist of the designation of a data protection officer in some cases, the inclusion of several mentions to the records of processing activities (such as the reasons why the exercise of rights is likely to render impossible the archiving activities or the justification of the public interest of the archives), or the provision of additional information to the data subject in case of direct data collection.

The Belgian legislator has also made the choice to provide specific rules related to the diffusion and communication of archives that contain personal data.

We will come back to those specificities in the section about the consequences of personal data protection rules on web archiving.

Implementation in French law

As to the same elements we just explained regarding their regulation in Belgian law, the French law implementing some aspects of the GDPR seems to differ. Focusing on what is of interest to us in this chapter, when it comes to the derogations to the Regulation permitted in the case of processing of personal data for archiving purposes in the public interest, the law of the 20th of June 2018 provides derogations to all the rights for which the GDPR left freedom to Member States. Therefore, there are also derogations to the same rights as in Belgian law when personal data are collected for archiving or research purposes in France. However, as to the archiving purposes side, those derogations are strictly limited to the processing of personal data made by public archives services, which excludes any private archiving initiative from their scope. This is not the case in Belgian law: any institution archiving personal data in the public interest can benefit from the derogations.

Regarding the safeguards aimed at balancing the derogations to several rights of the data subject, the French legislator has provided several obligations if the processing of personal data is likely to give rise to a high risk for the rights and freedoms of people. On top of that, the part of the French Heritage Code on the public archives provides additional safeguards

The importance of legal requirements for web archives studies

when their archiving activities involve the processing of personal data: it provides that the exception to the storage limitation principle (which we will go into more detail later on) entails a selection, once the duration of the conservation initially planned is expired, of the personal data which should be preserved or eliminated from the archives. This is an application of the data minimization principle.

This is, therefore, a good example of a variation of the Regulation of personal data processing between states that are both members of the European Union, despite them being neighbours and having legal systems that were historically common. It helps to show the trickiness of archiving personal data and having to comply with the various legislations applicable to the matter.

Consequences for web archiving and web archives research

As we have seen, the wide scope of application of the GDPR as well as the massive presence of personal data on the internet, and therefore the web archives, implies that archiving institutions aiming at archiving the web and researchers aiming at using it will most likely have to respect the obligations imposed by the European Regulation, often even if they operate outside Europe. This will have quite logically a greater impact on social media archiving, as social media is obviously filled with personal data.

The application of the GDPR to web archiving has several consequences, both from a pure preservation point of view and from access to or reuse of the web archives. Fortunately, those two activities can benefit from the application of some derogations, allowing the archiving institutions and researchers to avoid some obligations imposed by the GDPR for the processing of personal data they operate.

Regarding the preservation stage of web archiving, it constitutes a collection and storage that match the definition of a processing, which, if it involves personal data, will have to comply with the GDPR (Vanreck, 2018, p. 840).

It appears that the compliance with the principles and obligations of the GDPR for every single preservation of web content would make it extremely difficult, or even impossible to achieve. Fortunately, some derogations to these principles and obligations exist, but one must make sure to be able to benefit from them and to comply with the other obligations that ensue in their application.

One of the main derogations provided by the GDPR is the exception for the processing of personal data for archiving purposes in the public interest awarded by article 89, which allows companies, including some archiving institutions, to avoid many complexities involved in the Regulation. A few conditions must be met to enter the scope of this derogation. First of all, the data processing achieved for archiving purposes must be made “in the public interest”. This could include private structures that pursue a public interest (Vanreck, 2018, p. 852). As mentioned previously, some conditions can be added by national legislation as safeguards aimed at balancing the risks caused by the derogations (see the additional condition imposed by France which limits their scope to public structures).

If those conditions are met, the web archivist will not have to comply with several obligations imposed by the Regulation. Some of these derogations are directly provided in the GDPR, while others must have been implemented in national law (as referred to in the previous sections).

Regarding the first category, archiving institutions processing personal data will not have to respect the purpose limitation principle, the storage limitation principle, the interdiction

to process special categories of personal data (such as health-related or judicial data), the right to information in case of indirect collection of personal data, nor the right to erasure. Those controllers will however have to put the necessary technical and organizational measures in place to ensure respect for the principle of data minimization and to allow, where appropriate, the data subjects to exercise their rights. The definition of effective procedures made to handle data subject requests or staff awareness and training are examples of such measures.

Regarding the second category, as they vary from Member State to Member State, we already mentioned the rights of data subjects that could be derogated from as well as the obligations imposed to controllers who benefit from those derogations in Belgian and French laws.

This means that any web archive institution intending to archive personal data will still have to respect several obligations, despite the fact they benefit from derogations to the “classic” application of the GDPR. They also have to keep in mind that those derogations and alternative obligations can differ from one Member State to the other inside the European Union.

As to the access stage of the web archives, it gets a bit more complicated. Indeed, giving access to an undetermined amount of people to archives containing personal data can be quite risky, as it could deprive the person whose personal data are archived from the possibility of erasing them from the internet, and therefore from the human tendency to forget information after a certain amount of time (especially considering the derogation to the right of erasure applicable in that case that we just mentioned). It is then quite logical that the rules are a bit more constraining.

This particular stage of the archiving is not regulated as such by the GDPR. Therefore, as it constitutes a different type of processing, it will trigger the application of all the rules and derogations we already discussed regarding the preservation of the archives (Michel, 2021, p. 185). However, some states, such as Belgium, have decided to adopt specific rules regarding the access given to personal data contained in archives. Those rules make a distinction between the act of dissemination (publication of personal data to unidentified third parties) and the act of communication (communication of personal data to identified third parties) (Michel, 2021, p. 198). To diffuse personal data contained in archives in an “unpseudonymized” format, one of four conditions must be met: the consent of the data subject, the fact that the personal data were made public by the data subject themselves, the fact that the data have a close link with the historical or public nature of the data subject or the fact that the data have a close link with the historical or public nature of the facts with which the data subject is involved. Otherwise, the data will have to be pseudonymized before giving access to them (Michel, 2019, p. 9). To communicate personal data contained in archives in a clear way, one of the same four conditions applicable to the dissemination must be met. Otherwise, the communication can be done with the data in the clear, but the third parties having access to them must not be able to reproduce them, except in a manuscript way.

As to French law, there are also some restrictions to the communication of the public archives, but they do not refer to personal data protection and the GDPR as such.

When considering the communication or the diffusion of their web archives that most likely contain personal data, archiving institutions will therefore have to be careful of the specific rules that could be imposed in the states they target. An easy solution is often to pseudonymize (or anonymize if possible, but it is almost impossible regarding web content) the personal data contained in the archives to avoid those potential restrictions.

Finally, when it comes to the research that can be done within those web archives, researchers who reuse personal data contained in the archives will also have to respect all

The importance of legal requirements for web archives studies

those obligations if they intend to publish the results of their research. They can however also benefit from the derogations for archiving purposes and scientific, historical, and statistical purposes that we previously studied. The same rules as the ones stated for the diffusion and communication of those data do apply here as well.

Now, one could wonder what would be at stake if all those tricky conditions and obligations were not respected when archiving or reusing web content. Well, the GDPR has also attracted that much attention by being particularly relentless regarding any contravention of its rules (Tambou, 2020, p. 10). It has provided quite severe financial sanctions, sufficiently important to worry the GAFAM giants. Those penalties go as follows: each European Member State data protection authority can impose administrative fines on the companies and people submitted to the scope of the GDPR. Depending on the nature, gravity, and duration of the infringement, the intentional or negligent character of the infringement, or the degree of responsibility of the controller, they can impose fines going up to 20 million euros or 4% of the company's total worldwide annual turnover (whichever is higher).

Considering the amount of money at stake, personal data protection shall not be taken lightly by any web archiving institution or research institution entering the GDPR scope.

Copyright law as an obstacle to web archiving

Far away from any privacy considerations, copyright protection is an element that must also be kept in mind by all web archivists.

Copyrights have been created to encourage creation in general by rewarding it financially and protecting it with some degree of exclusivity. The general principles of copyright law tend to be very similar across the world, although there obviously exists some divergence between states.

As is the case for personal data protection, copyright is also regulated at the European level, although the harmonization is not maximum. This means that European Member States have way more room for manoeuvre when it comes to the copyright rules imposed at a national level. This can therefore complicate quite a bit the compliance with copyright law when it comes to web archiving, as everything web-related can be difficult to tie up to one state in particular. This however does not prevent archiving institutions or researchers from respecting the copyrights protecting the works contained in web pages they aim at archiving or reusing.

We will expose the main copyright rules imposed at the European level and we will then study their implementation in Belgian and French laws. We will end the second part of this chapter by highlighting the main implications of copyright law for web archiving, from preservation, access, and reuse perspectives.

Scope of the European framework

One main European Directive (Directive 2001/29/CE) regulates copyrights across the European Union. This Directive has recently been modified and somewhat adapted to the current digital reality by another Directive (Directive (UE) 2019/790).

To portray the scope of copyright law in broad terms, it is firstly important to know that copyright, which is a kind of intellectual property along with trademarks and patents, is automatically awarded to the author of a given work once two conditions are respected: the work must be “expressed” (meaning that it is not a simple idea but it has actually been put in a

tangible form) and original. This entails that copyrights are very often awarded to authors for four reasons. First, the copyright is automatically awarded to the author of a work, which means that no administrative process has to be conducted to obtain it. Second, the notion of work is extremely wide: it goes from texts to images, songs, or videos. The length of the work does not matter, a sentence or a very short song or video could just as well be protected by a copyright as a big book or a film. The Court of Justice of the European Union even decided that a ten-word text could be awarded such a protection (EUCJ, Judgment *Infopaq International A/S c. Danske Dagblades Forening*, 16 July 2009). Third, the condition of “expression” of the work has the only effect of excluding mere ideas from the protection. An often-used example we can mention is the following: the scenario of a book portraying a young orphan wearing glasses and a lightning scar that just happens to also be a wizard, is not protected. The Harry Potter books, on the other hand, express that idea in a concrete form and are protected. Fourth, the condition of originality is not that hard to meet, or at least not as hard as one would think: no need here to take into account the amount of time spent on the creation of the work, nor its quality: it is enough to offer something that constitutes an intellectual creation expressing its author’s creativity. Regarding texts, such intellectual creation can be achieved through the choice, sequence, and combination of words. For pictures, the author must only have been able to express their creative abilities by making free and creative choices regarding the lighting, the positioning of the model, or the angle of the picture for example. Consequentially, the internet is filled with copyrighted work that is not necessarily the most obvious.

Regarding the territorial scope of European legislation applicable to copyright, it has been implemented in national law by all Member States. Regarding which work is protected by a given State’s copyright law, another European Regulation containing private international law rules must be applied, which we will not go into in this chapter.

Being the owner of a copyright involves the benefit of several rights, financial on the one hand, and moral on the other. Regarding the economic rights of the copyright holder, there are the reproduction rights, the right of communication to the public, and the distribution rights. As to the owner’s moral rights, they are not provided by European law. However, they tend to be the same across the world: the right of disclosure, the right of paternity, and the right to the integrity of the work are the main ones. The rules vary depending on the type of rights, some being allowed to be transferred or licensed, others not, but we will not go too much in-depth here.

Obligations of third parties regarding a protected work

When a work is protected by a copyright, one cannot use it in any way they like. Certain rules must be respected to reproduce or communicate a protected work, due to the exclusive nature of the rights awarded to the owner of the copyright.

Regarding the right of reproduction or communication to the public in particular, they entail, as their names suggest, that the copyright owner has to expressly authorize any reproduction or communication of their work.

However, some exceptions have been provided by European legislators as well as in many other countries, as it quickly appears that the consequences of copyrights would be too complicated to handle in many cases. For example, the exceptions of quotation or illustration for teaching (provided that the source and author are indicated) are internationally

The importance of legal requirements for web archives studies

recognized by the 1979 Bern Convention for the Protection of Literary and Artistic Works. At the European level, we can find exceptions for private use, parody, political speeches, and many others. Some exceptions are particularly interesting when it comes to archiving activities, we will come back on those particular applications later on.

Implementation in Belgian and French laws

Copyright law is regulated in the European Union by Directive 2001/29/EC on the harmonization of certain aspects of copyright and related rights in the information society, which in essence is meant to be implemented in every Member State's national law with a certain amount of leeway, it is set to vary across Europe. As we did for the data protection rules, we will focus here on the implementation of the copyright Directive in Belgian and French laws.

Among the elements of copyright law for which European states had some freedom and which are of interest to our subject, we can mainly mention some of the exceptions to the reproduction and distribution rights. Indeed, the European legislator has simply set a list of exceptions from which Member States could choose. The exceptions for “specific acts of reproduction made by publicly accessible libraries, educational establishments or museums, or by archives, which are not for direct or indirect economic or commercial advantage” and for “communication or making available, for the purpose of research or private study, to individual members of the public by dedicated terminals on the premises of publicly accessible libraries, educational establishments or museums, or archives of works and other subject-matter not subject to purchase or licensing terms which are contained in their collections” are particularly interesting with regard to web archiving practices and research within web archives and have the main impact on their feasibility. The first exception mentioned has recently been made mandatory and reworded by the European legislator in a new copyright Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. However, the precise scope of application of the exception is not imposed and can vary from state to state. Finally, the mandatory exception to the reproduction right for text and data mining for the purposes of scientific research has been added to the list by the same Directive and could be very useful to researchers working with web archives.

As to both Belgian and French laws, the exception to the reproduction right provided to the benefit of archives as well as the exception to the right of communication to the public for research purposes have been implemented. The first exception has not been modified following the adoption of the new Copyright Directive in either State. Its scope of application is however very limited in French law by articles L132-2 and L132-3 of the Heritage Code: it can only be used by the French National Library regarding its legal deposit obligation, while it can be used by all the establishments targeted by the European text in the Belgian Economic Law Code. Further specificities exist in French law regarding the application of the exception to the reproduction right to legal deposit, it is however quite complicated and covers a very particular case of web archiving, *i.e.* the web archiving achieved by the French National Library.

Regarding the text and data mining exception to the reproduction right, it has been implemented in both French and Belgian laws in similar ways, as the European legislator did not leave much freedom to Member States.

Consequences for web archiving

Given the easiness for a work to be protected by copyright, it is understood that the internet is filled with copyrighted works of many natures (Patel, 2007, p. 3). Any image or text that is sufficiently original is protected. As we have previously seen, exceptions do exist in some cases to avoid the complication of having to systematically request the copyright owner's authorization to reproduce or communicate their work. However, those exceptions did not initially consider the impact of digitalization on all copyright-related matters, which makes them useless in some cases.

As to the step of the preservation of the web, if some protected content is contained in the archiving institution's selection, it undoubtedly represents a reproduction of said work, requiring in principle the authorization of the copyright owner (Michel, 2020, p. 55). Considering the amount of protected work, as well as the fact that the copyright owner might not be easy to find, the application of an exception is crucial to make such activity manageable. The exception to the reproduction right for heritage preservation could therefore be useful in this precise case. The big question is however if it can apply to the reproduction of digital content. There is a controversy among scholars: according to a large part of them, this exception is only meant to allow the digitization of paper documents for heritage preservation, but it does not cover web harvesting. For others, it does. This is confirmed by one recital of the initial Directive (2001/29), according to which "an exception or limitation should not cover uses made in the context of online delivery of protected works" (40th recital). The new copyright Directive (2019/790), which makes this exception compulsory for all Member States and rewords it a bit, provides that it should "allow cultural heritage institutions to make copies of any works or other subject matter that are permanently in their collections, in any format or medium, for purposes of preservation of such works or other subject matter and to the extent necessary for such preservation" (article 6). While it takes the digital side of things into account by mentioning that such preservation can be done "in any format or medium", it also entails that the work must already be in the institution's collections for the exception to apply, which seems to exclude web harvesting from its scope, as the web content is precisely not in those collections before the harvesting takes place, making it a bit of a chicken and egg situation. Even though we could consider that an evolutive construction of this exception would be very much needed and could be justified in practice, there is no legal certainty on this matter and we could not predict how a judge would decide — which is an important hindrance to web archiving that should be resolved by the European legislator.

Regarding the access that could be given later down the line to the web archives, it represents a communication to the public of protected works if some are contained in the archives (Michel, 2020, p. 54). Again, the authorization of the copyright owner should be obtained in order for archiving institutions to give access to their web archives, which is hugely impractical. The exception to the right of communication for research purposes could be a nice solution here. However, some difficulties come from the very limited scale of the access provided, i.e. the premises of the establishment. This therefore excludes any kind of online communication of web archives without the copyright owner's authorization. On top of that, such communication must be limited to "purposes of research or private study" and must be made to individual members of the public, excluding a larger diffusion of the archives.

The importance of legal requirements for web archives studies

Regarding the consequences for researchers who intend to reuse those web archives, they will also have to be careful when reproducing protected content found in the archives, and later on giving access to the results of their research if they contain such content as well. However, they can benefit from one particular exception to the reproduction right that has been provided in the new Copyright Directive: an exception for reproductions and extractions made by research organizations and cultural heritage institutions to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access (Bottis, Papadopoulos, Zamakolas & Ganatsiou, 2019). One will however have to be aware the meet the scope of this exception, although it is larger than the exception to the reproduction right for heritage preservation.

If an online protected work is reproduced and archived by an archiving institution or a researcher without respecting the scope of one of the two exceptions mentioned (which is very limited regarding the exception for heritage preservation, and a bit wider for the text and data mining exception), or if it is made accessible to the public in a wider way than what is authorized by the applicable exception, the institution is at risk of being sued by the copyright owner, claiming damages for the violation of their economic rights. At a large scale and applied to an important amount of web archives, this can have a significant impact on any organization. It is therefore important to be aware of those risks while developing any web archiving process.

Conclusion

While this chapter is obviously not as precise and detailed as a purely legal article would be, we hope it still had the effect of highlighting the importance of personal data protection rules and copyright rules for any person or institution intending to archive web content or conducting research within web archives. We are aware that any legal subject may seem trivial to laymen, but the objective of this chapter is mainly to explain, in a practical way, what can go wrong if the law is ignored. Indeed, the famous adage “*nemo censetur ignorare legem*” (or “nobody is thought to be ignorant of the law”) is the rule in many states, including Belgium and France, but also Canada, Brazil, or the United States. And yes, it even applies to a very complicated and boring law.

Although web archiving is far from being a new practice, it appears that European legislator, as well as national legislators, do not always take its importance into account by adopting laws that would facilitate archiving institutions’ tasks as well as the research activities conducted within the web archives.

As the internet is filled with personal data and work protected by copyrights, it must be kept in mind that the process of archiving it has to respect the rights of the people involved in that content in one way or the other. Even though archiving activities and research pursue noble goals most of the time, they have to remain lawful and protective of others nonetheless.

References

- Bottis, M., Papadopoulos, M., Zampakolas, C., & Ganatsiou, P. (2019). Text and data mining in directive 2019/790/EU enhancing web-harvesting and web-archiving in libraries and archives. *Open Journal of Philosophy*, 9, 369–395.
- Michel, A. (2019). Le traitement des données à des fins archivistiques dans l’intérêt public. *DPO News*, 5, 6–9.

The Routledge Companion to Transnational Web Archive Studies

- Michel, A. (2020). Tour d’horizon sur les aspects légaux de l’archivage du web. *Cahiers de la Documentation*, 2020/2.
- Michel, A. (2021). Web archiving in the public interest from a data protection perspective. In J. Herveg (Ed.), *Deep Diving into Data Protection: 1979–2019: Celebrating 40 Years of Research on Privacy Data Protection at the CRIDS* (pp. 181–200). Larcier.
- Patel, K. (2007). “Authors v. Internet Archives”: The copyright infringement battle over web pages. SSRN. https://web.archive.org/web/20240308082716/https://papers.ssrn.com/sol3/papers.cfm?abstract_id=965199.
- Tambou, O. (2020). *Manuel de droit européen de la protection des données à caractère personnel*. Bruylant.
- Vanreck, O. (2018). Impacts du Règlement général sur la protection des données dans le domaine de l’archivage. *Le règlement général sur la protection des données (RGPD/analyse approfondie)*, *Cahiers du CRIDS*, 44. Larcier.

27

PUBLIC POLICIES, TECHNOLOGICAL INFRASTRUCTURE, AND USES OF WEB ARCHIVES BY THE DIGITAL HUMANITIES IN BRAZIL

Moisés Rockembach

Introduction

Websites play a vital role in sharing information with the public, and their content is often updated regularly. However, this constant change also means that valuable information can be lost or permanently removed. This can have negative consequences for research purposes and even governmental accountability. In today's digital age, we generate a massive amount of data in various formats. Despite this, only a few organizations truly appreciate the importance of preserving websites. Thankfully, there are a few notable initiatives that aim to tackle this problem.

Web archiving has been appreciated in several studies ranging from its specific theories and practices, such as its methodological approaches (Bragg et al., 2013; Németh & Drótos, 2019; Khan & Rahman, 2019; Gomes et al., 2021), and the specific ethical aspects of preserving web pages considering their presence in social life (Glanville, 2010; Rockembach, 2017; Mackinnon, 2021), to subjects related to the knowledge of the Digital Humanities (Brügger, 2018; Milligan, 2016, 2019; Weber, 2020; Fritz et al., 2021) and the use of web archives as a primary source for historical research (Brügger, 2011; Brügger, 2012; Rogers, 2018; Milligan, 2019).

The objective of this study is to identify the documents and actions that are related to the development of the website preservation policy, address legal and research challenges in web archives and Digital Humanities, using bibliographic and documentary research, as well as legislation analysis and case studies in the Brazilian scenario. Brazil has a variety of technical resources and specific legislation that addresses the need to preserve documents produced by the government (Brasil, 1988, 1991; Conarq, 2010, 2013) and has recently been working on a recommendation of minimum elements of a digital preservation policy for websites

and social media (Conarq, 2021b). In this sense, the National Council of Archives, which is associated with the National Archive and the Ministry of Justice, formed the Technical Consultative Chamber in 2021. The purpose of this chamber is to establish guidelines for conducting studies, presenting proposals, and finding solutions for the preservation of websites and social media. These documents were approved in December 2022 and will be published as a federal resolution. The research aims to demonstrate the advances, challenges, and gaps in the preservation of the web in the Brazilian context.

This chapter provides a comprehensive exploration of web archiving and preservation, focusing specifically on Latin America, notably Brazil. Divided into three sections, the first delves into diverse web archiving initiatives across Latin American countries, showcasing their challenges. The second section navigates the legal landscape governing web archiving in Brazil, analysing existing policies and regulations. The final section explores the intersection of web archiving and digital humanities in Brazil, highlighting the collaborative potential between these fields for innovative research and cultural heritage preservation. Together, these sections contribute to an understanding of web archiving's significance, challenges, and future prospects within the Latin American context.

Web archiving: Cases in Latin America

It is crucial to emphasize that even in countries with widespread internet usage, such as Brazil, there is still a need for more significant efforts in systematically archiving the web. Although examples of national sites can be found preserved in international initiatives such as the Internet Archive, studies reveal a lack of systematic preservation of Brazilian sites and the positive impacts of adopting projects on a national or regional scale (Ferreira et al., 2018; Rockembach, 2018; Rockembach & Serrano, 2021).

The scale of digital information production is exemplified by data from the Brazilian Institute of Geography and Statistics (IBGE). According to their findings, in 2016, 116 million people in Brazil, representing 64.7% of the population aged ten years and older, were connected to the internet (IBGE, 2018). Similarly, the Regional Center for Studies for the Development of the Information Society (Cetic.br), operated by CGI.br/NIC.br, reported in 2021 that 82% of Brazilian households had internet access, with 76% of the population utilizing the network (NIC.br, CGI.br, & Cetic.br, 2021). Additionally, the same institute revealed that all Federal and State public bodies in Brazil use the internet, with 90% of them having their websites (CGI.br, 2017).

Web archiving initiatives with international coverage, like the Internet Archive or Library of Congress, play a crucial role in preserving web content from various domains, including those from Latin American countries.¹ These resources and platforms provide valuable access to a wide range of web materials. However, it is equally important to encourage and support national initiatives that focus specifically on preserving local and national web domains. National initiatives can ensure comprehensive coverage of important cultural, historical, and government information for a specific country. By promoting and investing in such initiatives, nations can safeguard their digital heritage and maintain control over the preservation and accessibility of their web content.

There are some interesting examples in Latin America. In Colombia, the publication of the Mini/Manual on Social Media Archiving: Basic Concepts, Strategies, and Best Practices (Archivo General de la Nación, 2015a) offers a theoretical-practical approach to social media archiving, aiming to serve as a reference for the structuring and implementation of

preservation strategies for records produced in social media tools. This manual is aligned with the strategies of modernizing document management and E-Government in the country, providing essential guidelines, basic concepts, and best practices for the proper archiving of this information, ensuring its future availability.

Furthermore, the publication “Web Archiving: Basic Concepts, Strategies, and Best Practices” (Archivo General de la Nación, 2015b) also addresses basic concepts, strategies, and best practices, but with a specific focus on web content. The publication is divided into two parts to facilitate the reader’s understanding, the first part covers and defines the concepts related to web archiving, including its types, categories, and main challenges. Additionally, it mentions the main technological tools used in this context, along with relevant success cases globally. The second part summarizes, in five steps, the best practices and strategies to structure a web archiving project, ensuring the capture, organization, preservation, continuity, and access to the heritage recorded on the web for current and future generations.

Both documents were produced by the Archivo General de la Nación of Colombia, an institution responsible for the management and preservation of the country’s national archives. By developing the Mini/Manual on Social Media Archiving and the document Web Archiving, it demonstrates its commitment to providing relevant guidelines and orientations for archiving and preservation of records produced in social media and on the web. These documents serve as resources for professionals and institutions involved in document management and the preservation of Colombia’s historical and cultural heritage.

Another example of a Latin American website is the Latin American Government Documents Archive (LAGDA),² which preserves and provides access to ministerial and presidential documents from 21 Latin American and Caribbean countries (Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Guyana, Honduras, Jamaica, Mexico, Nicaragua, Panama, Paraguay, Peru, Uruguay, Venezuela). It contains copies of approximately 300 government websites, including official documents, videos, and speeches, where each country is followed by a list of ministries, departments, and agencies along with their respective website links. LAGDA is a collaboration between the University of Texas Libraries, The Nettie Lee Benson Latin American Collection, and the Latin American Network Information Center. Users can search or browse the archive by country or specialized collections and if a relevant website is missing, users can nominate it for inclusion.

México 2010 (Archive-It, 2010) is another example and represents a curated compilation of web-based resources, encapsulating the commemoration of Mexico’s bicentennial celebration of independence and the centennial observance of the Mexican Revolution in the year 2010.

The Venezuelan Political Discourse Archive (ARVEPODIS) has an extensive set of web-based resources, including over 100 websites intrinsically linked to the Venezuelan political scene. Within its repertoire, ARVEPODIS covers a diverse range of platforms, including blogs, media, official government portals, political opposition outlets, as well as non-partisan and civil society groups, among others. The archival process involved the daily capture of numerous websites over a concentrated period of three months, specifically in late 2005. ARVEPODIS stands as a repository for the exploration and analysis of the multifaceted Venezuelan political discourse.

The Archive of Political Parties and Elections in Latin America — APPELA (Archive of Political Parties and Elections in Latin America, 2005) endeavours to uphold the preservation

and accessibility of web-based resources, documents, and other materials from a diverse array of Latin American political parties. Currently, APPELA has an extensive collection that covers more than 100 political party websites, alongside those belonging to national electoral authorities from multiple countries. These archival captures were executed on various occasions throughout the year 2005.

The National Library of Chile³ is responsible for collecting and preserving the nation's printed, audio-visual, and electronic productions. Since 2008, the library has been working on various strategies to collect, preserve, and provide access to electronic newspapers. This includes implementing an electronic deposit policy, a procedural manual, and an FTP server for receiving archives. The web archiving initiative at the National Library began in late 2013 with a pilot project focusing on 10 websites related to presidential campaigns and citizen campaigns. In subsequent years, the scope expanded to include 23 registered media outlets, with a focus on regional press from 15 cities across the country.

While countries in Latin America, like Chile, have implemented several web archive initiatives, Brazil currently lacks a comprehensive website preservation policy (Melo et al., 2019; Melo & Rockembach, 2021; Rockembach, 2018). Nevertheless, progress has been made recently toward the creation of a national policy for website preservation in Brazil.

Policies and regulations on web archiving and web preservation in Brazil

The initial establishment of legal deposit regulations in Brazil can be traced back to Imperial Decree No. 433, enacted on July 3, 1847 (Camara dos Deputados, 1847). Nowadays, the legal framework that regulates the practice of legal deposit is referred to as the “*Lei Brasileira do Depósito Legal*” or the Brazilian Law of Legal Deposit, formally designated as Law No. 10.994/2004. This legislation was formulated and enacted by the executive branch of the Brazilian government with the objective of regulating the deposition of publications in the National Library. The primary aim of this legal deposit system is to ensure the systematic cataloguing and safekeeping of domestic intellectual works while facilitating the management, production, and dissemination of contemporary Brazilian literature. Furthermore, it serves to safeguard and preserve the national language and cultural heritage. However, current legal provisions do not cover the preservation of national websites and domains.

According to a study conducted by Luz (2016), it was found that 9 out of the 27 Brazilian capitals had either deleted websites or websites that were not publicly accessible. This highlights the need to safeguard the digital collection, and official websites created by public authorities. Therefore, the proposed legislation that we describe below aims to address this critical issue by protecting our digital heritage and ensuring that society has access to significant historical records.

In another study, Luz (2021) highlighted issues related to the preservation and access of official information. The research also discussed the preservation and archiving of web content and concluded that the official website of the Presidency of the Republic of Brazil does not fully meet the expected standards for information and memory rights in democracies. This emphasizes the need for implementing a Communicational Memory policy.

We must emphasize the importance of Bill no. 2.431/2015, authored by Luizianne Lins (PT-CE), which deals with the issue of institutional digital public heritage on the internet, such as official websites. This proposed legislation aims to deem the unauthorized modification of the original structure of government websites by officials as an illegal act. The

primary goal of this bill is to safeguard public access to information generated by the Brazilian government.

To preserve the history and culture of society, the bill presented aims to protect the digital collection produced by the Government, which includes images, videos, audio, and texts. The bill sought to solve the problem of erasing the websites of their predecessors by managers, which, in addition to problems of public accountability, also creates difficulties for citizens in accessing important historical records (Lins, 2015).

An official website, as defined by the bill, refers to any internet site linked to direct or indirect government agencies. According to the proposed legislation, the heads of public powers would be responsible for preserving and maintaining institutional digital content. It is worth mentioning that, although Law no. 12.527/2011, also known as the Access to Information Law (LAI), guarantees access to information produced by public bodies, currently there is no legal regulation that guarantees the maintenance of this information online, such as institutional websites.

The official websites of public powers play a crucial role in facilitating communication between the government and citizens, as well as providing access to information of public interest and contributing to the construction of institutional memory. However, preserving web pages presents a significant challenge due to the absence of legislation ensuring their maintenance online and with access over time.

Certain bills currently under analysis in Brazil are also concerned with the preservation and custody of online content, such as PL 3050/2020.⁴ This proposed law aims to amend article 1,788 of Law no. 10,406, enacted on January 10, 2002, which establishes the Civil Code. The objective is to address the succession of assets and digital accounts belonging to the deceased, ensuring the transfer of all valuable digital content to the rightful heirs. According to the proposed legislation, the heirs will inherit all content contained in the deceased's digital accounts or digital files. This provision ensures that valuable online assets are effectively transmitted to the rightful beneficiaries, enabling them to access and manage the digital legacy left behind. The same applies to PL 3051/2020,⁵ which, in this instance, proposes amendments to the Civil Rights Framework for the Internet. Its objective is to address the allocation of Internet application accounts following the death of the account holder.

In 2010, the National Archives of Brazil introduced the Permanent Program for Preservation and Access to Digital Archival Documents, commonly referred to as AN Digital. As part of this initiative, the program published the Digital Preservation Policy on its website, with versions available from 2012 and 2016. Both these documents emphasize the significance of safeguarding digital documents. However, they also specify that the preservation efforts will initially focus on specific types of documents, such as text, images, audio files, audio-visual content, email messages, slideshows, spreadsheets, and relational databases. However, the document called AN Digital recognizes that in the near future, more complex digital document formats, including multimedia and web pages, will also be considered for preservation (Arquivo Nacional do Brasil, 2016).

In 2021, the National Council of Archives of Brazil (CONARQ) within the national executive government sphere took the initiative to establish a technical chamber (Conarq, 2021a). The primary aim of this chamber was to develop guidelines for conducting studies, presenting proposals, and finding solutions for preserving websites and social media (Conarq, 2021b). In line with its overall objectives, the technical chamber has produced two key documents: (i) the Website and Social Media Preservation Policy, and (ii) a recommendation

outlining essential elements for the digital preservation of websites and social media (Conarq, 2021b). To facilitate its fast implementation, the technical chamber has formulated a set of clear and concise guidelines in the form of regulations. Instead of providing minimal policy recommendations for this purpose, they have proposed the establishment of a Digital Preservation Policy for websites and social media, aligning with the concept of micropolitics (Conarq, 2022).

In addition to publishing a Digital Preservation Policy for websites and social media (Conarq, 2023a), specific minimum requirements for preserving these digital objects have been defined (Conarq, 2023b). Both documents provide detailed information on the preservation techniques specifically tailored to websites. It outlines a step-by-step process for the macro-level preservation of collections, taking into account the unique characteristics of websites and social media (Conarq, 2022).

Web archiving and digital humanities in Brazil

Digital humanities can be seen as a community of practices that combines computing and humanities (Terras, 2006). The concept of community is highly significant in the digital humanities, as it plays an important role in advancing best practices and conducting research on web archiving. Moreover, it contributes to enhancing the technologies used in the digital preservation process.

According to Brügger (2021), it is essential for web archives to provide their collections as research data in the field of Digital Humanities. Presently, many web archives offer restricted access to their files. However, Brügger suggests that by combining metadata and hyperlink data from web archives, relevant insights can be analysed regarding the structure of a web domain. The author emphasizes the need for web archives to be ready for the implementation of digital methods and also highlights the importance of transparency in providing information about the origin of web archives. This transparency is necessary for a more comprehensive understanding of analysis results. At the same time, one should not overlook the issues related to data protection when dealing with preserved web content (Rockembach & da Silva, 2021).

In an age characterized by an abundance of information, the methodologies and technologies employed are important in the analysis of digital materials. According to Pimenta (2022), this ongoing transformation has had a significant impact on the practices of historians, challenging traditional notions of reading and writing. Pimenta argues that linear and textual approaches are no longer enough, as the act of reading and writing now includes visual elements, non-linear structures, and multidimensionality.

The advent of the “digital turn” has brought about a profound transformation in the field of Humanities, especially in History. It calls for contemporary historians to possess the skills necessary to navigate diverse digital sources and integrate different forms of information. To adapt to these changes, historians must move away from the two-dimensional paradigm of relying solely on paper-based or word-processing methods to construct historical narratives. In this sense, According to Pimenta (2022), the contemporary historian must embrace a multidimensional approach that transcends traditional boundaries and embraces the complexity of digital records.

Balbi (2024) emphasizes that the relationship between the web and history is explored by three distinct subfields: the “History Web,” focused on assisting historians through digitized sources; the “Web History,” centred on websites, the material dimension of the past web,

and its archive; and the “History of the Web,” which concentrates on the political, economic, technical, and cultural history of the web itself. The author examines arguments both in favour and against the combination of these subfields, pointing out differences in theoretical foundations, methodologies, and disciplinary categorizations. Despite disparities, similarities, such as an interest in the historical dimension of the web and the perception of marginalization within their respective fields, suggest potential benefits of integration. The chapter advocates for interdisciplinary collaboration and the formation of an autonomous field, potentially named “History and the Web,” to enhance research questions, theoretical frameworks, and the utilization of sources across these subfields (Balbi, 2024).

Dantas (2014) is recognized as one of the pioneering works that utilized web archives as a research subject in a Brazilian thesis. This study investigated the examination of contemporary practices related to web archiving, preserving, and accessing the historical records of the Brazilian web. The research provided insights into the mnemonic processes entailed in operations supported by digital means and brought attention to the interconnectedness of technical and cultural factors that influence how we archive and access the recent history of the web. Dantas asserts that the convergence of digital memory and digitization initiatives in Brazil has tangible consequences, such as the absence of web page collections in national institutions. To tackle this problem, the research team developed the “Buscas.br” collection, which comprises search engine pages from Brazil covering the years 1997 to 2013. This collection offers a means to engage with the materiality of digital sources.

Creating digital infrastructures to preserve national web content is an important step in the development of web archives. In this sense, some work has been carried out in the field of academic research. The Research Group of Web Archiving and Digital Preservation⁶ (NUAWEB), established by the Federal University of Rio Grande do Sul in 2017, has been actively engaged in various areas of web archiving and Digital Humanities. The group has been dedicated to exploring national and international initiatives, policies, and technologies involved in web archiving processes (Nuaweb, 2023). Their research endeavours primarily focus on critical aspects such as preservation, long-term accessibility, and the effective utilization of digital objects on the web. Drawing on the expertise from disciplines including Archival Science, Librarianship, Information Science, Communication, and Computer Science, the group’s investigations are enriched by diverse perspectives and multidisciplinary approaches.

One of the projects of the Research Group of Web Archiving and Digital Preservation (Nuaweb) is the Arquivo.ong.br⁷ which aims to foster a community focused on web preservation practices, including the capture, storage, and accessibility of web archives in accordance with international standards. Employing crawlers and using ISO formats for storage (WARC files), it can be considered the first web archiving initiative in Brazil. The activities also include the dissemination of software used by the international community and the exchange of knowledge and practices in digital preservation.

In Brazil, the initiative to establish a collaborative network for public discourse on the present and future of Brazilian Digital Humanities, known as the Collaborative Network for Digital Humanities (Colab HD+), was launched in 2023. It aims to promote the utilization of digital sources in scientific research, including web archives.⁸

Some works have been using web archives as a source of research in Digital Humanities. Laitano’s (2021) research focused on the analysis of two distinct types of digital sources: digitized documents and digital native files. The study first delved into the practice of digitizing document collections and establishing online repositories to ensure their

preservation and accessibility. As an illustrative example, Laitano highlighted the project “Brasil: Nunca Mais Digit@l,” which gives access to documents referring to the military dictatorship in Brazil.

The research also examined the archives of digital-born records, including publications found on social networks, websites, emails, and exclusively internet-published scientific articles. To exemplify this aspect, Laitano cited the website “Memórias da Ditadura,” which offers information and content related to the period of dictatorship in Brazil. Throughout the study, significant emphasis was placed on exploring the relationship between digital humanities and the preservation and accessibility of historical sources, with particular attention given to the potential of web archiving.

The study conducted by Fazano (2022) emphasized the importance of having technical guidelines to effectively guide the web archiving process. This need is even more evident when Fazano compares it with the policies of institutions in the USA, England, France, and other countries in Europe and Asia, which already have policies in place for capturing and preserving online content.

The methodology employed in this study involved analysing existing collections developed by the Library of Congress. These collections included materials such as the 2018 presidential elections, a collection on *Cordel* literature, and the 2016 Olympic and Paralympic Games. Fazano considers that web-based content differs from traditional collections found in museums, archives, or libraries due to its digital nature. Therefore, archiving procedures must be adapted to ensure the proper capture of information, this adaptation includes considerations such as the number of captures for a specific page and the desired level of navigability. For instance, a newspaper article may only need to be captured once, while a portal covering a particular subject may require more comprehensive captures, including all available links. Additionally, scheduled captures for different dates may be necessary.

Overall, Fazano’s study underscored the need for technical guidelines in web archiving, especially when compared to the existing policies in institutions around the world. The study’s methodology involved analysing collections from the Library of Congress, highlighting the challenges posed by web-based content and the necessity to tailor archiving procedures accordingly.

Another example can be mentioned in relation to the Brazilian elections. During the 2018 presidential election campaign, Facebook experienced a remarkable surge in popularity as a leading advertising platform, often surpassing the designated airtime of candidates’ television programs. Recognizing the significance of the published content, some authors have chosen to delve into web archiving of social media in Brazil as a research topic (Rockembach & Ferreira, 2018; Ferreira & Rockembach, 2019). Their studies were primarily focused on digitally preserving the videos produced by the presidential candidates during the 2018 elections. As the elections are a public event, all publications on this platform held public interest, impacting society both within and beyond the country’s borders.

Facebook was selected as the primary subject of analysis by Rockembach and Ferreira (2018) due to the presence of fan pages for all candidates, with content published on the platform being deemed freely available for use. The candidates’ deliberate choice to heavily utilize social media highlights a growing trend of personalization among politicians as they strive to establish a closer connection with the public.

Web archiving entails the preservation of all the content accessible on a web page with the assistance of specialized software. Typically, such preservation initiatives are undertaken

by independent efforts originating from universities and libraries in other countries. These initiatives aim to safeguard materials for future research. Remarkable examples include the Internet Archive platforms in the United States, Arquivo.pt in Portugal, and the Library and Archives of Canada. It is extremely important to preserve the contents published exclusively on the web and social media, as they constitute a vital component of today's societal history, vulnerable to modification or deletion as if they never existed.

What proves the rise of debates about web archives and their uses in the humanities is that, in recent years, several significant events in Brazil have focused on the topic of web archiving, most of them linked to the Humanities. These include the Brazilian Internet Forum 2019 — 9th Edition — Preservation of Brazilian Web Content, the 1st and 2nd Symposia on Policies and Strategies for Content Preservation on the web, which were part of the International Congress on Information Technology and Organization (TOI) in 2020 and 2021 respectively, the webinar titled “Archiving and Preservation of the web — Changing Experiences” organized by the research group Driade, and the Brazilian Teaching and Research Meeting in Archival Science VII REPARQ 2022 — Web archiving and digital preservation: from websites to social networks, organized by the National Forum for Teaching and Research in Archival Science — FEPARQ. Brazil has also participated in international discussions, mainly in the latest events of the International Internet Preservation Consortium (IIPC) through the Web Archiving Conference, in the latest editions in 2019 in Croatia, in 2021 online, and in 2023 in the Netherlands.⁹

Conclusion

From the 1990s onwards, a multitude of endeavours have emerged to safeguard segments of the internet's content. This includes initiatives such as the Internet Archive, national libraries and archives, university research institutions, laboratories, and governmental establishments (Brügger, 2018; Ogden, 2022). Given the scientific, cultural, and intellectual importance of web archiving, several countries have begun to establish their own programmes to preserve published web content. Web archives are widely regarded as indispensable interventions to maintain accessibility to web pages and online social networks, serving as invaluable resources for the field of Digital Humanities. Furthermore, there exists considerable research exploring approaches to web archiving, as well as investigating the ethical dimensions associated with it.

The internet challenges geographical boundaries due to its nature as a global network of interconnected computer systems and devices. Consequently, web content transcends local or regional limitations, allowing anyone, anywhere to create and access web pages with information and resources from across the globe, not restricted to a national domain.

Furthermore, Rockembach and Pavão (2024) point out the significant geographical inequality in the preservation of web resources and content, underlining the need for a more integrated approach that considers both national and transnational strategies to effectively archive the web. This highlights the importance of developing collaborative frameworks that can accommodate the web's borderless nature while addressing the disparities in digital preservation efforts.

There are several important issues that need attention, with the main concern being the lack of web archiving initiatives in Latin America compared to other regions. Although Brazilian content is occasionally preserved through international efforts, preservation work lacks systematization and national and regional coverage. Therefore, it is necessary to have

national initiatives to address this gap. Getting involved in website preservation does not have to be limited to institutionalized web archiving initiatives, but it is highly recommended. By taking a community-oriented approach aligned with the field of digital humanities, there are various ways to contribute. This includes collaborating with research groups on scientific investigations or conducting case studies related to web archiving practices. These practices involve carefully selecting content, capturing and storing it, digitally preserving it, and making the archived material accessible. Digital humanities focuses on interdisciplinary and transdisciplinary methods, promoting productive discussions between different areas of knowledge. This approach enhances our understanding of contemporary digital phenomena. Research in digital humanities requires knowledge of software and methodologies that may differ from traditional humanities approaches, including the adoption of digital research methods. Within this context, web archives provide a valuable opportunity to document the present and gain insights into society from a retrospective perspective.

Notes

- 1 For instance, <https://web.archive.org/web/20231115163945/www.loc.gov/web-archives/index/location/>.
- 2 Some websites can be retrieved at <https://web.archive.org/web/20231115162914/http://lanic.utexas.edu/project/archives/lagda/>.
- 3 Biblioteca Nacional Digital de Chile. (2023). Archivo web [Web archive]. Retrieved from <https://web.archive.org/web/20231115162329/http://archivoweb.bibliotecanacionaldigital.cl/>.
- 4 Please note that the legislation is still in progress, and you can access the stages of its legislative process through the website of the Brazilian Chamber of Deputies: <https://web.archive.org/web/20231115162042/www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2254247>.
- 5 Please note that the legislation is still in progress, and you can access the stages of its legislative process through the website of the Brazilian Chamber of Deputies: <https://web.archive.org/web/20231115161819/www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2254248>.
- 6 Ongoing projects and research can be found at <https://web.archive.org/web/20231115161719/www.ufrgs.br/nuaweb/>.
- 7 Further information, research findings, and materials regarding web archiving in Brazil are consolidated and can be accessed at <https://web.archive.org/web/20231115161325/http://arquivo.org.br/>.
- 8 Rede Colab HD+ (2023). [Brazilian Digital Humanities Network]. Retrieved from <https://web.archive.org/web/20231115164132/http://colabhd.org/>.
- 9 For a complete record of communications, all materials can be found on the website of Arquivo.org.br at <https://web.archive.org/web/20231115161325/http://arquivo.org.br/>.

References

- Archive of Political Parties and Elections in Latin America. (2005). Archive of Political Parties and Elections in Latin America — APPELA. Retrieved November 15, 2023, from <https://web.archive.org/web/20231115162523/http://lanic.utexas.edu/project/archives/appela/>.
- Archive-It. (2010). Mexico 2010 Collection. Mexico's bicentennial celebration of independence. Retrieved November 15, 2023, from <https://web.archive.org/web/20231115162621/www.archive-it.org/collections/1097>.
- Archivo General de la Nación. (2015a). Mini/Manual Archivamiento Web: Conceptos básicos, estrategias y mejores prácticas. Bogotá, Colombia. Retrieved November 15, 2023, from https://web.archive.org/web/20231115163300/www.archivogeneral.gov.co/sites/default/files/Estructura_Web/5_Conulte/Recursos/Publicacionees/ArchivamientoWeb.pdf.

- Archivo General de la Nación. (2015b). Mini/Manual Archívamento de Medios Sociales: Conceptos básicos, estrategias y mejores prácticas. Bogotá, Colombia. Retrieved November 15, 2023, from https://web.archive.org/web/20231115164037/www.archivogeneral.gov.co/sites/default/files/Estructura_Web/5_Conulte/Recursos/Publicacionees/ArchivamientoDeMediosSociales.pdf.
- Arquivo Nacional do Brasil. (2016). Política de preservação digital. Version 2.
- Balbi, G. (2024). “History Web,” “Web History,” and “History of the Web.” Three subfields and why (and why not) combining them. In S. Aasman, A. Ben-David, & N. Brügger (Eds.), *The Routledge Companion to Transnational Web Archive Studies*. Routledge.
- Bragg, M., Hanna, K., Donovan, L., Hukill, G., & Peterson, A. (2013). The web archiving life cycle model: White Paper. Internet Archive. Retrieved November 15, 2023, from <https://web.archive.org/web/20231115184553/https://blog.archive.org/wp-signup.php?new=ait>.
- Brasil. (1988). Constituição da República Federativa do Brasil de 1988, Brasília. Retrieved September 13, 2024, from www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm
- Brasil. (1991). Lei Federal 9.159/1991. Dispõe sobre a política nacional de arquivos públicos e privados e dá outras providências. Retrieved September 13, 2024, from www2.camara.leg.br/legin/fed/lei/1991/lei-8159-8-janeiro-1991-322180-norma-pl.html
- Brügger, N. (2011). Web archiving — between past, present, and future. In M. Colsalvo & C. Ess (Eds.), *The Handbook of Internet Studies* (pp. 24–42). Wiley-Blackwell.
- Brügger, N. (2012). When the present web is later the past: Web historiography, digital history, and internet studies. *Historical Social Research/Historische Sozialforschung*, 37(4), 102–117.
- Brügger, N. (2018). *The Archived Web: Doing History in the Digital Age*. MIT Press.
- Brügger, N. (2021). Digital humanities and web archives: Possible new paths for combining datasets. *International Journal of Digital Humanities*, 2(1/3), 145–168.
- Camara dos Deputados. (1847). Decreto nº 433, de 3 de Julho de 1847. Retrieved November 15, 2023, from <https://web.archive.org/web/20231115162233/www2.camara.leg.br/legin/fed/decret/1824-1899/decreto-433-3-julho-1847-560144-publicacaooriginal-82761-pl.html>.
- CGI.br. (2017). Pesquisa TIC Governo Eletrônico. Retrieved November 15, 2023, from <https://web.archive.org/web/20231115165558/https://cetic.br/pesquisa/governo-eletronico/>.
- Conarq. (2010). Conselho Nacional de Arquivos. Resolução nº 32, de 17 de maio de 2010. Dispõe sobre a inserção dos Metadados na Parte II do Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos — e-ARQ Brasil. Retrieved September 13, 2024, from <https://web.archive.org/web/20220331122220/www.gov.br/conarq/pt-br/legislacao-arquivistica/resolucoes-do-conarq/resolucao-no-32-de-17-de-maio-de-2010>
- Conarq. (2013). Conselho Nacional de Arquivos. Resolução nº 38, de 9 de julho de 2013. Dispõe sobre a adoção das “Diretrizes do Produtor — A Elaboração e a Manutenção de Materiais Digitais: Diretrizes Para Indivíduos” e “Diretrizes do Preservador — A Preservação de Documentos Arquivísticos digitais: Diretrizes para Organizações.” Retrieved September 13, 2024, from <https://web.archive.org/web/20240809134034/www.gov.br/conarq/pt-br/legislacao-arquivistica/resolucoes-do-conarq/resolucao-no-38-de-9-de-julho-de-2013>
- Conarq. (2021a). Portaria CONARQ no 131, de 9 de novembro de 2021. Institui a Câmara Técnica Consultiva para definir diretrizes para a elaboração de estudos, proposições e soluções para a preservação de websites e mídias sociais. Rio de Janeiro. Retrieved September 13, 2024, from <https://web.archive.org/web/20240123082819/in.gov.br/en/web/dou/-/portaria-conarq-n-131-de-9-de-novembro-de-2021-359118428>
- Conarq. (2021b). Requerimento ao Conselho Nacional de Arquivos instituindo câmara técnica consultiva para a elaboração de estudos, proposições e soluções para a preservação de websites e mídias sociais. Retrieved September 13, 2024, from <https://web.archive.org/web/20240913192631/www.gov.br/conarq/pt-br/assuntos/camara-tecnica-consultiva/para-definir-diretrizes-para-a-elaboracao-de-estudos-proposicoes-e-solucoes-para-a-preservacao-de-websites-e-midias-sociais-1/para-definir-diretrizes-para-a-elaboracao-de-estudos-proposicoes-e-solucoes-para-a-preservacao-de-websites-e-midias-sociais>
- Conarq. (2022, December 7). 104a reunião ordinária do CONARQ — 07 de dezembro de 2022, quarta-feira, das 15h às 17h. Retrieved May 19, 2023, from <https://web.archive.org/web/20231115161041/www.facebook.com/ConselhoNacionaldeArquivos/videos/678607297071001>.

- Conarq. (2023a). Resolução nº 52, de 25 de agosto de 2023. Estabelece a Política de Preservação de Websites e Mídias Sociais no âmbito do Sistema Nacional de Arquivos (Sinar). Retrieved March 17, 2024, from www.gov.br/conarq/pt-br/legislacao-arquivistica/resolucoes-do-conarq/resolucao-no-52-de-25-de-agosto-de-2023.
- Conarq. (2023b). Resolução nº 53, de 25 de agosto de 2023: Define requisitos mínimos de preservação para websites e mídias sociais no âmbito do Sistema Nacional de Arquivos (Sinar). Retrieved March 17, 2024, from www.gov.br/conarq/pt-br/legislacao-arquivistica/resolucoes-do-conarq/resolucao-no-53-de-25-de-agosto-de-2023.
- Dantas, C. G. (2014). Criptografias da memória: um estudo teórico-prático sobre o arquivamento da web no Brasil (Doctoral dissertation). Universidade Federal do Estado do Rio de Janeiro.
- Fazano, I. F. (2022). Arquivamento de páginas da web como recurso de preservação digital: estudo de caso das coleções de sites brasileiros da Library of Congress (Master's thesis). Fundação Getúlio Vargas.
- Ferreira, L. B., Martins, M. R., & Rockembach, M. (2018). Usos do arquivamento da web na comunicação científica. *Prisma.com*, 36, 78–98.
- Ferreira, L. B., & Rockembach, M. (2019). Preservação de mídias sociais e arquivamento da web: um estudo acerca das eleições presidenciais brasileiras de 2018. In *Encontro Nacional de Pesquisa em Ciência da Informação* (pp. 1–13). ANCIB.
- Fritz, S., Milligan, I., Ruest, N., & Lin, J. (2021). Fostering community engagement through Datathon events: The archives unleashed experience. *Digital Humanities Quarterly*, 15(1). Retrieved September 13, 2024, from <http://hdl.handle.net/10315/38257>
- Glanville, L. (2010). Web archiving: Ethical and legal issues affecting programmes in Australia and the Netherlands. *The Australian Library Journal*, 59(3), 128–134.
- Gomes, D., Demidova, E., Winters, J., & Risse, T. (2021). *The Past Web*. Springer International Publishing.
- IBGE. (2018). Pesquisa Nacional por amostra de domicílio. Retrieved May 19, 2023, from www.ibge.gov.br/estatisticas/sociais/populacao/9127-pesquisa-nacional-por-amostra-de-domicilios.html.
- Khan, M., & Rahman, A. U. (2019). A systematic approach towards web preservation. *Information Technology and Libraries*, 38(1), 71–90.
- Laitano, B. G. (2021). *Digitalizar o arquivo, arquivar o digital: a história e suas fontes diante das velhas e novas tecnologias* (Master's thesis). Universidade Federal do Rio Grande do Sul.
- Lins, L. (2015). Projeto de Lei 2431/2015. Dispõe sobre o patrimônio público digital institucional inserido na rede mundial de computadores e dá outras providências. In *Câmara dos Deputados do Brasil*. Retrieved September 13, 2024, from www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=1594241&fichaAmigavel=nao
- Luz, A. J. A. (2016). *Comunicação pública e memória das cidades: A preservação dos sistemas de comunicação nos sites das capitais brasileiras* (Master's thesis). Universidade Federal do Rio Grande do Sul.
- Luz, A. J. A. (2021). *Comunicação pública e memória comunicacional: revelações e apagamentos sobre o governo da presidenta Dilma Rousseff* (Doctoral dissertation). Universidade Federal do Rio Grande do Sul. Retrieved November 15, 2023, from <https://web.archive.org/web/20231115165857/https://lume.ufrgs.br/handle/10183/235405>.
- Mackinnon, K. (2021). Ethical approaches to youth data in historical web archives. *Studies in Social Justice*, 15(3), 442–449.
- Melo, J. F., Nunes, L. A. N. D. O., & Rockembach, M. (2019). Preservação de websites governamentais a partir do arquivamento da web: Abordagens e metodologias. In *Encontro Nacional de Pesquisa em Ciência da Informação* (pp. 1–14). ANCIB.
- Melo, J. F., & Rockembach, M. (2021). International Initiatives and Advances in Brazil for Government Web Archiving. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST* (Vol. 378). Retrieved November 15, 2023, from https://web.archive.org/web/20231115170115/https://link.springer.com/chapter/10.1007/978-3-030-77417-2_6.
- Milligan, I. (2016). Lost in the infinite archive: The promise and pitfalls of web archives. *International Journal of Humanities and Arts Computing*, 10(1), 78–94.

- Milligan, I. (2019). *History in the Age of Abundance? How the Web Is Transforming Historical Research*. McGill-Queen's University Press.
- Németh, M., & Drótos, L. (2019). A Blended Learning-Based Curriculum on Web Archiving in the National Széchényi Library. *Digital Library Perspectives*.
- NIC.br, CGI.br, & Cetic.br. (2021). Resumo executivo — Pesquisa TIC Domicílios 2021. Retrieved November 15, 2023, from <https://web.archive.org/web/20231115170553/https://cetic.br/pt/publicacao/resumo-executivo-pesquisa-sobre-o-uso-das-tecnologias-de-informacao-e-comunicacao-nos-domicilios-brasileiros-tic-domicilios-2021/>.
- Nuaweb. (2023). NUAWEB — Núcleo de Pesquisa em Arquivamento da Web e Preservação Digital. Retrieved Nov 15, 2023, from <https://web.archive.org/web/20231115161719/www.ufrgs.br/nuaweb/>.
- Ogden, J. (2022). “Everything on the internet can be saved”: Archive Team, Tumblr and the cultural significance of web archiving. *Internet Histories*, 6(1–2), 113–132.
- Pimenta, R. M. (2022). Fontes demais, tempo de menos: uma breve crítica à escalada tecnoinformacional para a escrita da História do tempo presente. In T. L. Nicodemo, A. Rota & I. K. Marino (Eds.), *Caminhos da História digital no Brasil*. Ed. Milfontes.
- Rockembach, M. (2017). Inequalities in digital memory: Ethical and geographical aspects of web archiving. *The International Review of Information Ethics*, 26.
- Rockembach, M. (2018). Arquivamento da web: Estudos de caso internacionais e o caso Brasileiro. *RDBCI: Revista Digital de Biblioteconomia e Ciência Da Informação*, 16(1), 7–24.
- Rockembach, M. (2019). Arquivamento da Web no contexto das humanidades digitais: Da produção a preservação da informação digital. *LIINC Em Revista*, 15(1) (Maio 2019), 131–139.
- Rockembach, M., & da Silva, A. M. (2021). Web data and the relationship between the general data protection regulation in Europe and Brazil. In P. F. Anunciação, C. R. M. Pessoa, & G. L. Jamil (Eds.), *Handbook of Research on Digital Transformation and Challenges to Data Security and Privacy* (pp. 222–233). IGI Global.
- Rockembach, M., & Ferreira, L. B. (2018). Web archiving of elections and Brazilian possibilities. In *Challenges and Opportunities for Knowledge Organization in the Digital Age, Proceedings of the Fifteenth International ISKO Conference 9–11 July 2018 Porto, Portugal* (pp. 969–972). Ergon-Verlag.
- Rockembach, M., & Pavão, C. M. G. (2024). *Arquivamento da web e preservação digital*. Pimenta Cultural. Retrieved March 17, 2024, from <https://doi.org/10.31560/pimentacultural/2024.98119>.
- Rockembach, M., & Serrano, A. (2021). Climate change and web archives: An Ibero-American study based on the Portuguese and Brazilian contexts. *Records Management Journal*, 31(3), 222–239. <https://doi.org/10.1108/RMJ-11-2020-0039>
- Rogers, R. (2018). Periodizing web archiving: Biographical, event-based, national and autobiographical traditions. In N. Brügger & I. Milligan (Eds.), *The SAGE Handbook of Web History* (pp. 42–56). SAGE.
- Terras, M. (2006). Disciplined: Using educational studies to analyze “humanities computing.” *Literary and Linguistic Computing*, 21(2), 229–246.
- Weber, M. S. (2020). *Web Archives: A Critical Method for the Future of Digital Research*. WARCNet Papers, https://web.archive.org/web/20230902214753/https://cc.au.dk/fileadmin/user_upload/WARCnet/Weber_Web_Archives_A_Critical_Method.pdf.

28

CONVERSATION 5

The future

*Friedel Geeraert, Michael Kurzmeier, Lise-Anne Denis,
Valérie Schafer, Márton Németh, Moisés Rockembach,
Saskia Huc-Hepher, Susan Aasman, Matthew Weber,
Jane Winters, and Janne Nielsen*

How do you imagine the future of transnational web archive studies?

- Friedel Geeraert:* I hope that we will be able to continue the work that was started in the WARCnet network in the context of a new research framework in Europe, and beyond, thereby helping to promote the value of web archives for research. As an historian, I already look forward to the first comparative transnational historical studies based (partly) on archived web content that will appear in the coming decades.
- Michael Kurzmeier:* The growing unavailability of content, largely through continuing platformisation and algorithmically curated feeds will be challenges for future web archive studies. As a response, I think we are going to see smaller and more focussed web archives. Additionally, the question of how to deal with AI-generated content in web archives will become increasingly important for curation work. These challenges will hopefully be addressed by a European web archive research framework.
- Lise-Anne Denis:* I imagine a proper legal framework taking web archiving and web archive studies into account and lifting all obstacles law currently involves in this field.
- Valérie Schafer:* Globalisation has ushered in an era where the transnational flow of information, user-generated content, and digital cultures is extremely prevalent. Recent or ongoing crises, whether they be health-related, environmental, or geopolitical, further emphasise the interconnect-edness of our planet and the imperative of transnational cooperation and analysis. While significant efforts have been made in this direction, such as the International Internet Preservation Consortium's (IIPC) initiatives in curating international collections on various transnational themes, there remain several critical steps that can be taken. These

Conversation 5: The future

include facilitating data access among heritage institutions, enhancing interoperability, rethinking legal frameworks within transnational teams, securing funding for cross-professional, cross-disciplinary, and cross-border collaborations, and fostering the sustainability of existing initiatives like the WARCnet and RESAW networks. It is essential to create an environment where institutions can seamlessly share resources and expertise across borders, breaking down barriers to collaboration. Furthermore, legal frameworks must adapt to accommodate the complexities of transnational projects, ensuring that intellectual property rights and other legal considerations are addressed effectively. Moreover, initiatives like WARCnet and RESAW exemplify the potential for sustained collaboration and knowledge exchange across borders. Encouraging the growth of such networks and ensuring their long-term viability is crucial in fostering a culture of transnational cooperation. Financial and institutional support plays a pivotal role in enabling this transnational cooperation.

Márton Németh: Fully agree with Valérie Schafer. I hope that sooner or later a Central-European network can be realised with adequate funding from various countries and stakeholders to run joint projects by the collaboration of historians, social scientists, and IT people to describe the reminiscence of our common past in web archives and make comparative research projects of the reflections of various events, social movements etc.

Moisés Rockembach: I see the future of transnational web archive studies being crafted through collaboration across disciplines and borders, emphasising the importance of developing frameworks that facilitate knowledge and resource sharing across borders. The sustainability of research networks like WARCnet and RESAW is essential, requiring funding and institutional support to foster innovation and adaptability to new technologies and digital dynamics. Additionally, quickly adapting to technological shifts and reassessing collection strategies are crucial to ensuring fair and comprehensive representation of digital cultures. Countries without extensive national web archiving coverage can also significantly benefit from these transnational efforts, not only by accessing shared resources and expertise but also by contributing unique perspectives and critical data to the global understanding of digital culture and heritage.

Saskia Huc-Hepher: In my vision of the future of transnational web archiving studies, we will not only be working in the seamless, technically interoperable and legally universal framework set out above, but we will also be working across sectorial borders, with university researchers and teachers designing and delivering courses in collaboration with their GLAM counterparts, and GLAM professionals drawing on the cultural, linguistic and historical expertise of their university colleagues to inform their collecting and curating practice. And these partnerships will not take place on a national level, but researchers and practitioners

The Routledge Companion to Transnational Web Archive Studies

from educational and memory institutions will be collaborating across north-south and east-west regional spaces, as well as across class divides, to develop courses and collections that are genuinely representative of the world's data in all its cultural, linguistic and social diversity, and to nurture the transnational, translocal even, web archivists of tomorrow.

Susan Aasman:

Transnational web archive studies has not yet addressed the issue of climate change, the challenge of the impact on preserving ever more data in the archive will need to be addressed. Will the solution be in more effective storage capacities, so technological ones, or do we need to rethink collection strategies and aim for international collaboration? There is not an easy way out, but it inevitably will be an urgent point for conversations.

Matthew Weber:

The study of the web, and the internet more broadly, focuses on an object of study that often blurs across boundaries, necessitating transnational research. For instance, examining major global events through an analysis representation on web archives would be limited without a transnational perspective. The future of transnational studies is one of growth, and this edited volume illustrates the many ways in which research in this space can contribute to future societal understanding. Valerie's comments point to the importance of research networks, but this is also part of a grand research challenge of our time. The sharing of resources and knowledge across institutions is a critical part of future growth in this space.

Jane Winters:

I completely agree with Susan that the future of transnational web archiving has to be an environmentally conscious and sustainable one. Collaboration across borders has the potential to address current problems such as duplication of data, infrastructure and tools and to bring web archiving into growing debates about the environmental impact of large-scale digital preservation. Web archive studies have long been characterised by collaboration, partnership and the sharing of good practice, all of which will be essential for shaping a sustainable transnational future.

Janne Nielsen:

I agree about the importance of sustainable international networks for collaboration and about the need for legal frameworks, which allow and facilitate transnational collaborations, as mentioned by several others. Adding to Michael Kurzmeier's important point about the need to address how we deal with AI-generated content in web archives: In the current debate about the fast-increasing amount of AI-generated material on the web, the web archive studies community can offer important insights into historical as well as current understandings of what the 'content' of the web is (in relation to what is important to preserve in web archives), and how this is influenced by different factors (legal, economic, cultural etc.). For instance, machine

Conversation 5: The future

translated content seems to be growing very fast (Thompson et al., 2024), and transnational web archive studies could study such content to help address the consequences of these new technologies and tools for how we perceive and interact with the web and web content — in web archives as well as everyday use of the web.

Reference

Thompson, B., Dhaliwal, M. P., Frisch, P., Domhan, T., & Federico, M. (2024). A Shocking Amount of the Web Is Machine Translated: Insights from Multi-Way Parallelism. <https://arxiv.org/pdf/2401.05749.pdf>.

GLOSSARY

A

Application Programming Interface (API): A set of protocols, tools, and standards for building software applications. APIs facilitate programmatic communication and the exchange of information between distinct applications. In the context of web archiving, APIs allow web-archiving tools and services to programmatically access, retrieve, and integrate archived web content and metadata. This facilitates tasks like bulk data harvesting, thematic collection building, or extracting web archive files and metadata data for research purposes.

Archives Research Compute Hub (ARCH): A research and education platform designed for analysing web archives. It provides researchers with access to advanced computing resources and tools for building, processing, analysing, and preserving archived web datasets efficiently and at scale, including text analysis, machine learning, and data visualisation. <https://archivesunleashed.org/arch/>.

Archive Unleashed Tools (AUT): A suite of software designed for web archive analysis. It enables researchers to explore and analyse large-scale web archives by facilitating tasks such as data extraction, metadata analysis, and visualisations.

ArchiveIT: A subscription web-archiving service by the Internet Archive that allows institutions and individuals to capture and preserve websites and social media content, to create and manage digital collections and lend access to them.

C

Cascading Style Sheets (CSS): A language describing the presentation of a document written in a markup language such as HTML. It is used to manage the layout, formatting, and visual appearance of web pages, by allowing developers to separate the content and presentation.

CDX Internet Archive Index File: A file format for exchanging metadata about web content archived in web archives. A CDX file provides a standardised way to summarise a single web document and thereby enable sharing of information, including URL, timestamp, MIME type, and checksums.

Glossary

Close reading: A mode of analysis method that employs a systematic reading of an individual text. The aim is to come to an interpretation based on a deep focus on words, sentences or formal structures.

Country code Top-Level Domain (ccTLD): The two-letter suffix of internet addresses identifying a country or a territory, such as .dk for Denmark, .jp for Japan or .ru for Russia. The Internet Assigned Numbers Authority (IANA) manages the assignment of ccTLD based on the two-letter ISO 3166-1 alpha-2 country codes, derived from the United Nations list of recognised countries and territories. Many national web archives rely on ccTLDs to define the part of the web they are mandated to preserve.

D

Distant reading: A method for analysing large digital data sets by applying computational methods. It originated in literary studies to indicate a new approach different from the more traditional form of close reading but is now used more broadly in the Digital Humanities.

G

GDELT: A massive database that monitors global news media, extracting information about people, events, themes, and emotions. It tracks millions of news articles daily in multiple languages, offering insights into global events, societal trends, and media coverage. www.gdeltproject.org.

Generic Top-Level Domain (gTLD): A top-level domain that is not associated with a specific country, organisation, or industry, such as .com, org, or net.

H

Heritrix: An open-source web crawler for web archiving developed by the Internet Archive. The tool was further refined and enhanced through the collaborative efforts of the IIPC, making Heritrix a widely adopted standard for systematic and automated web preservation within the web-archiving community.

HyperText Markup Language (HTML): HTML is the standard markup language for creating and structuring web content. HTML uses tags to define the content, layout, and semantics of web documents. These tags guide web browsers in rendering and presenting web pages to users.

I

IIPC Collaborative Collection: A public curated collection of web content contributed by member institutions of the International Internet Preservation Consortium (IIPC). It consists of thematic collections covering the web activity related to diverse topics, themes or events, such as politics, culture, and technology. The IIPC Collaborative Collection is administered by the IIPC Content Development Group (see below).

IIPC Content Development Group: The IIPC Content Development Group pursues collaborative collections by member institutions of the International Internet Preservation Consortium (IIPC). By leveraging expertise and resources from diverse IIPC member institutions, the group contributes to the establishing of collections that do not map to any one member's responsibility or mandate, and that is transnational in scope.

Glossary

The International Internet Preservation Consortium (IIPC): A global organisation dedicated to preserving and providing access to web content for future generations. It brings together libraries, archives, and other institutions to collaborate on web-archiving initiatives, develop best practices, and promote the long-term preservation of online information.

Internet Archive: A non-profit digital library aiming to provide “Universal Access to All Knowledge,” founded in 1996 by Brewster Kahle. It archives and provides free access to a vast array of digital content, including web pages, books, music, and videos. In 2001, the Internet Archive launched its “Wayback Machine” (see below), an interface allowing users to view archived versions of websites.

J

JSON (JavaScript Object Notation): A lightweight, text-based data format used for data interchange in many web applications. It is designed to be easy for humans to read and for machines to parse and generate. JSON structures represent collections of name– value pairs and ordered lists of values.

M

Memento Protocol: A web standard (RFC 7089) that facilitates accessing previous versions of a web page by integrating with the web’s time dimension feature. It allows browsers and other tools to retrieve archived versions of a web page from various archives.

MIME type (Multipurpose Internet Mail Extensions type): A standard that indicates the nature and format of a document or a file on the internet. It helps determine how to process, transmit, or display the data. For example, the MIME type for HTML documents is “text/html”, and “image/jpeg” for JPEG images. Web archives use MIME types to accurately categorise and store digital content, ensuring that archived files can be correctly identified and rendered when accessed.

P

PHP: A programming language widely used for developing websites. It is known for its ability to produce dynamic and interactive web pages. PHP executes code on a web server to generate customised web content before sending it to the viewer’s browser.

R

RESAW (A Research Infrastructure for the Study of Archived Web Materials): A community established in 2012, aiming at promoting a collaborative European research infrastructure for the study of archived web materials. Since 2015 RESAW has organised a biennial series of conferences. <http://resaw.eu>.

S

Scalable reading: An approach that combines close reading with distant reading (see above). The aim is to zoom in and out in large data sets by pairing computational distant viewing methods to more traditional qualitative interpretative methods.

Seed list: A list of known URLs (seeds) used as the starting list for web crawling to systematically discover and collect web pages. From the seed list the web crawler recursively visits links on the discovered web pages to discover new content.

Glossary

SolrWayback: A web archive replay software package built on top of Apache Solr and the Wayback Machine. It offers fast and flexible indexing of and access to archived web content (ARC/WARC files), and the integration with Solr enables powerful full-text search capabilities, making SolrWayback a valuable tool for researchers and archivists. SolrWayback is developed by the Royal Danish Library. <https://github.com/netarchivesuite/solrwayback>.

T

Top-Level Domain (TLD): The highest level of domain names in the internet's Domain Name System (DNS). TLDs are the segments at the end of domain names, such as .com, .org, and .edu. The Internet Corporation for Assigned Names and Numbers (ICANN) manages and coordinates TLDs to ensure stability of the internet's naming system.

TROVE: A collaborative resource developed by the National Library of Australia and hundreds of Partner organisations around Australia which provides access to the archived web materials and other digitised resources.

W

WARC (Web ARChive): A standard file format for storing and exchanging web archives. It allows for the collection and preservation of web content, including HTML pages, images, videos, and other resources, along with associated metadata, enabling efficient storage, retrieval, and analysis of web-based information.

WARCnet: An international research network (2020–2023) aiming at promoting national and transnational research to better understand the history of (trans)national web domains and transnational events on the web, drawing on the increasingly important digital cultural heritage held in national web archive (funded by the Independent Research Fund Denmark | Humanities, grant no 9055-00005B).

Wayback Machine: An interface that enables users to explore historical versions of websites collected by the Internet Archive and how these have changed over time. By entering a URL, users can see a calendar view of the dates when snapshots of that page were taken, and, upon selecting a date, they can view how versions of that page have changed over time. The Wayback Machine also allows users to save URLs to the Internet Archive. To date, the Wayback Machine offers a limited keyword search option for URL discovery. The Wayback Machine was launched in 2001 by the Internet Archive.

Web crawler: A digital tool that automatically navigates the internet to collect data from websites by following links from one web page to another. In the context of web archiving, web crawlers are used to visit and capture snapshots of web pages systematically.