Exploring the Archived Web during a Highly Transformative Age

Sophie Gebeil, Jean-Christophe Peyssard

Exploring the Archived Web during a Highly Transformative Age

PROCEEDINGS OF THE 5TH INTERNATIONAL RESAW CONFERENCE
Marseille, June 2023

edited by
Sophie Gebeil
Jean-Christophe Peyssard

FIRENZE
UNIVERSITY
PRESS

ISSN 2704-601X (PRINT) | ISSN 2704-5846 (ONLINE)

Exploring the Archived Web during a Highly Transformative Age

Proceedings of the 5th international RESAW conference, Marseille, June 2023

edited by
Sophie Gebeil
Jean-Christophe Peyssard

FIRENZE UNIVERSITY PRESS
2024

Exploring the Archived Web during a Highly Transformative Age : proceedings of the 5th international RESAW conference, Marseille, June 2024 / edited by Sophie Gebeil, Jean-Christophe Peyssard. – Firenze : Firenze University Press, 2024.

https://books.fupress.com/isbn/9791221504132

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 979-12-215-0412-5 (Print)
ISBN 979-12-215-0413-2 (PDF)
ISBN 979-12-215-0414-9 (XML)
DOI 10.36253/979-12-215-0413-2

Front cover: View over Marseille from the park Émile Duclaux, 11 February 2018, by Velvet CC BY-SA 4.0.

Published with the support of the TELEMMe laboratory of Aix-Marseille University and the CNRS, the French National Research Agency (Project PICCH ANR-21-CHIP-0003), the Institut Universitaire de France, and the MMSH.

Peer Review Policy
Peer-review is the cornerstone of the scientific evaluation of a book. All FUP's publications undergo a peer-review process by external experts under the responsibility of the Editorial Board and the Scientific Boards of each series (DOI 10.36253/fup_best_practice.3).

Referee List
In order to strengthen the network of researchers supporting FUP's evaluation process, and to recognise the valuable contribution of referees, a Referee List is published and constantly updated on FUP's website (DOI 10.36253/fup_referee_list).

Firenze University Press Editorial Board
M. Garzaniti (Editor-in-Chief), M.E. Alberti, F. Vittorio Arrigoni, E. Castellani, F. Ciampi, D. D'Andrea, A. Dolfi, R. Ferrise, A. Lambertini, R. Lanfredini, D. Lippi, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, A. Orlandi, I. Palchetti, A. Perulli, G. Pratesi, S. Scaramuzzi, I. Stolzi.

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

The online digital edition is published in Open Access on www.fupress.com.

Published by Firenze University Press
Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

This book is printed on acid-free paper
Printed in Italy

Table of contents

Introduction 9
Sophie Gebeil, Jean-Christophe Peyssard, Charles Riondet, Xavier Daumalin, Maryline Crivello, Fabien Borget

SECTION 1. WEB ARCHIVING IN THE MEDITERRANEAN: CHALLENGES FOR DIGITAL HERITAGE FACING THE CRISIS

"Just like home." The Words of online hospitality 25
Dana Diminescu, Quentin Lobbé

Web archiving in Tunisia post-2011: The National Library of Tunisia's experience 37
Raja Ben Slama

SECTION 2. RETHINKING COLLECTION CREATION FOR CULTURAL AND SOCIETAL CHANGE

Bridging the gap: Capturing UK trans health discourse in the Archive of Tomorrow 45
Alice Austin

Making social media archives: Limitations and archiving practices in the development of representative social media collections 57
Beatrice Cannelli

SECTION 3. THE WEB AS HERITAGE IN THE MAKING: DEBATES AND CHALLENGES

Challenges in archiving the personalized web 79
Erwan Le Merrer, Camilla Penzo, Gilles Tredan, Lucas Verney

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, © 2024 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

Mapping the archival horizon: A Comprehensive survey of COVID-19 web collections in European GLAM institutions 95
Nicola Bingham

A Network to develop the use of web archives: Three outcomes of the ResPaDon project 113
Sara Aubry, Audrey Baneyx, Emmanuelle Bermès, Laurence Favier, Alexandre Faye, Marie-Madeleine Géroudet, Benjamin Ooghe-Tabanou

SECTION 4. WEB ARCHIVE AS A MATERIAL FOR UNCOVERING WEB HISTORY

Time, bits, and nickel: Managing digital and analog continuity 129
Julie Momméja

A Decade of transformation discourse: Sociotechnical imaginaries of the Dutch web between 1994–2004 141
Nathalie Fridzema, Susan Aasman, Tom Slootweg, Rik Smit

Flirting and the web: The case study of Luxusbuerg 163
Carmen Noguera

The Online presence of the Danish public sector from 2010 to 2022: Generating an archived web corpus 185
Tanja Svarre, Mette Skov

SECTION 5. MULTI-LEVEL METHODS FOR STUDYING WEB ARCHIVES

Web archives and hyperlink analyses: The case of videnskab.dk 2009–2022 201
Niels Brügger, Katharina Sølling Dahlman

Do user comments belong to journalistic articles? A brief visual history of user interaction on selected German and American news websites 1996–2024 223
Johannes Paßmann, Martina Schories, Paul Heinicker

Multi-level structure of the First Tuesday communities after the 2000 dot-com crash: A social network analysis of economic actors based on web archives 247
Quentin Lobbé

Semantic analysis of web archive historical data: 1983 “Marche pour l’égalité et contre le racisme” 259
Davide Rendina, Sophie Gebeil, Mathieu Génois, Patrice Bellot

SECTION 6. BODY AND HEALTH STUDIES IN A DIGITAL CONTEXT

Food, cooking, and health in a selected corpus of websites and

connected YouTube channels in France. Collecting and archiving the audiovisual web 277
Christian Bonah, Solène Lellinger, Caroline Sala

We're all experts now? Archiving public health discourse in the UK Web Archive 295
Alice Austin, Leontien Talboom

SECTION 7. STUDYING MEDIATIZED MEMORIES

Websites as historical sources? The benefits and limitations of using the websites of former repatriates for the history of schooling in colonial Algeria 311
Christine Mussard

A Social media archive for digital memory research 321
Costis Dallas, Ingrida Kelpšienė

CONCLUSION

A Highly transformative age for web archives 343
Nicola Bingham, Valérie Schafer, Jane Winters, Anat Ben-David

Introduction

Sophie Gebeil, Jean-Christophe Peyssard, Charles Riondet, Xavier Daumalin, Maryline Crivello, Fabien Borget

The recent transnational political and health crises serve as a stark reminder of the need to closely scrutinize recurring information disseminated across web and social media platforms. These digital documents, both sublime and sinister, have become ubiquitous in human activity, encompassing everything from ironic memes about lockdown measures to tragic videos documenting violent actions against civilians by armed forces. The nature of web archiving systems is geared towards preserving these born-digital sources, which are crucial for studying recent phenomena. Consequently, they constitute invaluable materials that not only elucidate the history of the current context since the 1990s but also contribute to the construction of a digital heritage—a memory for future societies and an indispensable source of contemporary facts for researchers.

Nevertheless, the short timeframe during which transnational events (pandemics, political crises, etc.) have unfolded should not overshadow the profound societal changes that have characterized the period since the democratization of the web in the 1990s. This includes grappling with the

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)
Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, © 2024 Author(s), CC BY 4.0, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

climate challenge, asserting individual fundamental rights, navigating the effects of economic liberalism, and observing shifts in religious practices. These transformations have been accompanied by far-reaching socio-technical developments and an expansion in digital cultures. The global history of these developments is a narrative yet to be fully articulated. The first web archiving initiatives, originating in the 1990s, not only provided a valuable source base but also spurred a large international following, comprised of organizations and individuals such as GLAM (Galleries, Libraries, Archives, Museums), scholars, and civil society. The aim was to collect, safeguard, and make the past web accessible to all. This archived web forms a unique resource for the study of recent phenomena, offering the means to preserve the corpora and retrieve old websites that are currently unavailable online. However, using web archives as research material demands epistemological and methodological reflection, and consequently, has emerged as a distinct field of study.

In recent years, the use of the web to investigate social, cultural, and political phenomena has experienced significant growth. This has contributed to advancements in both the history of the web and research in the social sciences and humanities. Moreover, the web and digital cultures have themselves become subjects of study, sometimes by trial and error, through the exploration of methodologies, whether they be grassroots approaches or projects rooted in the digital humanities.

Since 2012, RESAW[1], an organization established to develop a pan-European research infrastructure, has played a significant role in this dynamic focused on the study of web archives and the past web. This underscores how a deeper understanding of the web's history can assist us in facing the challenges of a world in constant flux. RESAW has also created the opportunity for epistemological and methodological reflection on the creation of web archives. It has thereby solidified its position as a major player in web archives studies—an interdisciplinary research field undergoing radical restructuring.

Eight years after the inaugural RESAW conference, which sparked groundbreaking debates on technical, scientific, and archival aspects, the conference in Marseille, organized by Aix-Marseille University, aims to assess web archives studies in relation to research conducted on the internet, social media, web archives, and reborn digital heritage. Based on the selected papers from the conference, this book examines the development of web archiving, highlighting the ways in which technical, cultural, geopolitical, social, and environmental changes are affecting the conception,

[1] Research Infrastructure for the Study of Archived Web Materials, https://web.archive.org/web/20220824102113/http://resaw.eu/

study, and dissemination of this reborn digital heritage[2][3].

The first part of the introduction describes the scientific and cultural context that fosters the development of web archiving practices by giving voice to the actors from the TELEMMe laboratory, Aix-Marseille University, and the Mucem. These partners play a significant role in advancing this research field both nationally and internationally. The second part outlines the structure of the book and the primary questions addressed by its various contributions.

1. Understanding the global challenges of web archiving from the shores of the Mediterranean

The core reflections at the RESAW 2023 conference stem from a research dynamic initiated in the 2000s within the TELEMMe laboratory and the Maison méditerranéenne des sciences de l'homme (MMSH). This initiative has expanded across Aix-Marseille University through collaborations with national partners (Mucem, BnF, INA) and international partners (RESAW).

Presenting this scientific ecosystem will help the reader grasp the unique approaches to web archive studies highlighted during the Marseille conference in June 2023. The TELEMMe laboratory—Time, Spaces, Languages, Southern Europe, Mediterranean[4]—under the leadership of Sophie Gebeil, played a crucial role in organizing the conference. Gebeil, who directs the transversal workshop on Visual Studies and Digital Humanities in the Mediterranean, explored the web as a source for cultural history in her 2015 thesis supervised by Maryline Crivello. This early exploration of web archives was part of a broader reflection on digital practices initiated by the the Image Sounds and Digital Practices pole at the MMSH in the early 2000s, in connection with the MMSH Sound and Media library[5]. Gebeil's work, in collaboration with the INA and the BnF, aligned with the RESAW collective founded by Niels Brügger in 2013. The idea of

---

[2] The book was coordinated by Sophie Gebeil and Jean-Christophe Peyssard, with assistance from Mara Bertelsen (coordination, translation, and editing). These proceedings also result from a collective work since 2012 within the framework of the Scientific Committee of the RESAW network, which includes Niels Brügger, Jane Winters, Valérie Schafer, Susan Aasman, Anne Helmond, and Sebastian Giessmann. The chapters are based on a selection of the best proposals from the RESAW 2023 international conference https://resaw2023.sciencesconf.org archived version: https://web.archive.org/web/20240112135604/https://resaw2023.sciencesconf.org/

[3] All of the figures in the book can be viewed online here: Gebeil, S., & Peyssard, J.-C. (2024). Figures. *Exploring the Archived Web During a Highly Transformative Age. Proceedings.* edited by. S.Gebeil & J.-C. Peyssard. Exploring the Archived Web During a Highly Transformative Age (RESAW 2023), Marseille. Zenodo. https://doi.org/10.5281/zenodo.11263165

[4] UMR 7303 TELEMMe (AMU-CNRS), under the direction of Xavier Daumalin from 2020 to 2024.

[5] Image Sounds and Digital Practices in SSH pole, https://imageson.hypotheses.org, version 2/12/2010, https://web.archive.org/web/20101202094753/http://imageson.hypotheses.org/

organizing the fifth conference emerged in 2019, garnering the interest of Jean-Christophe Peyssard, who became the head of the MMSH Media Library in June 2020. With a background in Arabic studies and five years of experience in the Middle East, Peyssard specialized in collecting digital traces from Arab and Muslim societies, contributing notably to the *Lexique vivant de la révolution et de la guerre en Syrie*[6] project and co-organizing an international conference on digital archives in Lebanon[7]. The conference also aimed to develop web archiving practices and raise awareness within the Aix-Marseille academic communities, leading to the creation of the WebLab at the MMSH Media Library. The WebLab is hub for reflection and practice in studying archived web content and new media[8]. It will facilitate the development and sharing of knowledge on web archiving, including best practices for collection and preservation tools and methodologies, exploratory approaches to corpus analysis, ethical and legal considerations, and the use of the archived web in academia among students, researchers, and support staff in the social sciences and humanities.

The conference is thus the result of these multiple encounters, from Aarhus to Beirut via Paris, but would not have been possible without the decisive support of the TELEMMe laboratory, Aix-Marseille University, and the Mucem, all interested in the interdisciplinary and innovative nature of web archive studies. We wanted to give a voice to the individuals and institutions that contributed to the collective reflection and successful organization of the conference, culminating in this publication.

1.1 Prof. Xavier Daumalin, Directory of the TELEMMe research laboratory

The TELEMMe laboratory is particularly pleased to contribute to the opening of this conference, especially in partnership with the Mucem, a museum celebrating its 10th anniversary and long-time collaborator in the laboratory's activities. This collaboration goes beyond hosting to a co-constructive approach, as seen in the foresight day on urban geography organized with the CNRS's National Institute of Humanities and Social Sciences in collaboration with the MMSH. Another example is the day focused on creating colonial imaginaries in Naples and Marseille, based on the MARS IMPERIUM project. This project, fully integrated with the Mucem, includes digital deliverables such as a website, web documentary with nearly 90 videos exploring as many imperial themes, digital showcases

[6] https://web.archive.org/web/20231014024837/https://syria-lexicon.pubpub.org/
[7] New Digital Archives in the Near East, https://nanpo.sciencesconf.org/
[8] WebLab, Bibliothèques et Archives à la Maison méditerranéenne des sciences de l'homme, https://pba.mmsh.fr/?page_id=1465

analyzing collection and exhibition methods, and the journey of objects that belonged to the *Institut colonial de Marseille*, a critical film on the 1922 colonial exhibition, digital urban walks, and a documentary platform gathering all the archives used in the project for public access, which requires rigorous referencing and rights management[9].

The conference theme, "Exploring Archived Web in a Highly Transformative Age", reflects the innovative capacity of the TELEMMe laboratory. From the initial avenues opened by Maryline Crivello to the creation of a transversal axis on Visual Studies and Digital Humanities in the Mediterranean—and soon an advisory cell for Open Science—and including the first French history thesis on web archives defended by Sophie Gebeil in 2016, this research field has flourished and gained prominence within the unit. TELEMMe has been reflecting deeply on how these new tools transform our professions, both methodologically and in terms of research valorization[10].

The issue of archiving digital information disseminated on the web has become crucial, establishing itself as a key observatory for contemporary knowledge creation and transmission. It raises numerous issues: ethical, legal, heritage, technical, environmental, political, accessibility, training, public use, and the sustainability of digital creations. For researchers engaged in producing digital deliverables disseminated via the web, it is essential to integrate these considerations from the project's inception and budget for them. The creation, dissemination, and sustainability of digital objects have costs that extend beyond project durations, engaging laboratory finances for decades. Ensuring the longevity of deliverables in a fragile and evolving technical and institutional environment is challenging. Additionally, rights for disseminating documents must be renewed every five, ten, or fifteen years, complicating financial planning in a context of short-term resource visibility and increasing commodification of documentary resources. The field of web archiving thus contributes to a broader reflection on the sustainability of digital documents produced by research laboratories in the humanities and social sciences.

1.2 Charles Riondet, Mucem

Hosting the RESAW conference provided a superb opportunity for the Mucem to delve into the wealth of research on web archives and explore a

[9] MARSIMPERIUM, Imperial Marseille: History and (Post)Colonial Memories, 19th–20th Centuries, TELEMMe, https://marsimperium.hypotheses.org

[10] The laboratory thanks the organizers—Sophie Gebeil, Maryline Crivello, and Jean-Christophe Peyssard—as well as the MUCEM, and the entire TELEMMe research support team—Maja, Agnès, Delphine, Caroline, and Mireille—who have been working for several months to ensure the success of this event.

less-charted segment of its history, opening new avenues for collaboration.

As successor of the *Musée national des arts et traditions populaires* (MNATP), founded in 1937 and closed in 2005, the Mucem houses ethnographic collections compiled since the late 19th century and enriched through field surveys and collections known as *enquêtes-collectes*. This approach allows for the preservation of the context in which objects are produced and used, enriching the collections through extensive documentation such as interviews, sound and video recordings. From the 1960s to 2005, the MNATP maintained close connections with the world of research and was envisioned by its creators and managers as a museum-laboratory. Today, the principle of e*nquêtes-collectes* is still central to the Mucem's acquisition process, albeit with a new focus on the Mediterranean. The Mucem's collections comprise:

- Museum collections, with acquisitions validated by a Ministry-approved committee
- Public and private archives
- Library collections, serving as documentation for the museum and former laboratory, as well as a heritage library and a public reading library

The Mucem has a long history of information and digital technology, dating back to the MNATP's pioneering efforts in digital museographic management. The impetus provided by Georges-Henri Rivière and later Jean Cuisenier led to the development of exacting standards for collection management, subsequently incorporating the use of information technology to document collections and manage vocabularies, with initial efforts commencing in the late 1960s. The advent of the World Wide Web in the 1990s did not go unnoticed, leading to the creation mid-decade of the MNATP's first websites, both institutional and corpus oriented.

Between 2005 and 2012, the closure of the MNATP and the establishment of the Mucem marked a turning point for documentation, mediation, and digital promotion policy. Alongside extensive efforts to review inventories, complete records, conduct conservation operations, and acquire photographs, the Ministry of Culture facilitated the online publication of certain MNATP databases, notably those pertaining to postcards or photograph collections.

However, due to the limitations of physically showcasing the collections at that time, digital technology became the primary means of public access to the collections. Over ten thematic websites presented corpora linked to the new scientific project (centered on the Mediterranean and European civilizations), highlighting field surveys related to the Mediterranean area (such as the olive tree, glass craft, industry and trade, Armenian diaspora, etc.), as well as more 'traditional' MNATP themes adapted to the Mucem

program, such as ethnomusicology with the website dedicated to bagpipes of Europe and the Mediterranean.

The inauguration of the Mucem in 2013 posed a significant challenge for the online presentation of collections. A federated search engine was designed to amalgamate the various document types mentioned above, providing a unified search experience and aligning vocabularies across museum, archive, and library collections.

With the commemoration of the Mucem's ten-year anniversary in 2023, there is renewed contemplation on how to develop its digital productions and online access, aligning with contemporary constraints and opportunities such as linked data, artificial intelligence, and digital sobriety. Within the array of new challenges to be addressed, some are specifically related to web archives: As a civilization museum, is it the museum's role and responsibility to collect, document, and preserve digital practices? How should the museum handle the digital output created by itself and its predecessors, particularly the oldest, and what role does it play in safeguarding, curating, and promoting it?

Addressing these questions, raised during the opening of the RESAW conference at the MucemLab in Marseille in June 2023, has led to the initiation of a collaborative research project with Sophie Gebeil (AMU) and Alexandre Faye (web archive curator at the BnF). The project aims to preserve, analyze, and present archived thematic websites from the MNATP and the Mucem, contributing to web archive research and serving as a reciprocal benefit to the field.

1.3 Prof. Maryline Crivello, Vice President of the Board of Directors, in charge of developing the interdisciplinarity mission, Aix-Marseille University

Hosting this international conference on the archived web is a tremendous honor, marking the first of its kind in France, made possible through the collaboration of long-standing partners Aix-Marseille University (AMU) and the Mucem.

AMU, the largest Francophone multidisciplinary university, boasts a community of 80,000 students and nearly 8,000 staff members across multiple campuses, notably in Marseille and Aix-en-Provence. Recognized as a 'research-intensive university', AMU hosts 122 research structures affiliated with major national research organizations. Embracing interdisciplinarity, AMU established a dedicated Interdisciplinarity Mission in late 2020. Unique in its positioning and ambitions within the French university landscape, its purpose is to facilitate an integrated, crosscutting approach to support interdisciplinary projects and communities.

Naturally, the AMU Mission has lent its support to this conference, developed with the aim of fostering vital dialogue across disciplines. The program exemplifies this collaboration, uniting fields such as communication sciences, sociology, and history, as well as computer, archival, and library science. These fields and professions converge to advance tools and methodologies crucial for understanding the web, following the lead established by its pioneer, Niels Brügger. Moreover, this conference holds particular significance to us as it resonates with the scientific heritage cultivated by AMU researchers in the humanities and social sciences since the 1980s. Initial research focused on television media, their archival methods by the INA, and the importance of preservation within Legal Deposit frameworks. Today's exploration of archived web materials aligns with the questions raised by earlier studies on these communication platforms. A notable milestone was the colloquium organized by MMSH researchers in 2005, titled “Screen Writings: History, Practices, and Spaces on the Web”, which anticipated the emergence of web archiving as an internationally recognized field of study.

Taking place symbolically on the shores of the Mediterranean, this conference evokes the spirit of the 2011 Tunisian Revolution, which utilized social networks as crucial mobilization tools. Because, as we know well, websites and the web serve as digital conduits for human experiences and sensitivities, fundamentally shaping our perception of the world.

Thank you to the organizers for seizing the opportunity to convene us around these invaluable reflections. And special acknowledgement is due to Sophie Gebeil, whose dedication over the years has been instrumental in making this project a reality.

1.4 Fabien Borget, Open Science officer, Aix-Marseille University

Aix-Marseille University is hosting this year’s RESAW conference, which will foster rich exchanges on web archiving, its questions, tools, and methods.

Web archiving involves capturing a snapshot of the web at a specific moment and preserving it. However, for these archives to be easily findable, accessible, interoperable, and reusable—in other words, FAIR—all the methods used must facilitate these qualities.

Making research data FAIR is a cornerstone of Open Science, a movement initiated a number of years ago and now widely adopted in the institutional policies at national[11] and international[12] levels. Open access

[11] 2nd French Plan for Open Science, available on https://www.ouvrirlascience.fr/deuxieme-plan-national-pour-la-science-ouverte/, viewed 16 April 2024

[12] UNESCO Recommendation on Open Science available on

allows research resources to be freely accessible online, promoting the unrestricted flow of knowledge and information. The link between web archiving and Open Science is thus evident.
Web archiving also presents an opportunity to diversify academic dissemination models, which are currently undergoing significant changes. When editorialized, we must ensure bibliodiversity, crucial for maintaining the scientific integrity of scholarly output.
Ensuring the sustainability of the infrastructures hosting this diversity is also a challenge, necessitating policies that account for this aspect.
Open Science policies currently revolve around four main points:

1. Generalizing open access.
2. Structuring, sharing, and opening research data. To some extent, web archiving falls under this point: how can archived web content be considered 'research data'? What does FAIR (Findable, Accessible, Interoperable, and Reusable) mean for web archives?
3. Opening and promoting source codes, research software, and algorithms.
4. Transforming research practices to make Open Science a default principle.

One of the key challenges of the RESAW conference is to share web archiving practices to ensure that web archives can indeed be considered 'research data'. Another challenge is to recognize that web archiving checks all the boxes of open science while retaining its unique nature as data that will (in the future) become research data.

One of our ambitions at Aix-Marseille University is to make research data FAIR[13]. To achieve this, we offer support to colleagues producing data through the Data Desk of Aix-Marseille (GDsAM), a network of experts in research data present on-site. This initiative, qualified by the ministry as a Data Workshop, involves key university actors such as our Common Documentation Service (SCD) and the Center for Research Data Support and Training (CEDRE)[14]. We provide specialized training on research data management and support operational implementation in data storage, visualization, analysis, and more, leveraging local and national infrastructures such as our Data Center, our mesocenter, and HumaNum.

Recognizing its strengths, our university is organizing itself around research data, which includes archived web data. This structuring involves synergizing existing strengths and competencies covering all aspects of the research data lifecycle.

In the research data lifecycle, it is legitimate to ask where web archiving fits in. These digital traces are intended to be the digital footprints for future

---

https://unesdoc.unesco.org/ark:/48223/pf0000379949, viewed 16 April 2024

[13] Aix-Marseille University Open Science charter, https://www.univ-amu.fr/system/files/2021-09/Politique_Science_Ouverte_AMU.pdf, viewed 16 April 2024

[14] Part of the IDeAL (Integration and Development at Aix-Marseille through Learning) project, France 2030

researchers and are thus fully-fledged research data. Legitimate questions about their FAIRness arise; but beyond the very concept of ‘data’, the issues surrounding digital tools, which are growing day by day but remain extremely fragile, are also becoming clearer. How can we ensure the conservation, preservation, and quality of these digital traces over time?

The expertise of participants in the RESAW conference is invaluable in stabilizing knowledge around these questions. We probably also need to develop new tools, methodologies and, potentially, technologies, to consider web archiving as research data. These questions will undoubtedly be central to the discussions at the RESAW meeting. These digital traces will serve as evidence for understanding the societal impact of the digital revolution we are experiencing. Ultimately, isn't transparency in research data also a means to rebuild society's trust in science?

Therefore, this book is the result of a collective effort driven by researchers from various disciplines, librarians, and archivists who have contributed to the development of practices and studies related to web archiving.

2. Exploring the archived web in a highly transformative age

Through 20 chapters written by a collective of authors from diverse backgrounds—students, researchers, archivists, and librarians—supported by a wide variety of disciplines ranging from history to media studies and computer science, the book is organized into seven sections that address the main challenges associated with web archiving amidst a significant period of transformation, as underscored during the June 2023 conference in Marseille.

The first part of the book proposes examining the Mediterranean region as a crucial laboratory for studying the challenges confronting web archiving within a context of crisis, particularly migration and political upheaval. It serves as a platform for stimulating discussions based on approaches employed across different cultural spheres and at various levels to reflect on web archiving within the Mediterranean and its environs. The book opens with the testimony of Professor Raja Ben Slama, former director-general of the National Library of Tunisia (BnT) from 2015 to 2023. A university professor, psychoanalyst, and a key intellectual figure of the Jasmine Revolution of 2011, she shows how the BnT collaborated with public institutions, associations, and volunteer researchers to gather and archive documents dispersed on the web and mobile phones of citizens involved in the popular uprisings between December 17, 2010, and January 14, 2011. From the northern shore, this contribution is juxtaposed with a chapter by Dana Diminescu, addressing the issue of migration. Indeed, the

Mediterranean served as both a transit route and a graveyard for migrants desperately attempting to reach European shores. A pioneer in the field of digital diasporas through the e-diasporas project, which employed web archiving and data analysis tools like Gephi as early as 2004, Dana Diminescu examines digital traces that reveal notions of hospitality and probes the role of emotions in our research based on born-digital sources.

The second part of the book presents two initiatives aimed at tackling the challenge of inclusivity in web archiving. Alice Austin highlights the obstacles encountered by the team of the project "The Archive of Tomorrow" (AoT), particularly in capturing UK trans health discourse, while Beatrice Cannelli offers a brief overview of representation issues in the social media archiving landscape based on the geographical distribution of social media archiving initiatives, highlighting under-represented areas and emerging trends in content collection from social platforms.

The third part sheds light on the creation of collections that absolutely must be documented. Collection organizations are faced with new challenges posed by web developments, as illustrated by Camilla Penzo, Gilles Tredan, Lucas Verney, and Erwan Le Merrer through the case of the personalized web. The following contributions highlight the dedication of archiving institutions to recent dynamics. Nicola Bingham delineates the varied landscape of Covid-19 web collections in European GLAM (Galleries, Libraries, Archives, Museums) institutions. Comprising Emmanuelle Bermes, Laurence Favier, Audrey Baneyx, Benjamin Ooghe Tabanou, Sara Aubry, Alexandre Faye, and Marie-Madeleine Géroudet, the ResPaDon project (Network of Partners for the analysis and exploration of digital data) shows how the BnF's efforts to foster usage among students, primarily from Paris and Lille.

The fourth part offers several case studies that demonstrate the importance of web archives as sources for studying the history of the internet and digital cultures. Julie Mommeja delves into the origins of archiving history in the 1990s, comparing projects and imaginaries carried by the Internet Archive and the Long Now Foundation. The question of representation is also addressed by Nathalie Fridzema, Susan Aasman, Tom Slootweg, and Rik Smit through the study of the Dutch web and the importance of analyzing socio-technical narratives to understand more recent developments. Tanja Svarre and Mette Skov focus on the development of the Danish government's web services in the late 1990s and shows the role played by websites' labels. Beyond representations and service development, the web is also the site of new sociabilities and cultural practices, as recalled by Carmen Noguera's chapter illustrating how the hybrid model of the Luxusbuerg site, combining online chat with offline events, influenced the evolution of flirting.

The fifth part delves into the methods and practices used by researchers

exploring archived web data, offering insights into future perspectives. Research methodologies utilize archived web data in various dimensions, including URLs, visual organization of pages, and text. Niels Brügger and Katherina Sølling Dahlman investigate how the holdings of a national web archive can be used to shed light on the hyperlinks related to one individual website. They complement Quentin Lobbé's application of actor-network theory through a corpus on the history of the First Tuesday community. Johannes Paßmann, Paul Heinicker, and Martina Schories discuss the advances of the "Technograph", a tool the authors are developing for visualizing and analyzing web archive data, applied in an online press study. These approaches emphasize interdisciplinary, as illustrated by Davide Rendina, Mathieu Génois, Patrice Bellot, and Sophie Gebeil's chapter on analyzing text from web pages archived by the INA (Institut National de l'Audiovisuel) on the March for Equality and against Racism, blending computer science and cultural history.

The final two parts address more specific themes emerging as significant issues in studies on web archives within a context of globalized mutations. Part 6 focuses on questions of representation of the body and health within born-digital heritage. Christian Bonah, Solène Lellinger, and Caroline Sala explore how audiovisuals shape eating habits and their connection to individual health concerns and healthy eating. For this, they propose a study on food, cooking, and health in a selected corpus of French websites and connected YouTube channels, in relation to the BnF's web archives. Meanwhile, Alice Austin and Leontien Talboom focus on a singular British initiative among the many collections dedicated to the Covid-19 crisis, aimed at increasing representation and diversity within the UK Web Archive. The pandemic has heightened awareness of the digital nature of traces produced by contemporary societies, necessitating an inquiry into the processes of preserving this digital heritage. Memory concerns are not new in digital studies, as illustrated by the vitality of memory studies that focus on multiple online past narratives and to which Part 7 of the book will be dedicated. Memory discourses are often a clear concern for historians of the contemporary period. Christine Mussard, a historian of the Algerian War, illustrates how she found websites of children educated in Algeria during the colonial period and transformed them into a genuine object of questioning about the relationship to the past, transcending their role as mere access points to contacts serving historical inquiry. Exploring a different cultural and memorial boundary, Costis Dallas, Ingrida Kelpšienė plunge us into Lithuanian social media, revealing how memory practices on Lithuanian social network sites mediated by contested heritage shape cultural identities at transnational, national, and intersectional levels, while building on digital curation approaches to archive Lithuanian social digital memory and save it from future obsolescence.

This dynamic panorama shows the essential role of studies on and using web archiving in understanding contemporary phenomena and addressing future challenges. With its significant impact extending beyond researchers and GLAMs, web archiving also has a significant impact on civil society as a whole, particularly through amateur archiving initiatives in countries lacking established collection systems. The conclusion offers a forward-looking vision by four major figures in the transformation of the web archives landscape 30 years after the first initiatives. Exploring the extent to which web archives participate in and are influenced by this highly transformative age, Nicola Bingham, Valérie Schafer, Jane Winters, and Anat Ben-David propose future challenges for the web archiving community to address.

SECTION 1
Web archiving in the Mediterranean: Challenges for digital heritage facing the crisis

“Just like home.” The Words of online hospitality

Dana Diminescu, Quentin Lobbé

Abstract: In 2015, the Singa association created Calm (Comme à la maison “Just like home”), an internet platform for connecting refugees who are looking for housing with private individuals. The analysis of its archives and different versions provides information on both the expressions of online hospitality and the role of digital tools in facilitating hosting and accommodation. While this innovative interactive directory is based on state-of-the-art tools, it also challenges their limitations, namely the algorithmic temptation that its implementation may reflect, meaning the attempt to automate what is contained within the fluctuating realm of human relations.

Keywords: refugee crisis, online hospitality, techno-solutionism.

Summer 2015. At that time, just like today with the war in Ukraine, the media in every country as well as social media were full of reports on suffering refugees on the borders of Europe[1]. Years ago, we saw images of Syrians, and today there are ones of Ukrainians charging their smartphones on the road to exile or using Facebook, WhatsApp, and Google Maps on their journey, which have circled the world and led to an unprecedented collective awareness of the importance of information and communication technologies (ICT) in the lives of migrants. Since then, a type of technophoria has swept the world of migrants[2]. Making digital technology a lever for integrating these vulnerable populations has become the number one challenge of many initiatives carried out by associations, companies in the social and solidarity economy, universities, public institutions, technology giants, or simply by anonymous individuals committed to the humanitarian cause.

Seen as a wide range of opportunities—in conjunction with the diversification of online uses and access to a stable, personal, and free connection—technological innovation intended to help refugees is occurring alongside the increasing dematerialization of public services for immigrants[3] while also going hand in hand with social innovation. On one

[1] This chapter is a translated and revised version of the article originally published here in French: « Comme à la maison ». Les mots de l’hospitalité en ligne. *Hommes & migrations*, 1337, 2022, https://doi.org/10.4000/hommesmigrations.13962

[2] We have identified around one hundred applications intended for refugees, see Table 1.

[3] For example, the ANEF portal: https://web.archive.org/web/20240427021017/https://administration-etrangers-en-france.interieur.gouv.fr/particuliers/#/ and also the excellent platform: https://web.archive.org/web/20240502093303/https://refugies.info/fr

Dana Diminescu, Télécom Paris, France, dana.diminescu@telecom-paris.fr, 0009-0002-7554-0266
Quentin Lobbé, CNRS, France, quentin.lobbe@gmail.com, 0000-0003-2691-5615

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

hand, it has its origins in the collaborative practices of hackers, but also in the resurgence of the ideal of community within entrepreneurial capitalism linked to digital technology and the online reappraisal of emotion.

In her remarkable book *Le sens de l'hospitalité* (The Meaning of Hospitality), Anne Gotman (2001, 300) shows that the war in Kosovo has already shown television viewers crowds of defeated faces and sparked an unprecedented surge of personal solidarity at the end of the 1990s. This "immediate," "human," "natural," and "self-evident" response is an emotional effect caused, yesterday as today, by the news, the effectiveness of images on TV, and the effect of social media. This has led to an unprecedented level of mobilization of "affective computing"[4] (Chavalarias 2022).

Addressing the needs of refugees with a geek approach, donating their digital skills to serve their cause, proposing coding (not just help, support, clothing, accommodation, or food) to help migrants are all emerging forms of solidarity that we examine in this article.

Should Data4Goods, FabLabs, Techfugees, and other refugee hackathons be understood as new ways of connecting with others or are they building/configuring a new way of organizing hospitality and donation? How effective is it? Is it a form of technological utopia or social logic? A utilitarian approach or a humanitarian action? Or is it a technology company trying to take hold of the shock of the migration crisis, as researcher Evgeny Morozov (2014) describes when he talks about "technological solutionism"[5]?

In this article, we will look at the case of the Calm 'Just like home' platform[6] now renamed 'J'accueille' or 'I welcome'[7]. It proposes connecting "refugees seeking temporary accommodation with citizens who have a room to host them", i.e. an immersion, lasting 3 to 12 months, set up and monitored by Singa, the association behind the Calm platform and which runs the back-office software. This involves analyzing a new form of personal hospitality, between private individuals, mediated by a digital system, and supplemented, or even corrected, by a human mediation protocol (employees and volunteers at the Singa association) that integrates a training program and face-to-face meetings between refugees and French people wishing to host refugees at their homes.

---

4 "Affective computing" is generally used either to manipulate the emotions of users in individual feedback loops or to lead them through methodological individualism to crowdsourcing practices by coordinating the action of digital crowds.

5 Technological solutionism consists of imposing so-called 'intelligent' industrial systems on populations in order to solve social and political problems. These industrial mechanisms are automated and often employed to benefit a data economy.

6 https://web.archive.org/web/20180209063317/http://calm.singa.fr/

7 https://web.archive.org/web/20240502093417/https://www.jaccueille.fr/

1. Research and methodology

We conducted our research in two stages. First, we carried out a detailed study between 2015 and 2017 of the content and operation of the Calm platform. We have created a collection of 11,892 host forms who have registered online for the program. This large, valuable collection provides information about the age or geographical location of the participants as well as the space offered at the host's home, including their 'motivations', and also contains an initial self-introduction (the 'Tell us about yourself' field). Then, we conducted a new series of interviews in 2022 to understand the association's development since its launch. We were able to explore the Singa association's digital archive again over the entire 2015–2022 period.

Our hybrid methodology is at the intersection of sociology and IT. We used a computational system for extracting, cleaning, and analyzing the Calm pool of information. Our results and reflections presented in this article are based on various ad hoc visualizations from the questions that emerged from the exploration of this data.

Our analysis is supported and enhanced by a series of interviews (conducted to orient the questions of the quantitative study beforehand and to validate its findings through feedback from the field) with hosted refugees, families that have housed them, and members of the Singa association.

2. Refugee hospitality: The algorithmic temptation

Since 2015, the Singa association has offered to connect French families who want to host a person who has submitted an asylum application through its Calm program, now called J'accueille!. The hospitality proposed by Singa is not trying to respond to an emergency situation but to register hosts over the long term to facilitate the integration of refugees (with the average hosting duration being 8 months) through an immersion process. Singa is not the only association to have combined digital technology with refugee hosting.

Today, almost all refugee associations—as well as state institutions in charge of foreigners —have "platformed" (Casilli and Posada 2019, 293–30), meaning organized all procedures via a platform.

Among these first initiatives are those undertaken by members of the Jesuit Refugee Service (JRS) who use an email contact and registration system for host families. In Belgium, members of the Bxlrefugees citizens platform have organized their efforts in an autonomous and decentralized way through Facebook surveys. They then meet every evening in Maximilien Park in Brussels to set up particularly efficient, emergency accommodation for refugees on a daily basis.

The Singa association connects hosted refugees with private individual

hosts by having them fill out an online form. Between 2015 and 2017, registering on the platform was done in the following manner: when logging in, you first needed to indicate whether you are registering as a potential host or as someone in need of accommodation, and the system then directed you to a dedicated form. Both the hosts and those seeking accommodation registered individually, even if, in the case of the hosts, it was more generally understood to refer to as a single household (a family, a couple, or even a single person). These forms are based on the free Google Forms technology and are presented as a series of questions calling for open (free-form) or closed (multiple choice or listed) answers.

In its design, Calm's original intention was to build a kind of refugee-specific Airbnb, meaning to match a person who would like to host a foreigner with refugees looking for housing without using any other intermediary. In fact, the form for those seeking housing (refugees) was actually never set up, except for a few days at the beginning of fall 2015. During this period of time, the service was inundated with several hundred entries and broke down soon after. The members of the various associations who were assisting the refugees in the field completed the form on Calm for them. Thus, because the platform could not manage the demand, the form became inaccessible almost immediately, and the Calm team then proposed that those seeking housing should go to the association offices so they could meet them directly.

As for the host form, hosts could still access it and was regularly updated by the association. We count—through the platform archive—4 successive versions of this form: 1) before June 2015 2) from June 2015 to November 2016 3) from November 2016 to June 2017 4) after June 2017. Outside the archive, the platform continued to evolve after 2017 (in particular, its name changed) up to the present version: https://www.jaccueille.fr/. First, we can see that the Calm form is structured to provide continuity. New questions do not appear per se, but open questions generally tend to fragment and become more specialized ('tell us about yourself', 'type of room', etc.), leading into a set of narrower fields ('What languages do you speak?', 'How many rooms can you provide?').

Throughout the changes to the form, the descriptors with which the hosts are qualified in the database are transformed and streamlined. From an IT point of view, it is thus easier to manage a clearly defined and quantifiable parameter (age, number of rooms, list of languages) on a large scale than a field for free expression from which potentially poorly formulated or even unusable information has to be extracted.

What we observed later was in fact, a change in engineering implemented to improve the management and use of the registration data. Therefore, this discretization of an individual into a set of standardized

descriptors is based on a form of "algor[8]ithmic governmentality" as defined by Antoinette Rouvroy and Thomas Berne (2013, 163–196). In view of this, we asked ourselves the following question: why did the Singa association want to streamline the way in which it manages its registration data to this extent? Why model the profiles of host families in this way?

3. Developing digital ways of hosting

Born out of the *Réfugiés Connectés (Connected Refugees) hackathon*, the Calm program was initially managed by people who do not consider themselves IT developers. By choosing an accessible and simple technological infrastructure (Google Forms and Excel spreadsheets), which serves as the basis for manually linking hosting/hosted pairs, the association first promoted human, mediation, and qualitative work based on the experience acquired by its team. Members of the association manually searched their database for good hosting candidates. Each mediation was the subject of a unique discussion between families and refugees.

However, a few months later when facing an increase in registrations, Singa decided to hire a developer for Calm. In interviews with the developer, we learned that he was then working on the design of an "automated matching" system for the hosts and those seeking housing that could support, or even replace, the tedious work of the members of the association. By encouraging precise and useful questions in the forms, he thought that he would identify "around thirty parameters" that acutely describe the profile of the hosts. The database would, in fact, be adapted for future algorithmic processing.

His automated suggestion system would then propose "a qualified subset of host families" who, once validated by the person seeking housing, would be able to communicate "in a private online lounge, hosted on the platform". In doing so, the developer pushed the logic of automating the connection to the maximum, with mediation becoming purely algorithmic and supposedly neutral or fair because it was invisible and had no human intermediary.

But the developer and the association ended their collaboration because they could not reconcile their respective visions of the tool, and the system's launch was canceled. Indeed, Singa quickly understood that the connection between refugees and host families needed to remain human, not become

[8] "This is a certain type of (a) normative or (a) policy-based rationality based on the collection, aggregation, and automated analysis of mass data in order to model, anticipate, and preempt possible behaviors." Algorithmic governmentality is a strategy for streamlining and transcribing reality into digital data that can be used by IT systems. The uncertainty generated by human behavior is removed (in this case, the free expression fields) in favor of spontaneous, objective information, which no longer needs to be assessed. (Rouvroy and Berns 2013, 163–196).

automatic (via an algorithm). It realized that, in order to succeed in its refugee immersion project, putting people in contact with one another and matching them via a platform, like Airbnb, would not be enough.

"Hospitality cannot and must not be delegated to a machine." Singa realized that the core of its expertise (logistics, communication, diverse support, civic engagement, and collaboration with different government actors in charge of asylum procedures) was ultimately the best solution for finding accommodation for refugees and working towards their integration.

This experience led the Calm team to move towards a strategy that favors long-term accommodation (minimum 3 months), not the search for shelter in an emergency situation, and towards the set-up and multiplication of local teams throughout France. Today, Singa still provides online registration for hosts, but this is only a preliminary step. An offline protocol is now taking over from online registration in order to favor human selection or matching. Singa then provides training for the hosts and organizes meetings between those hosting and those who will be hosted.

After being tempted and experimenting with different algorithms and collaborations with developers, Singa is now refocusing on its know-how: giving logistical and organizational value to digital technology and letting volunteers do the rest. Their
hospitality expertise has taken the lead as a vector of trust and connections between the two parties and is proving particularly effective in managing the Ukraine crisis, seven years after Calm was founded.

4. The words of online hospitality

Once filtered[9] and anonymized, the Calm archives were explored through different dimensions (geographical, socio-demographic, etc.) and the changes in the vocabulary used by the hosts in the 'motivations' and 'tell us about you' fields were analyzed in greater depth.

In the following, we visualize the main dynamics of the question of hosting formulated by the actors themselves, from 2015 to 2022, between responses to emergency situations and a more thought-out desire to integrate a refugee. This observation first goes through a detailed exploration of the Calm archives from June 2015 to June 2017, and then we step back for an overview so we can examine the development of the entire semantic landscape extracted from the Calm archives until February 2022.

Between June 2015 and June 2017, Singa facilitated the hosting of nearly 300 refugees while registering 11,892 host registrations on its platform. The time distribution of these registrations is far from uniform, with almost

[9] Here we merge some fields adjacent to the successive versions of the form and remove exact names and addresses.

10,000 hosting proposals in September 2015 alone.

In August 2015, the Calm program had not been officially launched yet and the platform was only known to associations and activists. Nevertheless, the number of guest registrations exploded on September 3, 2015, rising from a few dozen to several thousand within a week. The worldwide coverage of the photograph of the young boy Alan Kurdi, found drowned on the Turkish coast, sparked a huge emotional response and generated an unprecedented wave of inscriptions.

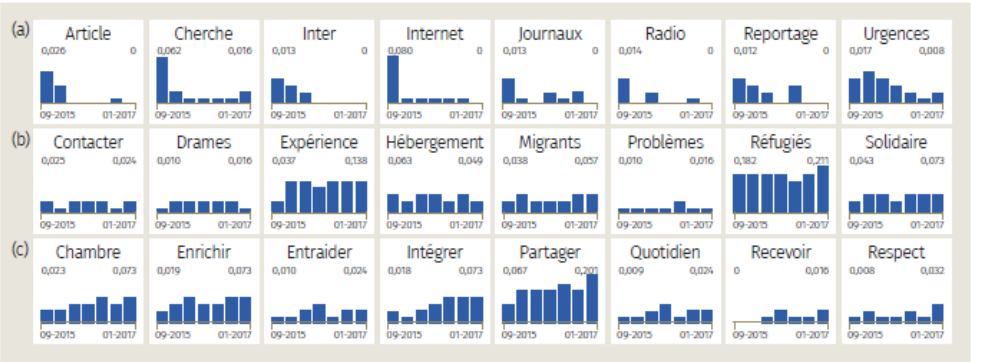
The quantitative analysis of the closed fields (age, place of residence, housing characteristics, etc.) in the form allows us to characterize these new hosts. They live in large French urban homes and offer simple rooms in small apartments or even just the living room with a sofa. Those who live in the countryside sometimes provide refugees with entire houses or vacant outbuildings that can accommodate much more than one refugee at a time. The hosts are young people in their thirties or parents who find themselves alone after their children have left home and some 50% of them are executives, retirees, or people with an intellectual profession (teachers, journalists, artists, etc.).

5. Media impact

The media coverage of Singa, in particular by France Inter, played a key role in the success of the Calm program. Between September 3 and 15, 2015, Singa was mentioned several times on the national radio, with a report, an interview, and a debate. In the ‘motivations’ field, the hosts then mentioned France Inter as the main vector of discovery for the association. The expression ‘I heard about it on France Inter’ becomes the common denominator of more than 500 registered via the online form. Listening to France Inter thus acted as a sociological marker at least as important as the age or profession of the hosts.

While hosting refugees for some families is a continuation of an already existing philosophical or political commitment (1968 protest participants, humanitarians, etc.) or even ties into a past family history (for descendants of immigrants), it seems that a majority of them said that they responded, first and foremost, to an emergency situation. For the first period of time, reports on the situation in the Aegean Sea broadcast daily on television and radio were pushing a non-activist segment of the population to look for, by their own means, a quick and concrete solution to the crisis.

Figure 1. Evolution of the frequency of use of words from the ‘Motivations?’ field.



This impression, which was reflected in the qualitative observation of the 'Motivations?' field, could be confirmed by an overall semantic analysis. To do this, we built a new collection based on all the statements of the hosts' motivations. We then browsed through these texts using automated 'scripts' and natural language processing tools to extract a vocabulary representative of the words most commonly used by the families. For each of these words, we drew a quarterly frequency curve normalized by the number of new registered users over the same period. This allowed us to track the evolution, over time, of the frequency of use of all the selected words (Figure 1).

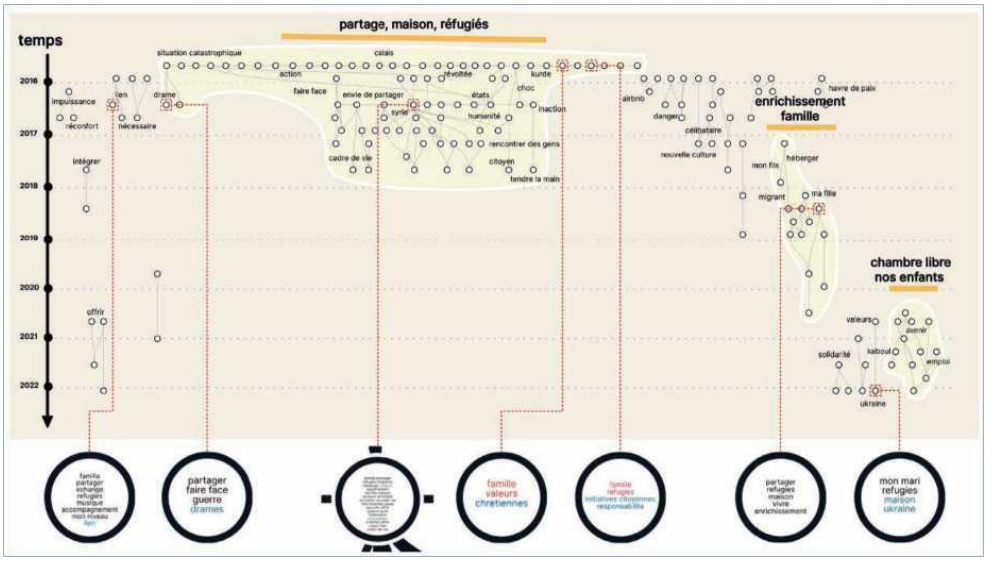
6. From Calm to J'accueille: Vocabulary development

These words can be classified into four categories: 1) words frequently used at the end of summer 2015 and whose proportion decreased afterwards (Figure 1, a), 2) words used in a stable manner from June 2015 to June 2017 (Figure 1, b), 3) words that became more important as the media coverage of the crisis decreased (Figure 1, c), and 4) words used in a non-characteristic manner. In fact, a vocabulary specific to each category can be identified: if the desire to remain in solidarity with the crisis and distress of refugees by offering them accommodation remained stable over time, it was first in response to the media shock (television, radio, reports) that made many hosts join the Calm program, as well as a response to the urgency of the situation. On the other hand, the more chronologically we move away from the epicenter of the crisis, the more well thought-out and developed the hosting experience becomes over time: the hosts say they want to integrate the refugees, host them comfortably and respectfully, help each other, and have a richer, more meaningful experience. The vocabulary becomes calmer, reassuring ('warmth', 'happy', 'beautiful', 'heart'), and is used as ways to express care and hospitality. Thus, the room shared in 2017 became more prominent than the living room sofa offered in 2015.

Has there been another emergency comparable to the summer of 2015?

To answer this question, we expanded our analysis and visualized on a macro scale the evolution of more than 600 words used in the field of ‘Motivations?’ from the Calm database (which has since become J'accueille) up to February 2022. To do this, we used a computational approach called phylomemy reconstruction. With the help of textual data search algorithms, we identified recurring and common patterns for the thousands of responses recorded in the form. These motifs or themes can be seen as groups of words frequently used together by different people in the ‘Motivations?’ field. They are represented in Figure 2 by circles. It was then possible to trace the evolution of these themes over time by linking them from one month to another using a semantic similarity measure. The structure thus obtained is called a phylomemy (Lobbé, Delanoë, and Chavalarias 2022) (Figure 2). It consists of a series of branches representing a coherent lineage of themes that respond to each other and change over the months.

Figure 2: Evolving semantic landscape (i.e. phylomemy of the entire Calm platform archive between 2015 and 2022



Valérie Mustapha Djamel, screenshot, 2022. © Chantal Capelli/TelecomParis.

First, the phylomemy confirms the importance of the 2015 crisis, marked by a record number of registrations on the platform and by a “semantic outpouring”: it was between 2015 and 2017 that the majority of welcoming words were invented. Beyond 2017, few new words enter into the phylomemy, except on the margin as historical milestones of migration events (‘Kabul’, ‘Ukraine’, etc.), which apparently did not have as much influence (in terms of the content of the database).

The main branch ‘sharing, home, refugees’ is characterized by the emergence of two major themes: on the one hand, the vocabulary of urgency, which is not included in the rest of the phylomemy, and on the other hand, the vocabulary of accommodation and sharing, which continues until February 2022. Starting in 2017, the hosting vocabulary stabilized around the concepts of encounters and enriching experiences. A central place is then given to children from host families (‘my daughter’, ‘our son’, etc.) either as a moral trigger (‘What will my children say if I do nothing?’), or as a fully-fledged driver for hospitality: by interacting with the people hosted, the child will encounter a new culture.

This analysis via phylomenies is only a first step towards a more complete search of the archives, but at this stage of our study, we can already note an interesting advance for human and social sciences in general, while remaining in the field of digital humanities and migration. In his 2015 article "The social sciences and the traces of big data”, Dominique Boullier[10] rightly wrote that digital traces only allow us to capture the “high-frequency vibrations” of our societies and not long-term social structures, nor even medium-frequency movements of opinion. This was absolutely true: overwhelmed by an unprecedented flow of digital traces, SHS researchers, supported by computational methods, focused primarily on detecting events, peaks, and variations of great amplitude in the present. We now believe that we can overcome this limitation thanks to the new methods introduced by complexity science (such as phylomemy reconstruction) and the methodical constitution of thematic archives by SHS researchers such as the Calm collection on which we have worked. It is now possible to study shifts in opinion over time, while changing the scale between high vibrations (emergency moments) and low frequencies (semantic invariants).

Conclusion

Exploring the millennium-old research field of hospitality, which has given rise to such extensive literature, with our methods from the digital humanities is not without risk. While recognizing the limitations of such research, we want to emphasize that—from a perspective that goes beyond this case study—refugees’ access to digital technology has changed hospitality practices in an unprecedented way, starting with the ‘place’ that is given to a foreigner and that this foreigner occupies inside the home. Today, hosting a connected refugee means bringing home a whole connected environment and its communication practices with its family

[10] URL: https://halshs.archives-ouvertes.fr/halshs-01141120/document

network, geographically distant but constantly nearby on a daily basis via ICT.

Thanks to these digital technologies, their presence is ubiquitous. How to host and help someone integrate who is present more or less completely (physically and remotely) in several locations? And to add to the complexity, what is this house that is called ‘Just like home’ when the hosts, via ICT, are themselves travelers on their own remote communications platform?

References

Gotman, Anne. 2001. *Le sens de l'hospitalité: Essai sur les fondements sociaux de l'accueil de l'autre*. Paris: PUF.

Morozov, Evgeny. 2013. *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: PublicAffairs.

Chavalarias, David. 2022. *Toxic Data: Comment les réseaux manipulent nos opinions*. Paris: Flammarion.

Casilli, Antonio A., and Julian Posada. 2019. "The Platformization of Labor and Society." In *Society and the Internet: How Networks of Information and Communication Are Changing Our Lives*, edited by Mark Graham and William H. Dutton, 293–306. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198843498.003.0018.

Rouvroy, Antoinette, and Thomas Berns. 2013. "Algorithmic Governance and Prospects of Emancipation: Disparateness as a Precondition for Individuation through Relationships?" *Réseaux* 177. https://doi.org/10.3917/res.177.0163.

Lobbé, Quentin, Alexandre Delanoë, and David Chavalarias. 2022. "Exploring, Browsing, and Interacting with Multi-Level and Multi-Scale Dynamics of Knowledge." *Information Visualization* 21. https://doi.org/10.1177/14738716211044829.

Web archiving in Tunisia post-2011: The National Library of Tunisia's experience

Raja Ben Slama

Abstract: The National Library of Tunisia had undertaken the digitization of written heritage and the "heritagization" of digital documents. To meet the second requirement, a web archiving unit was created outside the legal deposit service, as the organic law regulating this procedure only provided for the voluntary deposit of digital documents produced by publishers and authors, while social networks were becoming a virtual agora and a stage hosting several forms of literary and artistic creativity.

Keywords: BnT, heritagization, legal deposit, web archiving.

Tunisia's national library (Bibliothèque nationale de Tunisie – BnT) began to digitize its written heritage for preservation back in the 2000s and, from 2016 onwards, to offer free of charge, open access to Tunisian works already in the public domain. Digital technologies had proven their worth as a document preservation and dissemination tool. However, preserving native digital documents and *moving from the digitization of heritage documents to treating digital documents as heritage*, was one step further for which the BnT was not really prepared. Firstly because its thirty or so librarians, most of them trained at the Higher Institute of Documentation (Institut Supérieur de Documentation), gained their degrees before a web archiving module was introduced in 2018 and, secondly, because the Organic Act n°2015-37 of 22 September 2015 on registration and legal deposit makes no reference to this particular preservation method. Clause 4 of this act states that "Works of the types listed below that are made available to the public, in any form and on any medium, must undergo the registration and legal deposit procedures:

- all written matter and printed, engraved, illustrated, audio, audiovisual or multimodal documents, drawings, maps, photos, digital artworks, abstract words, and other content aimed at the general public
- software packages, databases, and the associated web pages and news sites."

However, the act does not specify the authorized body with whom websites are to be 'deposited' and makes no mention whatsoever of web archiving in the sections relating to the BnT. Despite its quest to be exhaustive, its openness to the new reality of document and artistic production, and its implicit abolition of the prior screening of publications,

Raja Ben Slama, Manouba University, Tunisia, raja.benslama@flah.uma.tn, 0000-0003-3059-7154

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Raja Ben Slama, *Web archiving in Tunisia post-2011: The National Library of Tunisia's experience*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.05, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 37-42, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

it simply places digital in the same media category as books. It falls short of the new socio-political and cultural reality brought about by the internet revolution and the closely associated Tunisian revolution.

Admittedly, by inventing the notion of *digital heritage* and encouraging the preservation of digital documents, the 2003 UNESCO charter on The Preservation of Digital Heritage goes some way to fill this legal loophole. However, where it falls short is that native digital documents are not elevated to the same status as classical heritage documents in every country. This involves a lengthy, almost involuntary process and there is generally a gap between manuscripts and books, on the one hand, and what archivists place in the 'non-book documents' category. Digital documents are short-lived, hence the need to archive them, but because of this ephemerality they are unfavorably compared to the great intellectual works which, when mentioned in Arab literature, are often accompanied by the adjective 'immortal' or considered to be the 'mothers of books'. It would take a shift in politics and document procedures for web archiving to become necessary and possible without the use of force. And, for Tunisia, the turning point was the revolution that broke out in the country on December 17, 2010. Triggered by a traveling salesman named Bouazizi setting himself on fire, it brought down the Ben Ali regime on January 14, 2011. How did this political event become a document and archiving issue without which web archiving could not have been contemplated?

Before the revolution, the executive government was using the legal deposit procedure for the purposes of prior censorship, thus preventing the recognition and circulation of the books published by opponents. The Head of the Legal Deposit Office at the BnT was a member of a Reading Committee that sat within the Ministry for Internal Affairs. The censored books were kept in a locked cupboard and could not be consulted. It was as if they had been imprisoned. Because of the revolution, they were once again made available to readers. The Reading Committee was abolished, as was prior censorship. It was on its own initiative that the national library extricated itself from the control of the political authorities. The web archiving initiative would not have been possible had this not happened.

There are several factors that contributed to the web becoming a key political issue in Tunisia and to the Tunisian revolution taking on archiving and digital significance. Before the uprising was sparked in December 2010, there was a revolt against internet censorship that some of its instigators referred to as a "virtual revolution". In May 2010, seven months before the people's revolt, young digital natives organized a campaign against internet censorship after the government closed down several dozen websites and blogs. The campaign was known as Ammar 404, referencing the error code 404 (file not found) and linking it to the first name of a public figure

(Ammar Bouzwir) who had been contacted to obtain the release of Salah (another public figure), the imprisoned madman depicted on an old, early-20th-century postcard.

After this pre-revolution for the liberation of the internet, the cyber activists, who could get round the censorship by using proxies (there's a Bendirman song about this), used blogs and social media to garner support, primarily circulating videos of the victims of repression and capitalizing on the fact that the written press and TV channels were under full governmental control. As such, they were almost the only source of media coverage at the local and international level.

In 2016, the threat was no longer the issue of government censorship but the extreme fragility of digital heritage. It was established that the lifespan of a domain name was around 3.8 years. This brings us to the notion of loss of heritage and to the findings of a 10th-century poet and critic from Kairouan, Ibn Rachiq (d. 1064 A.D.), who wrote in the book *Al-Umdah*: "It is reported that there is more good prose spoken by Arabs than good poetry. Of this prose less than a tenth survives; of poetry less than a tenth has been lost." Ibn Rashiq's aim was to commend poetry and demonstrate that rhyme and rhythm were factors of permanence and preservation. Moreover, I demonstrated that this was over-optimistic, that much of his own work had been lost and that only a tiny portion of Medieval Arab-Islamic works survive, particularly those related to Tunisia which was a juncture of struggles, influences, and migratory movements (Ben Slama 2022). We wanted to counter this fear of loss and had noticed that social media was becoming a virtual forum and a stage for several types of literary and artistic creation and that fewer and fewer periodicals were being published and websites were taking over. So we decided to create a selective archiving unit and a web manual without using automated extraction software. Several months later, we heard about the digital heritage of the revolution and the need to archive it. "Stored away in mobile phones, on desktops, on the web, and in foreign TV archives, these digital sources are perishable. Casualties of piracy and falsification, dispersed among amateur video-makers and photographers, Facebook accounts, dissident websites such as *Takriz*, image-sharing platforms such as *YouTube* and *DailyMotion*, websites of bloggers or news sites such as *Nawaat*, these digital revolution documents were doomed to disappear. The situation is alarming and, because it threatens the preservation of the nation's memory, it has not failed to attract the attention of so many archivists and human rights defenders" (Ben Achour et. al. 2021, 16).

Two courses of action were required: the retrospective archiving of revolution-related materials and day-to-day 'current' archiving. So the BnT joined forces with the multi-disciplinary group formed in 2016 of several public institutions (the National Archives of Tunisia (les Archives

nationales de Tunisie), the Higher Institute of Documentation (Institut supérieur de documentation – ISD), the Higher Institute for the History of Modern Day Tunisia (Institut supérieur d'histoire de la Tunisie contemporaine – ISHTC)), associations (FEMDH, the Euro-Mediterranean Foundation of Support to Human Rights Defenders) and voluntary researchers to collect and archive these documents that were dispersed over the internet and on the phones of ordinary people.

Not only had the BnT extricated itself from the prior censorship that had been secretly carried out in collusion with the executive government, but it was also able to fully fill its preservation function again and found itself inheriting revolutionary content it had to collect. It so happens that the word for revolution in Arabic (thawra) is an anagram of heritage (wartha).

The man who took on the role of representing the library within the group, Faycel Hamdi, was a young opponent of the Ben Ali regime and one-time member of the Progressive Democratic Party that was founded in 1983. As he himself had played a part in the revolution, he knew the Tunisian web well and also the ins and outs of publishing on dissident blogs and social media. The National Archives of Tunisia had been tasked with pulling together the videos and photos so Faycel set about collecting caricatures, sketches, graffiti, songs, slogans, poems, tweets, and Facebook posts. The National Archives collected, dated, authenticated, cataloged, and saved 1,100 videos and as many photos. The parties got together on 11 March 2017 to take stock of the situation and came up with the idea of staging an exhibition of the "digital heritage, but also the documentary, artistic and satirical heritage of the revolution" entitled *Before the 14th, instant tunisien* (Archives of the Tunisian Revolution) and producing a trilingual catalog. The exhibition opened at the Bardo National Museum, under the patronage of the Head of State, and ran from January 14 until March 31, 2019. It opened at the Mucem (Museum of European and Mediterranean Civilizations) in Fort Saint-Jean in Marseille on March 19, 2019.

The internet holds documents with a limited lifespan that need to be archived and saved, but it also conceals a matrix of things that no longer exist in real life. For instance, the walls of the sumptuous villas of the family and stepfamily of President Ben Ali were defaced and covered in humorous revolutionary graffiti. Unfortunately, to the great dismay of the 'archivists', the authorities confiscated and repainted these walls. However, there was still some evidence of this graffiti on the web and we were able to reproduce at least part of it.

Once the revolution heritage had been archived, the TnT undertook other targeted initiatives, among them the archiving of the 2019 *hirak* movement in Algeria, given its African and Maghreb remit.

During lockdown, 143 webinars (a total of 194 hours) were recorded.

Manual, selective daily web archiving began on November 1, 2016. The decision was made to archive the opinion pieces published on the news sites and the Facebook and Twitter pages and accounts of intellectuals, activists and opinion leaders and also caricatures and photos with informative and/or artistic value that were in the public domain. To reconcile copyright, privacy protection, and preservation, the following restrictions were agreed:

- Only public (and not private) posts would be kept. For Facebook, for example, the security setting must be public and not friends only.
- Until there was clear legislation in place, the information would not be made open access. It would be available for consultation at the library.

On February 15, 2017, a further two archivists joined the group and the number of websites, blogs, and Facebook and Twitter pages and accounts scrutinized each day increased. According to the group's 2022 report: around 120,000 documents have been saved, equating to 88.6 GB (gigabytes) of files. These documents have been archived as non-editable files and saved in two locations:

- Storage arrays for long-term retention,
- NAS servers for ease of access and use.

I remember the debate surrounding the technology to be used to catalog the documents. A technical solution was to be provided by the supplier of the new BnT platform that went online in May 2022. However, the archivists had already created a tree structure for the web archiving catalog with a menu linking the different categories (videos, caricatures, graffiti, posts, tweets, etc.) to the authors' names and a calendar, which provides the browser with several entries and consultation options.

Conclusion

Many documents archived by the BnT since 2016 had already disappeared from the net by 2023. Others had been changed, for instance the videos of comic singers had had the sound removed. The new challenges facing national, and more generally, archive libraries are the 'digital deluge'—and the volatility of its fallout—and the risk of partisan manipulation that has grown with the progress of artificial intelligence. These will change their role and *raison d'être*. In the past, librarians fought a constant battle against book-destroying insects (cockroaches, termites, booklice, etc.). The threat today comes from faceless, unspecified protagonists in a rapidly growing and increasingly sophisticated document-related environment.

References

Ben Slama, Raja. 2022. “Parmi les vestiges du patrimoine écrit”, Inaugural talk of the Rencontres humanités numériques, *Insaniyyat Tunis*, Cité de la Culture, Thursday 22 September 2022. Translated by Kmar Bendana. https://hctc.hypotheses.org/3732

Ben Achour, Rabâa, Bel Haj Yahia, Fathi, Ben Mhenni, Sadok, and Mkada, Amina. 2021. *Before the fourteenth, Instant tunisien*, directed by Rabaa Ben Achour, Tunis: Ministry of Cultural Affairs.

SECTION 2
Rethinking collection creation for cultural and societal change

Bridging the gap: Capturing UK trans health discourse in the Archive of Tomorrow

Alice Austin

Abstract: The barriers trans and non-binary people in the UK face when accessing healthcare have been well documented in recent years, and a proliferation of sites produced by and for trans communities have emerged to bridge the gaps left by suspended services and growing waiting times. Concurrently, a number of high-profile legislative cases and public debates have underscored the extent to which the provision of information about trans* health is defined and shaped by societal and political contexts. This chapter discusses the challenges of collecting online trans* health information in a rapidly changing and hotly contested environment, and explores the questions around representation and the ethical implications of collecting online health discourse.

Keywords: ethical collection, representativeness of collections, health information, contentious collecting.

In traditional conceptions 'the archive' functions from a position of neutrality, operating as a storehouse for the passive accumulation of information about the past that is maintained for the benefit and use of the future. More recently, however, it has been acknowledged that rather than reflecting our present reality, archival preservation recreates and reaffirms it, or, as Eric Ketelaar (2001) has argued, "the archive reflects realities as perceived by the 'archivers'" (Ketelaar 2001, 133).

This position has been hugely informed by the 'memory boom' that characterized the late twentieth century, and the attendant rise in community archiving initiatives (Miztal 2010). As the archival profession has begun to attend more closely to the social, mnemonic, and affective aspects of archives, the concept of *representation* has become central to our understanding of the function that archives and archival collections play in society. Examining the impact of archival representation on communities who have traditionally been "ignored, misrepresented, or marginalized" by mainstream repositories, Michelle Caswell et al. (2016) have argued that feeling represented in an archives "has an ontological impact":

> …it changes [the viewer's] sense of being in the world; she can 'discover' herself 'existing' in ways she did not before this record was created and made accessible. Representation in community archives catalyzes this ontological shift from not being/not existing/not being documented to being/existing/being documented, with profound personal implications.
>
> (Caswell, Cifor, and Ramirez 2016, 61)

Being cognizant of these implications, then, and aware that if the archive attempts "to collect everything … it will soon succumb to entropy and

Alice Austin, University of Edinburgh, Scotland, United Kingdom, alice.austin@ed.ac.uk, 0009-0007-5586-2571

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

chaos" (Spieker 2017, xiii), archivists and curators have had to grapple with the question of which communities, which stories, which *realities*, to represent in their collecting. The difficulty of this is only magnified in the digital sphere, where digital records exist and operate within multiple networked realities simultaneously, and the short lifespan of web content means that archivists do not have the luxury of waiting for the flotsam and jetsam of documentary detritus to wash up upon archival shores.

These challenges were writ large as institutions in the heritage sector rushed to capture a picture of the Covid-19 pandemic from a mercurial and rapidly evolving digital landscape. The impacts of the pandemic were not evenly felt, and research has suggested that ethnic minorities and other already marginalized communities were more heavily impacted by the outbreak (Platt 2021): how can curators reflect and represent such a myriad of different experiences? Deep divisions emerged around what constituted appropriate medical, legal, and social responses to the pandemic: how can collecting respectfully and responsibly reflect the dissent and divisions in a moment without a single, unifying narrative? Key government and medical websites were updated on a daily basis as new information emerged, and social media reacted quickly to find, debate, and digest each new study or guideline: when moving at speed, what are the ethical implications of such 'rapid response' collecting?

Emerging from these observations, the Archive of Tomorrow project sought to explore these questions in more detail. After introducing the project in brief, this chapter will then focus on the trans*[1] health subcollection as an exemplary microcosm of the collection as a whole. It will detail how the subcollection evolved; shed insight on the ethical considerations that contributed to its development; and conclude by exploring what the experiences of this project can tell us about creating 'representative' collections.

1. Project background

The Archive of Tomorrow (AoT) project sought to build a collection of archived websites to reflect how online spaces were used to share, discuss, and debate issues around health in the aftermath of the pandemic. A Wellcome Trust-funded initiative that was led by the National Library of Scotland with extensive support from the British Library, AoT sought to explore best practices in preserving, describing, and enabling access to information captured from the web. The project team comprised three web

[1] Trans* is an umbrella term referring to a number of identities within the gender identity spectrum. The use of an asterisk expands the definition beyond binary trans identities (i.e. transmen/transwomen) to include non-binary and gender-fluid identities.

archivists (one based with each academic partner), a metadata analyst and a rights officer, and was led by a project manager (all of whom were based at the National Library of Scotland). As well as the tangible objective of preserving a collection of 10,000 targets relating to health, the project also had exploratory aims around how to ethically collect from the web; how to republish responsibly; and what is needed to increase research usage of web archives.

Collection was performed within the context of the UK Web Archive (UKWA), a partnership of the six Legal Deposit Libraries (LDLs) that performs the web function of the LDL's legislative responsibility to collect and preserve a copy of all material published in the UK and Ireland. The UKWA has been systematically collecting non-print material since 2013, with the majority of material being captured through an annual domain crawl that attempts to make a copy of any content published to a website with a recognizable UK top-level domain (e.g., .uk, .scot), or hosted on a server physically located in the UK (identified via a GeoIP lookup). The yearly crawl is supplemented by curated collecting which is achieved by manually adding targets to the Annotation and Curation Tool (W3ACT), a web-based interface that allows a user to create an entry for a specified URL, establish parameters such as depth or frequency of a crawl, and record metadata for description and rights-management purposes.

Curated collections are made available via the UKWA's public interface, where there are over 100 thematic collections available to browse. The collection resulting from the AoT project (since named Talking About Health) comprises around 3,500 targets, and has been further subdivided along various lines (such as source, form, focus, etc) to allow for navigation and discovery. The Regulations that govern legal deposit impose some constraints on collection and access: they only allow for the collection of material that has been made publicly available and do not cover material made available to a 'restricted group' (i.e. requiring an individual to provide credentials to access), nor do they cover material that is predominantly audio-visual in format. Additionally, access to archived material is restricted by default to users at computer terminals onsite in LDLs, unless permission for access has been explicitly granted by the website owner.

2. Trans* health in the UK

The subject of trans* health was selected for focused collecting as it exemplifies both how the digital sphere has transformed contemporary approaches to health information, and how the collecting of such information is complicated by the social and legal contexts in which it exists. Recent years have seen a sustained increase in the media coverage and commentary on the provision of gender-related care and treatment in

the UK. The treatment of gender dysphoria in children has become a particular area of debate, with a number of high-profile legal cases and inquiries being conducted into questions such as the competency of minors to consent to medical care and the long-term impacts of medical treatments such as puberty blockers. The barriers that trans* and non-binary people face when accessing healthcare in the UK were already in dire straits in 2018. An uneven geographical distribution of gender services both across and within the four nations results in many people being required to travel long distances to access healthcare services, and one study released in that year concluded that long wait times "exacerbate gender dysphoria and mental health problems, and increase risks of suicide and self-harm" (TransActual 2022). This was made significantly worse by reduced access to medication and transition-related care as a result of the Covid-19 pandemic, with the average waiting time for a first appointment at an adult gender dysphoria clinic rising to around 38 months in recent years.

Furthermore, there is an increasingly toxic culture of debate surrounding the issue of trans* health and rights, and indeed, on the question of whether trans* identities are or should be considered valid: a recent court ruling concluded that 'gender critical' beliefs—broadly put, that a trans* person's internal feelings about their gender identity has no basis in material reality—constitute a philosophical belief that is protected under the Equality Act and the European Convention on Human Rights (Forstater v CGD Europe 2021). In a 2021 report the Council of Europe observed a "baseless and concerning" level of transphobia, and noted that "rhetoric … which denies trans identities … is being used to roll back the rights of trans and non-binary people and is contributing to growing human rights problems" in the UK (Council of Europe 2021).

3. Producing the trans* health subcollection

It was against this backdrop that collecting took place. The trans* health subcollection is comprised of 76 URLs and includes information published by providers both within and external to the publicly funded healthcare systems; gray literature and guidance on the delivery of healthcare; campaign sites relating to the provision of trans* healthcare; peer-to-peer information sharing sites; and social media discussion. A 'top-down, center-out' approach to identifying material was employed, with initial efforts focused on material published by service providers operating at the national or top level (NHS, private providers), followed by the regional or local instances of those services. Next, material which addressed the delivery of those services was targeted: this included best practice guidance for individual providers 'on the ground', as well as monitoring and advocacy regarding service provision at a national and international level. From these

targets, key areas of discussion and debate emerged which could then be used as access points for the identification of peripheral or 'bottom-up' discourse on social media.

During collecting, a decision had to be made regarding the extent to which 'transceptical' or gender-critical sites would be included in the collection, and if/how these would be described to users. The project had initially adopted a framing of 'information vs. misinformation', but as collecting progressed and the complexities of the documentary landscape emerged it became clear that such a binary distinction was unhelpful: not only were the project team unqualified to make judgments about the veracity, reliability, or appropriateness of a source's content, it was also felt that attempting to distinguish between information and misinformation in this way would lead to a misrepresentation of the context in which health information is located, accessed, and understood. Instead of trying to determine information from misinformation, then, the project team instead sought to collect all relevant material that could be found on a subject in order to better reflect the documentary landscape at the time of collection.

A vibrant culture of 'information activism' has emerged around the subject of trans* health that the project team felt it was important to capture. Sites offer commentary, provide peer-to-peer support for trans* people, and generally seek to bridge the gaps in trans* healthcare provision by collating information on specific medicines (such as guidance on safely acquiring and self-administering hormones in the absence of a prescription) or by sharing first-hand accounts and experiences of treatments, procedures, and providers. The sharing of information about trans* healthcare therefore serves to counter perceived social and systemic barriers to medical treatment and support. One such site, Trans Healthcare Intelligence, sums up their mission thusly:

> Accurate and useful information about trans healthcare in the UK is difficult to come by, limited by a transphobic medical system as well as targeted harassment from hate campaigners… This resource aims to collect information about transgender healthcare and our community experiences of a system not designed to cope with our existence, ensuring it's as accessible as possible to our community.
>
> (Trans Healthcare Intelligence)

While there is a reasonable expectation that information published by official or authorized sources such as the NHS will be preserved through other channels, many of these peer-to-peer initiatives exist only in their web-based form with no supporting infrastructure and no regulatory record-keeping duties: they exist only as long as interested individuals have the means and motivation to maintain a website. It is in the capacity to capture and preserve these traces of a community and a documentary landscape that has routinely been excluded from the historical record that the value of web

archiving—as a route towards a more representative and diverse documentary record—can be most clearly observed.

4. Ethical considerations

In addition to the grassroots peer-to-peer sites that were targeted, the subcollection also includes captures of many UK-specific threads, accounts, and forums on social media platforms such as Twitter/X, Reddit, and Tumblr. When approaching this material the project partners had to carefully consider the need to balance the research value of capturing social media discourse against the risk of bringing harm or distress to individuals. There is a growing body of literature exploring the ethical challenges of collecting and using social media posts for academic research that the project team was able to draw on when conceptualizing these challenges and how to address them, and the issue of implied vs. informed consent required particular attention. As Hunter et al. (2018) have noted, although "consent for usage and collection of data are usually implied via [a] platform's terms of service" the extent to which this can be considered 'informed consent' is questionable, and social media users "may not necessarily expect their personal data to be used for research purposes" (Hunter et al. 2018, 345) Furthermore, there was a concern that collecting these social media sources may undermine the social logic by which such platforms and spaces operate. As Nicholas Norman Adams (2022) observes, "many Reddit forums position themselves as 'safe spaces' where users can discuss various struggles. Users posting on these forums do so in the knowledge that postings are contextualized within a wider, local topic board conversation: i.e. the 'safe space', which is policed by local online moderators" (Adams 2022, 52). To remove this content from this safe space, then, significantly changes the context in which any implied consent is given. This becomes an even more pressing concern when considering the potentially sensitive nature of the topics under discussion here. The project team recognized that the long-term archival implications of posting online may not be at the forefront of an individual's mind when turning to the internet for information on health-related topics, and particularly considering the possibility that posts were made at a time of crisis or distress.

Similarly, the nature of the topics under discussion within this subcollection required the project team to be mindful of the potential risk that capture and preservation might pose to creators. In their efforts to develop a framework and toolkit for the ethical collection and use of social media content, the *Documenting the Now* project team recognized that "while the benefits of social media to the democratization of information access are clear, the abundance of and access to social media content and

data by countless third parties also presents opportunities for some to 'weaponize' the platforms and the data they generate in ways that can cause harm to marginalized and already vulnerable communities" (Jules, Summers, and Mitchell 2018, 3). In October 2022 draft guidance was published indicating that young people in England who access medication or treatment for gender dysphoria without the support of NHS clinicians may be referred to safeguarding agencies, including the police (Topping 2022). This presents questions as to the extent to which the information being shared on these personal blogs, message boards, and other sites could conceivably be construed as promoting the use of controlled substances or—at the extreme—encouraging child endangerment. Adams' exploration of the ethical challenges of using social media content in scholarly research noted that "replication of Reddit user postings—verbatim—in scholarly publications can often lead to internet reverse-searching. In some cases, this could allow the original Reddit threads to be easily and rapidly located online, therefore risking invalidation of any assumed 'participant' anonymity and allowing the linking of specific isolated comments used in publications to specific user accounts and postings" (2022, 7). It is not inconceivable that a user may remove content they have shared in an attempt to protect themselves from potential legal action, but then "discover that their comments now exist in a permanent archive, for which they have no control over the ways in which such comments are used; no autonomy and decision over the deletion of these materials, nor access to any procedure from which to de-associate these comments with their Reddit username" (Adams 2022, 9). While there is a clear argument for the historic and social value of preserving such material, then, it is also important for curators and collecting initiatives to be aware of these issues and to consider what responsibilities the archive has in relation to content creators.

As noted above, a minimal amount of description has been applied to sites within the Talking About Health collection. Sites were assigned to the main collection using W3ACT's tagging function, and from there, could be further assigned to one or more sub-categories that had been chosen to aid navigation and discoverability. These low-level descriptors were largely intended to describe the publisher rather than the content—denoting a target as being NHS-published, or a social media resource, or a charity website, for example—and the tagging function was also used to group sites along a theme, producing subcategories like the trans* health subcollection.

However, even minimal levels of description and arrangement can influence how a resource is understood by a future researcher. In relation to this subcollection, the risks of descriptive choices exoticizing and 'othering' an already marginalized community were apparent. Historian Jules Gill-Peterson (2022) has argued that the "material difference between transgender healthcare and non-transgender healthcare…is transphobia".

> The medical resources needed to transition are not of a different species than the equally numerous ways that non-trans people's sex and gender are routinely medicalized. Yet they are treated fundamentally differently. Although they share the same clinical and scientific history, one is treated as new, experimental, and potentially dangerous, while the other is rarely the subject of sustained news coverage at all. One is treated as always arriving too quickly while the other is treated as so unremarkable it is as if it has always existed.
>
> (Gill-Peterson 2022)

This siloing of trans* health concerns has very real consequences, with Wall et al. exploring how 'trans broken arm syndrome' (a form of medical discrimination faced by transgender and gender diverse patients wherein healthcare providers "conceptualize patients through their transgender identity first, and chief complaint second") can adversely affect the level of care that trans* people receive (2023, 18). Recognition of this required the project team to consider whether by isolating 'trans* health' from 'cis health' our collecting practices might be compounding the othering and exclusion of a marginalized group of people and how our descriptive practices can better reflect the ways that these communities view, understand, and describe themselves.

5. Conclusions

In many respects, the trans* health subcollection can be understood as a microcosm of the Talking About Health collection as a whole. As the UK's legal gender recognition processes require a clinical diagnosis of gender dysphoria, this topic is particularly illustrative of how questions of health are entwined with debates in other areas such as politics, science, or law; and the tangled questions of authority and representation that arise as a result can be clearly observed in the subcollection.

The subcollection is also exemplary of the way that information ecosystems emerge around communities with particular health issues or concerns, and it is in the potential to capture this 'information activism' that the value of web archiving tools for producing a more representative and inclusive historical record can be observed. As Andrew Flinn (2007) has noted, where the conventional archive does document historically marginalized or excluded communities "it … rarely allows them to speak with their voice, through their own records". Instead, "traces are generally one-dimensional, often reducing individuals to statistics, appearing as problems, occupations, rigid ethnic or faith-based identities which minimize or ignore complexity and deny them their own voice" (2007, 152; 160). Web archiving can therefore be seen to offer a corrective to this, but it is important that we recognize that "the internet affords the luxury of a certain amount of distance to be able to observe people, consume information

generated by and about them, and collect their data without having to participate in equitable engagement as a way to understand their lives, communities, or concerns" (Jules, Summers, and Mitchell 2018, 3). Proponents of participatory archiving practices that invite communities to create or describe archives in their own ways have suggested that such approaches can "have an impact in diversifying and democratizing heritage" (Flinn 2007, 165) and while many mainstream organizations have experimented with inviting communities into the archival process through crowdsourced description projects or by soliciting contributions of material, it has been argued that such approaches reinforce the claim of the archive to 'speak for' communities: that is, in their control of the terms on which the community can engage, power over final decisions regarding appraisal, arrangement, and description still rests with the 'experts' (Eveleigh 2015). If web archiving programs are to engage such methods in search of a more representative record, then, we need to work with communities to find sustainable, respectful, and equitable avenues for participatory collection building.

Furthermore, the highly politicized atmosphere around the topic of trans* health made collecting this topic particularly challenging, and Eira Tansey's observation that the historical record should not be "a high priority while people are trying to keep their shit together and attempt to not die" is particularly pertinent in the context of building a collection like the trans* health subcollection (Tansey 2020). When the subject under debate is kidney stone treatment, very few would question the right of someone experiencing symptoms to access healthcare, and even fewer would express doubt about the existence of kidney stones or kidney stone pain in the first place. In contrast, discussion surrounding trans* health issues can (and regularly does) include questions over the legitimacy of trans* identities, and the extent to which they should be recognized and respected by law. Such questions can be distressing to witness even for those outside of the trans* community. Before we ask individuals to frame and examine their personal lived experiences in this way, it is crucial that we ensure they can be adequately supported in this work. Returning to the observation on the power of archival representation that opened this chapter, we must remain aware that this is a power that must be wielded responsibly—and consider what it means to 'discover yourself existing' in a context that constitutes your existence as deviation from 'the norm'. As Tansey cautions, archive and heritage professionals must recognize "that respecting people's privacy and right to forget their own past means accepting that we will lose parts of the historical record that others may wish we had gone to great lengths to get" (Tansey 2020).

References

Adams, Nicholas Norman. 2002. “Scraping’ Reddit posts for academic research? Addressing some blurred lines of consent in growing internet-based research trend during the time of Covid-19.” *International Journal of Social Research Methodology* 27, no. 1: 47–62. https://doi.org/10.1080/13645579.2022.2111816.

Caswell, Michelle, Marika Cifor, and Mario H. Ramirez. 2016. “To Suddenly Discover Yourself Existing: Uncovering the Impact of Community Archives.” *The American Archivist* 79, no. 1: 56–81.

Council of Europe Committee on Equality and Non-Discrimination. 2021. “Combating rising hate against LGBTI people in Europe Doc. 15425.” Belgium: Council of Europe. Retrieved 7 March 2024. https://web.archive.org/web/20230228040752/https://pace.coe.int/en/files/29418/html

Eveleigh, Alexandra Margaret Mary. 2015. “Crowding out the Archivist? Implications of Online User Participation for Archival Theory and Practice.” PhD diss. University College London. https://discovery.ucl.ac.uk/id/eprint/1464116

Flinn, Andrew. 2007. “Community Histories, Community Archives: Some Opportunities and Challenges.” *Journal of the Society of Archivists* 28, no. 2: 151–76. https://doi.org/10.1080/00379810701611936.

Gill-Peterson, Jules. 2022. “Three Questions for Every Paper of Record That Publishes a Story on Trans Healthcare.” *Sad Brown Girl* (blog) *Substack*. 15 June 2022. Retrieved 8 February 2024. http://web.archive.org/web/20220630001614/https://sadbrowngirl.substack.com/p/three-questions-for-every-paper-of

Hunter, Ruth F., Aisling Gough, Niamh O’Kane, Gary McKeown, Aine Fitzpatrick, Bergis Jules, Ed Summers, and Vernon Mitchell. 2018. “Documenting the now-white paper: ethical considerations for archiving social media content generated by contemporary social movements: challenges, opportunities, and recommendations.” Documenting the Now. Retrieved 23 January 2024. https://web.archive.org/web/20240202062719/https://www.docnow.io/docs/docnow-whitepaper-2018.pdf

Ketelaar, Eric. 2001. “Tacit Narratives: The Meanings of Archives.” *Archival Science* 1: 131–41. https://doi.org/10.1007/BF02435644.

Platt, Lucinda. 2021. “Why ethnic minorities are bearing the brunt of COVID-19.” *Research for the World* (blog) *London School of Economics and Political Science*. 9 November 2021. Retrieved 17 Jan 2024.

https://web.archive.org/web/20230925103916/https://www.lse.ac.uk/research/research-for-the-world/race-equity/why-ethnic-minorities-are-bearing-the-brunt-of-covid-19

Spieker, Sven. 2008. *The Big Archive: Art from Bureaucracy*. London: MIT Press.

Tansey, Eira. 2020. "No One Owes Their Trauma to Archivists, or, the Commodification of Contemporaneous Collecting" (blog). 5 June 2020. Retrieved 15 Jan 2024. https://web.archive.org/web/20200627194117/http://eiratansey.com/2020/06/05/no-one-owes-their-trauma-to-archivists-or-the-commodification-of-contemporaneous-collecting/

Topping, Alexandra. 2022. "Young trans people accessing treatment outside NHS may get safeguarding referral." *The Guardian*. 14 October 2022. Retrieved 6 February 2024. https://web.archive.org/web/20221102130414/https://www.theguardian.com/society/2022/oct/14/young-trans-people-accessing-treatment-outside-nhs-may-get-safeguarding-referral

Trans Healthcare Intelligence. 2023, "Trans Healthcare Information." Retrieved 12 February 2024. http://web.archive.org/web/20230815183354/https://www.transhealthcareintel.com/

TransActual. "Bridging Prescriptions." Retrieved 9 February 2024. https://web.archive.org/web/20240121233504/https://transactual.org.uk/bridging-prescriptions/

Hunter, Ruth F., Aisling Gough, Niamh O'Kane, Gary McKeown, Aine Fitzpatrick, Tom Walker, Michelle McKinley, Mandy Lee, and Frank Kee. 2018. "Ethical Issues in Social Media Research for Public Health." *American Journal of Public Health* 108: 343–348. https://doi.org/10.2105/AJPH.2017.304249.

Wall, Catherine S.J., Alison J. Patev, Eric G. Benotsch. 2023. "Trans broken arm syndrome: A mixed-methods exploration of gender-related medical misattribution and invasive questioning." *Social Science & Medicine* 320.

Making social media archives: Limitations and archiving practices in the development of representative social media collections

Beatrice Cannelli

Abstract: Social media has become an important digital space where individuals can participate in ongoing global discussions and document instances of historical events. Social media offers marginalized communities a means to express their identities, voice their concerns, and tell their stories. Archiving institutions have started to include social media in their collections because of its enduring value. However, constraints set by legal and technical frameworks and limited resources available at single institutions can influence the overall representativeness of content archived on social sites. This chapter explores the impact these constraints have on the development of representative social media collections and illustrate participatory approaches that can help to mitigate concerns.

Keywords: social media archiving, representativeness of collections, participatory archive.

Social media has become an important digital space where individuals can participate in ongoing global discussions, offering at the same time a platform for sharing and documenting instances of historical events, as health and political crises in the early 2020s have demonstrated (Simon 2012; van Dijck 2011). Moreover, social media provides an opportunity for marginalized communities to express their identities, voice their concerns, and tell their stories (Bergis et al. 2018). The cultural value and historical relevance of social media content has been widely recognized (Henninger and Scifleet 2016; Pietrobruno 2013), leading cultural heritage institutions worldwide to include the material generated on these sites in their preservation strategies in order to ensure its safeguard and accessibility in the long term (Bingham and Byrne 2021; Fondren and Menard McCune 2018; Schafer and Winters 2021; Storrar 2014). Social media archiving initiatives have the ability to preserve fragments of our (online) present, passing down to future generations of researchers key information to understand the 21st century. For this reason, it is essential that the plurality of voices emerged on social platforms is adequately reflected in the resulting archive collections. However, developing social media archives has proved to be a difficult endeavor under many points of view. Although social media archiving inherits some of the challenges identified over more than 25 years of web archiving activities, web curators have been dealing with a series of new technical, ethical, and legal issues that are specific to social media sites and have been limiting the scale of collection, thus potentially influencing the granularity and representativeness of archived

Beatrice Cannelli, University College London, United Kingdom, beatrice.cannelli@postgrad.sas.ac.uk, 0000-0002-8645-9503

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Beatrice Cannelli, *Making social media archives: Limitations and archiving practices in the development of representative social media collections*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.08, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 57-75, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

social media collections (Thomson 2016). Also, appraising and selecting content out of the sheer amount of information generated daily on social platforms requires time and resources that web archiving teams often do not possess.

Scholarly literature has discussed concerns of comprehensiveness and representation in web collections focusing on archiving strategies, the influence of socio-technical infrastructures, and cultural perspectives (Bingham and Byrne 2021; Brügger 2018; Hegarty 2022; Maemura 2023), calling for a critical approach to web archives (Ben-David 2021). However, few have addressed the questions surrounding the representativeness of social media archived material (Chambers et al. 2021; Schafer et al. 2019). The unique dynamics that regulate social platforms, the ephemerality of content, and the distinctive curatorial challenges that archiving this born-digital material pose to collecting institutions, call for further examination of the limitations and practices related to the development of representative social media collections at a national level. Drawing from interviews[1] and fieldwork conducted as part of wider, ongoing PhD research investigating the challenges and opportunities related to the development of social media archives, this chapter explores the factors that may impact the degree of representativeness of social media collections and the actions taken by existing social media archiving initiatives to mitigate these concerns.

In the first section, I will delineate the context in which social media archives are embedded, drawing attention to platforms' representation, geopolitical dynamics, and archival narrative disparities. I will consider the popularity of certain platforms and how this is not always mirrored in the collections developed by existing web and social media archiving initiatives in the Global North. In the second section, I will discuss how the need to preserve this important historical resource often collides with the numerous constraints imposed by national legal frameworks, social media policies, technical challenges, and inadequate resources necessary to guarantee the long-term sustainability of archiving efforts at an institutional level, setting the stage for representation concerns, biases, and narrative gaps in national social media collections. In the final section of this chapter, I will identify some of the steps taken by existing web and social media archiving initiatives to mitigate representativeness and inclusivity concerns. I will conclude by arguing that the use of crowdsourcing strategies and participatory approaches are examples of good practices that can not only sustain the development of more comprehensive collections, but also offer a

[1] Semi-structured interviews were conducted between April and September 2022 with twelve web archivists at national archiving institutions currently archiving or planning to archive social media. Insights and examples that emerged during the interviews are referenced in the footnotes.

unique opportunity to raise awareness of the existence of social media archives and their cultural significance across diverse layers of society.

1. Questions about inequalities and representation in the social media archiving landscape

A rich body of post-modernist archival literature has discussed the meaning of representation in the archives specifically with regard to selection practices, highlighting the potential repercussions that archival choices may have on the history told through the cultural heritage thus preserved (Caswell et al. 2017; Yakel 2003). In particular, concerns have been raised about biases existing in mainstream narratives and how collecting practices have frequently led to documenting one side of history, silencing groups of people placed at the margin of society because of structural power dynamics (Harris 2002; Jimerson 2006; Schwartz and Cook 2002).

With the advent of social media platforms, plus a diffused democratization of mobile devices and access to the internet, many of those marginalized voices have found multiple virtual spaces where they could make themselves heard. As Barrowcliffe (2021) noted, social media offered minority groups a means to convey counter-narratives, documenting, from their standpoint, critical events related to their own history, as these unfold on both a national and a global scale. For this reason, archiving social media represents an unprecedented opportunity for memory institutions to preserve historical traces of the present that have the potential to portray the multi-leveled landscape of voices prompting or joining conversations on these sites. A kaleidoscope of stories coming from communities that have often been underrepresented or misrepresented in mainstream media and repositories.

However, while social media has been amplifying certain protests, movements, and events contributing to the online unfolding of viral phenomena such as those expressed through the hashtags #BlackLivesMatter or #MeToo, it still mirrors societal and geographical inequalities existing offline, if not exacerbating some of those differences (Jackson 2020; Lutz 2022). The roots of these inequalities, as Lutz (2022) explained, are to be found in the different layers of social media divide, which involves among other factors the uneven distribution of access to not only mobile internet, for example, but also to the plethora of existing social platforms that may differ from country to country, with subsequent repercussions on the empowerment of certain marginalized groups rather than others (Lutz 2022).

The social media divide stemming from geopolitical dynamics, as well as other aspects that will be discussed below, appears to have heavily influenced the geographical distribution and development of social media

archiving initiatives. While numerous web archiving initiatives have emerged at various national memory institutions, consolidating over the past twenty years techniques and collection strategies to safeguard national Top-Level Domains (TLDs), the preservation of social media is still finding its pace and space, with only a few countries consistently archiving this born-digital material. As I reported in a blogpost recently published on the International Internet Preservation Coalition (IIPC) blog, the preservation of social media material appears to be fundamentally located in Global North countries, especially in North America, Europe, and Oceania (Cannelli 2022). This uneven distribution has raised questions about the potential gaps in the overall preservation of the collective memory generated on social platforms (Cannelli 2022). The reasons behind these discrepancies include geopolitical factors and challenges that are still unresolved, which makes this material particularly difficult to collect and provide successful access to (Bruns and Weller 2016; Pehlivan et al. 2021; Thomson 2016). As emerged from the aforementioned preliminary study, imbalances are also to be found in the type of social media that are currently being preserved by Global North cultural heritage institutions. Among the most archived platforms to date there is Twitter (officially rebranded as X in April 2023), followed by Facebook and Instagram; conversely, sites such as YouTube and TikTok, which has become very popular in the past couple of years, only appear at the very bottom of the list (Cannelli 2022). However, if these data are compared to the list of social media sites counting the highest number of users in the past couple of years, some discrepancies emerge between the platforms that are being archived and the ones that users across different countries engage in. In fact, the high number of users active on Meta Inc. platforms confirms the interest of most archiving initiatives in taking steps to preserve these sites (Statista.com 2023). Contrastingly, YouTube, which counted over 2.5 billion monthly active users as of October 2023, seems to be infrequently included in social media collections for various reasons. On the opposite side of the spectrum sits Twitter, which is largely archived in North American and European institutions, despite the number of users active on this platform being considerably lower than on other, more popular social sites (Statista.com 2023). These trends generate concerns regarding the representativeness of the social media cultural heritage that will be passed down to future generations, also highlighting the need to create positive conditions that could ease the barriers that prevent the development of social media archiving initiatives, especially in countries of the Global South (Colin-Arce et al. 2023).

Moreover, the combination of these imbalances consequently raises questions regarding the actual granularity of social media content collected on a national scale. In the following section, I will offer an overview of the factors that may affect the type of social media platforms archived as well

as the level of representativeness of national social media collections, examining restrictions set by legal frameworks, technical aspects, and available resources.

2. Factors influencing representativeness of social media collections

Web and social media archives play a central role in shaping the image that future generations will be able to remember and study about present events. A complex set of elements intervening in the development of national social media collections should be taken into consideration, as these have a profound impact on the overall structure, gaps, and narratives preserved.

The making of traditional archives involves a series of selection and appraisal procedures that can only be applied to a certain extent to social media, due to its unique characteristics. Reflecting on web archiving practices, Masanès et al. (2021) observed how "archiving the 'whole' Web is not attainable, due to resources and time limitations, as well as its de facto infinite generative nature." Preserving social media appears to some extent even more challenging than traditional websites, owing to its highly ephemeral, dynamic nature and the sheer volume of content generated each second on an ever-growing number of platforms. For this reason, instead of striving to achieve an impossible and unnecessary level of comprehensiveness, archiving institutions aim at providing the best representation possible of events and discussions on social sites (Masanès et al. 2021).

One aspect to consider when it comes to archived social media content is that many national archiving institutions rarely distinguish between websites and social media, creating collections that indiscriminately include both types of artifacts. While this is relatively justified by the fact that social media is indeed part of the web, it is undeniable that the latter has evolved into a separate phenomenon. Unlike websites that are collected through multiple approaches combining broad-scope, annual, and several selective crawls, only a rather small selection of social media accounts or hashtags is included in existing web collections, organized around specific themes or events, and often captured in the context of emergency collection campaigns to document unexpected crisis (Schafer et al. 2019). Although archival practices and collection development policies may vary between institutions, there are several factors that affect almost all social media archiving initiatives and may have a profound effect on the granularity of collections.

2.1 Legal constraints

National legal frameworks, including digital legal deposit legislation and policies established by social media companies to regulate the use and reuse of content shared on their own platforms, are among the legal constraints that may impact the degree of representativeness of social media collections.

National legal frameworks have a significant influence on the content preserved as part of social media collections at national memory institutions, especially for those operating under digital legal deposit legislation. In an overview of existing non-print legal deposit legislation offered in a report compiled by researchers involved in the BESOCIAL project (Chambers et al. 2021), it emerged how the minimum common denominator of most of these regulations is that they define born-digital content as that which is related to or published within national borders. On the one hand, while this criterion coincides with the sovereignty that a government possesses over affairs within a territorial or geographical area, on the other hand it fulfills the need to preserve the digital history and cultural heritage of specific countries. This parameter implicates, however, geographical boundaries that tend to blur in the context of the World Wide Web and particularly social media. Because of the international interconnectedness that characterizes the web and even more so social media interactions, it is extremely difficult for web archivists to disentangle billions of threads of discussions and ascertain with absolute certainty content provenance on social sites. In a recent article discussing the archival strategies implemented in the development of the UK Web Archive, Bingham and Byrne (2021) reported the uncertainty surrounding the process of identifying content on social media that originates on national soil. As they explained, establishing the boundaries of the national web domain for websites is facilitated to a certain extent by the selection of sites bearing domain extensions assigned to the national TLD. Conversely, social media platforms are mostly hosted on .com domains and thus located outside the country in scope (Bingham and Byrne 2021). Moreover, assessing provenance of content shared on social platforms can be laborious and not always reliable. For example, relying on geolocalization data available on these sites has proved to be a challenging task as geographical information can be subject to high error rates and inaccurate (Graham et al. 2014). For this reason, archiving institutions have mostly resolved to hand-picking accounts of organizations or public figures for which provenance or pertinence to the country in scope can be determined with confidence. Similar archiving approaches, however, contribute to the formation of inevitable gaps in the collections, which, in some cases, cannot be filled due

to the low persistence, high ephemerality, and constant evolution of social media content (Richardson 2021; Ringel and Davidson 2020).

Digital legal deposit provisions and data protection legislation place another layer of restrictions on selection criteria. In order to safeguard individuals' privacy, the collection of born-digital material under the governing law is usually limited to content that is made publicly available on social sites. Such limitation, however, often leads to the exclusion of content that might be in scope but accessible only upon authentication. Particularly affected by this is the capture of platforms like Facebook, where users tend to share information among a selected group of 'friends' or among closed groups of people (Sinn et al. 2013). Especially in the latter case, the constraints imposed by existing regulations, although necessary to protect the users' privacy, can lead to the formation of important gaps in the cultural heritage preserved. For example, displaced or marginalized groups appear inclined to share information and communicate with members of their own community within private Facebook groups (Goldsmith et al. 2022; Good 2012). In order to capture this content, archiving institutions would be required to log into the platform or be invited to join said groups, which might not be authorized by national digital legal deposit legislation. This clearly constitutes a problem in terms of representativeness of collections as many of these communities are often only present on social media and have no website that could be archived instead (Ferré-Pavia et al. 2018). Besides, as observed by web archivists at the Luxembourg Web Archive[2], there are some additional dilemmas that come into play when trying to preserve social media, such as problems with online discoverability of a variety of small realities spread across the national territory, or concerns emerging from public figures' accounts that share personal information alongside public communication (Schafer and Els 2020).

In this already complicated panorama, social media policies add another layer of legal constraints that heavily affect the granularity of information collected on their platforms, particularly concerning access to data. Social media companies impose strict limitations, for example, on the quantity and frequency with which information can be captured within a set time frame. That, coupled with other technical challenges, may explain in part why many institutions archive certain platforms (e.g. X-Twitter) more than others. As mentioned in the previous section, Facebook appears among the most archived social platforms. However, when looking closely at the amount of Facebook materials included in existing collections, it becomes evident how some institutions only archive a limited number of relevant

[2] Ben Els (National Library of Luxembourg), interviewed via Zoom by Beatrice Cannelli, 12 April 2022.

profiles on this platform. Due to the numerous restrictions set by Meta Inc. on harvesting, many institutions have seen their accounts periodically blocked when exceeding the set rate limit. Web archivists' reduced ability to regularly capture information without having to worry about accounts being suspended or restricted, consistently affects Facebook's preservation, and specifically impacts all those organizations, communities, and public figures active only on this site.

Furthermore, platform acquisitions from third parties can lead to changes in social media policies, making sites more difficult if not impossible to archive. For example, problems surrounding the capture of platforms like LinkedIn can be connected to the implementation of stricter rules for web crawling following Microsoft's acquisition in 2016. The LinkedIn User Agreement explicitly states that users are forbidden to scrape data from the site using third party software[3]. Although it is not among the most archived social platforms, LinkedIn is still relevant to some institutions such as the UK National Archives that are left unable to capture potentially relevant content for their UK Government Web Archive because of the restrictions in place[4]. Similarly, the most recent acquisition that has had a major impact on existing social media archives and whose repercussions are yet to be fully assessed, especially for institutions collecting through the platform's official application programming interfaces (APIs), is the one concerning X-Twitter. The takeover in October 2022 of Twitter by Tesla Inc. CEO, Elon Musk (Clayton & Hoskins 2022), led to a series of changes that culminated—from a social media archiving perspective—in the upheaval of the Twitter API access system known until that point. The new leadership decided to end free access to its APIs, including the much-praised Academic Twitter API, in favor of a paid tiers system that, to date, does not include any access specifically designed for research or preservation purposes. According to information made available on the X Developer Platform[5], the only tier available that provides free access to data via the Twitter API v2 comes with several limitations in terms of the total amount of data that can be retrieved per month; whereas the tier that offers the highest level of access, including a full-archive search, requires the payment of a fee that many archiving institutions with already limited budgets will not be able to sustain both in the short- and long-term.

---

[3] LinkedIn, Prohibited software and extensions: https://web.archive.org/web/20231116062416/https://www.linkedin.com/help/linkedin/answer/a1341387

[4] Claire Newing (The UK National Archives), interviewed via Zoom by Beatrice Cannelli, 30 June 2022.

[5] Further information about access to the Twitter API v2 can be found at the following link: https://web.archive.org/web/20231208180144/https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api#v2-access-leve

Representativeness of social media collections can also be influenced by technical challenges encountered while using different archiving techniques. As there are no standard approaches to collecting and preserving social media material, archiving initiatives use different methods to capture information on social sites. The decision of which method to adopt is often based on requirements (e.g. preservation in the context of legal deposit), resources, and expertise available at single institutions. These methods include the use of Application Programming Interfaces (APIs) to access data on social media sites and traditional web crawlers (e.g. Heritrix).

As clarified at the beginning of this chapter, most archiving institutions do not aim for an exhaustive archiving of content on social media, but rather a representative snapshot of the discussions and digital cultural heritage generated on these sites. However, the restrictions applied by social media platforms through policies and terms of use pose several technical challenges to the development of representative collections. Pehlivan et al. (2021) provided a comprehensive overview of the archival challenges related to data collection via APIs, discussing the restriction rules imposed specifically by Twitter. Among the different types of Twitter APIs described, the authors pointed out how the Sample API allowed institutions to collect 1% of all public tweets selected randomly in real time, which did not include historical tweets as it was not Twitter's intention for the Search API to focus on exhaustiveness but rather on relevance to the chosen keywords (Pehlivan et al. 2021). Although archiving institutions might not aim for completeness when it comes to social media collections, it is important for them to understand how sampling mechanisms work and how representative those random samples are of the whole data available, so that this can be accurately documented. Numerous studies have shed light on the biases existing in sampling mechanisms (González-Bailón et al. 2014; Tromble et al. 2017; Wu et al. 2020), but only a handful of them have taken into consideration potential repercussions on institutional archiving and long-term preservation (Acker and Kreisberg 2020; Littman et al. 2018; Pehlivan et al. 2021). The opaqueness of criteria concerning sampling and the changes applied to algorithms that can occur at any time without making API users aware, can influence the representativeness of content collected using this method (Hino and Fahey 2019). The main risk lies in portraying and preserving potentially unbalanced perspectives on historical events and culture in the long term. For example, this can be particularly problematic in the case of controversial topics or election campaigns where the amount of social media content has increased significantly in the past decade and for which it is essential to preserve the different opinions of all the parties involved.

Additional challenges to shaping representative social media collections arise from the use of archiving tools, such as web crawlers, which were often initially developed to capture traditional websites. The majority of institutions archiving the national web domain at scale, including the UK Web Archive (UKWA) and the National Library of France (BnF), use the Internet Archive's Heritrix crawler (Aubry 2010; Bingham and Byrne 2021). Despite a few technical issues with more dynamic sites using JavaScript, web crawlers like Heritrix have become a widely accepted method to successfully preserve a comprehensive snapshot of national TLDs (Brügger 2018). However, archiving social media platforms using these tools still poses many challenges. Most web crawlers struggle to correctly interact with the complex layout of social platforms and thus capture the highly dynamic content shared on social media, with institutions reporting gaps in the materials collected. As remarked by curators at the BnF[6], the inability of web crawlers to interact with elements on the page, such as buttons to expand hidden sections or scroll down pages to prompt the loading of more posts in the feed, can lead to the loss of relevant information from public profiles selected for their enduring cultural value that tend to share numerous posts daily. When harvesting social media content for the special collection dedicated to the COVID-19 pandemic, the BnF registered, overall, a higher success rate on Twitter compared to other platforms (Gebeil and Schafer 2020). According to the BnF, Facebook crawling scored instead a particularly low success rate that required additional efforts to obtain adequate captures, so much so that the BnF has decided to temporarily pause collection activities on this site until new, more sustainable archiving solutions are found (Gebeil and Schafer 2020). Indeed, the reduced quality and persistently unsuccessful outcomes obtained when archiving social platforms can ultimately result in institutions deciding to focus their time and resources on other, easier-to-archive sites. This means, however, that important evidence about contemporary events and culture will be lost, potentially increasing already existing biases and gaps.

In terms of technical challenges surrounding the capture of popular social media platforms, the Internet Archive's Archive-it Help Center page offers an interesting summary of known archiving issues using Heritrix. Among these, it is worth noting how Facebook and Instagram, both owned by Meta, are identified as platforms that hinder the capturing of many organizational profile pages and some Facebook Groups pages. As a result, this leads to the exclusion from the collection of countless relevant accounts, including small groups that only exist on these platforms. Besides, after Elon Musk

[6] Vladimir Tybin (National Library of France), interviewed by Beatrice Cannelli, Paris, 19–20 April 2022.

acquired Twitter in 2022 and implemented various changes from both technical and rebranding perspectives, the Archive-it Team updated the Twitter archiving status. They initially reported issues with harvesting some Twitter seeds[7] in March 2023, and then advised Heritrix users to pause archiving activities on the site as "recent changes to visibility of content on Twitter present multiple archiving challenges."[8] The combined legal, curatorial, and technical challenges generated by the complexity of social media platforms require archiving institutions to constantly adjust and find bespoke solutions to the latest variations in the field. In addition, institutions have to gauge the scale of preservation activities often based on the limited funds.

2.3 Resources and sustainability

Resources available play an important role in the development of representative collections of social media material and its long-term sustainability. Social media platforms and technologies are always shifting, requiring a substantial amount of resources to support the improvement and implementation of strategies and technologies that can successfully capture these platforms (Bingham and Byrne 2021). It is important to consider that many institutions preserving social media content are operating under legal deposit mandates that require them to archive, preserve, and provide access to this material. However, public archiving institutions are often developing social media collections with financial resources that do not always commeasure with the scale of the endeavor.

Apart from some exceptions, national web and social media archives are the result of small teams' efforts, which sometimes comprise only a few curators tasked with sifting through the sheer amount of information published on social sites, and carefully selecting profiles or hashtags that fit the scope of the collections. Moreover, a considerable share of resources necessarily flows into the technical side of social media archiving: developing ad hoc tools or implementing existing ones requires engineers or highly specialized technicians that institutions with limited budgets fail to attract as they struggle to offer competitive salaries. The alternative is either to outsource collecting activities to third parties or subscribe to external web archiving services (e.g. Archive-it) that offer a set of tools, training, and technical support for preserving and providing access to the archived data,

[7] Archive-it Help Center "Social media and other platforms status", archived on 25/03/2023 https://web.archive.org/web/20230325144024/https://support.archive-it.org/hc/en-us/articles/9897233696148-Social-media-and-other-platforms-status-

[8] Archive-it Help Center "Social media and other platforms status", archived on 02/01/2024 https://web.archive.org/web/20240102174300/https://support.archive-it.org/hc/en-us/articles/9897233696148-Social-media-and-other-platforms-status-

both of which can still be expensive. Nevertheless, because of the many challenges and constraints that social media archiving entails, selecting and collecting social sites is still largely a manual process, which requires time, curators, and specially dedicated resources. Moreover, curators occupied with handpicking content from social platforms are often also simultaneously working on selecting websites for ongoing web archive collections, further stretching the capacity of curators to singlehandedly ensure a well-balanced, broad-spectrum representation of the various strata of society.

The long-term sustainability of (representative) social media archiving collections is an open issue. While recent studies, such as the one conducted by the BESOCIAL project[9] at the Royal Library of Belgium (KBR), have shed light on opportunities for the development of sustainable social media archiving strategies (Messens et al. 2021), questions persist on how to sustainably tackle inclusivity and diversity concerns in the context of social media collections.

3. Mitigating representativeness concerns through participation practices

To mitigate inherent institutional biases and concerns about representativeness of social media collections, many archiving institutions have experimented and consolidated specific participatory archiving strategies designed to make the most of the limited budgets, staff, and time available (Pendergrass et al. 2019). Web and social media archiving initiatives have been successfully using participatory collection practices to record specific events or topics, seeking the contribution of the public or researchers that could bring their own unique perspective to the archive.

Web archives have increasingly adopted participation to expand the catchment area of the web-based material to be archived as part of national collections and help address known problems of representativeness (Cui et al. 2023; Schafer and Winters 2021). When it comes to participatory approaches, web and social media archives tend to turn to forms of collaborative curation and crowdsourcing. In this context, it is important to reiterate that, as social media are frequently included in existing web archive collections, distinctions between social sites and traditional websites are often minimal in the development of participatory practices or campaigns.

Popular types of crowdsourcing practices related to the appraisal and selection of valuable web materials include open calls for suggesting content to be incorporated in specific collections. Some web archives have

[9] Further information about the BESOCIAL project can be found here: https://web.archive.org/web/20240214181009/https://www.kbr.be/en/projects/besocial/

dedicated pages on their portals where individuals can fill in a form providing information and the URL of websites they would like to nominate for preservation, such as the "Save a UK website" feature available on the UK Web Archive portal. However, most of these forms appear to be structured and formulated in favor of submitting website URLs rather than social media content, likely due to legal and ethical concerns surrounding the latter. Nevertheless, recent campaigns promoted between 2020 and 2021 in light of the COVID-19 outbreak have encouraged members of the public to nominate meaningful hashtags and social media content alongside traditional websites. Moreover, curators at the Luxembourg Web Archive[10] have noted that suggestions received through the campaign they launched at the beginning of the first lockdown helped them uncover small religious groups that were particularly active in disseminating official information among their members. These groups had not been included in their national web collection before (Schafer and Els 2020; Schafer and Winters 2021). Besides, institutions undertaking pilot or short-term projects to test the feasibility and sustainability of social media archiving have found in these practices a means to discover new themes and areas of interest to integrate the initial 'top down' approach. The BESOCIAL project at KBR[11], for example, decided to test different approaches including a crowdsourcing campaign to ask the public to nominate social media material (including text-based material, hashtags, and public accounts) that should be preserved as part of the online national heritage collection they were developing. The BESOCIAL team not only received hundreds of responses helping them fill in the gaps and mitigate representativeness concerns that emerged from the initial archiving approach, but also observed how the campaign supported the promotion of social media archiving activities in Belgium. Certainly, the effectiveness of such campaigns in terms of increasing representativeness of social media collections is linked to how they are promoted and among which communities. A meticulous dissemination strategy among specific target groups is indeed essential for obtaining contributions that can truly enrich the content already being preserved. While public involvement in these campaigns may vary, each individual URL can still be crucial for uncovering underrepresented themes or marginalized communities.

At some institutions, the selection of born-digital content to be added to the archive is the result of the combined effort of web and social media curators as well as a network of contributors identified both within and outside the cultural heritage institution. This is exemplified by the system established at the National Library of France (BnF) 12, where curators of

[10] Ben Els, interview.
[11] Fien Messens and Friedel Geeraert (Royal Library of Belgium), interviewed by Beatrice Cannelli, 13 July 2022.
[12] BnF, *Cooperer autour de l'archivage du Web*:

the digital legal deposit team, contributors from other BnF departments, and associated regional centers (e.g. regional archives, libraries, and research institutes) have been collaborating to support the capture of a diverse representation of the French territory and society. To facilitate the process and management of web and social media materials to be collected, the BnF has also developed an application called "BnF Collecte du web[13]"(BCweb), that allows contributors to independently perform actions such as entering, modifying, or deactivating URLs in the seed list.

Similarly, other institutions have invited collaboration or established partnerships with researchers who are both qualified information professionals and representatives of certain minority groups. Researchers-curators are often sought for their participation in the development of thematic and special collections focusing on capturing the many-sided reality of small communities on the national territory. For example, in the UK Web Archive[14] several thematic collections have been created through participatory curation practices, including the "Black and Asian Britain", "French in London" and the "LGTBQ+ lives online". These participatory practices aim to bring a diverse range of material into the web and social media archive, helping preserve communities' own viewpoints, experiences, and stories.

However, even in the context of co-curation practices, these collections might still face criticism due to certain curation choices. While curatorial decisions are made in collaboration with the archiving institution, documenting episodes of hate, discrimination, and violence can raise concerns among community members, despite these unfortunate occurrences often being integral parts of minority groups' lives. Nevertheless, constructive discussions between the parties involved can still lead to positive outcomes, such as the production of extensive collection descriptions featuring any potentially controversial aspects and content warnings, which are of great value to both the institution and users.

Conclusion

Social media has radically changed the way individuals interact and communicate online, offering unique insights into contemporary events and providing environments for minority groups to self-represent. While the inclusion of social media content of enduring value in national cultural

---

https://web.archive.org/web/20240107200657/https://www.bnf.fr/fr/cooperer-autour-de-larchivage-du-web#bnf-un-r-seau-de-partenaires-pour-encourager-les-recherches-sur-les-archives-du-web

[13] BnF, Collecte du web homepage: https://web.archive.org/web/20231208191553/https://collecteweb.bnf.fr/login

[14] Nicola Bingham (British Library), interviewed via Zoom by Beatrice Cannelli, 23 June 2022.

heritage preservation strategies is on the rise, numerous unsolved archiving challenges persist, prompting questions about representation and inclusivity.

In this chapter, I have provided an overview of the manifold limitations influencing the representativeness of social media collections. I began by considering the social media archiving landscape and how this is shaped by dynamics ascribable to geopolitical and social media divides. Following that, I have described how legal frameworks, technical matters, social media policies and their ephemerality can deeply affect the degree of representativeness of social media collections at a national level, including the type and rate with which different platforms are collected.

I have illustrated how, in order to mitigate representativeness concerns, many social media archiving institutions have adopted specific curatorial strategies that seek the participation of researchers, networks of contributors, and the wider public. Engaging with the public and external contributors has proved to be a valuable approach to uncover stories from underrepresented communities that might escape the large links of the net used by archiving institutions to sift through content in scope. The impact of these participatory practices on the overall enhancement of the representation of national social media collections, and especially their sustainability in the long term, still needs to be fully assessed. In the meantime, documenting how these participatory collections have been developed and making collection scoping documents publicly available or upon request would be a significant step towards helping researchers fully understand the potential implications of such curatorial processes.

Nevertheless, the participatory practices described in this chapter present a good opportunity to raise awareness of the significance of social media archives among the wider public, also contributing to continual engagement with these collections. Encouraging members from different groups of society to actively participate in developing national social media archives, can truly support the preservation of the multifaceted impact these communities have on the national cultural landscape, letting them tell their stories—through content they suggested—using their own voices.

References

Acker, Amelia, and Adam Kreisberg. 2020. “Social media data archives in an API-driven world.” *Archival Science* 20 (2): 106–123. https://doi.org/10.1007/s10502-019-09325-9.

Aubry, Sara. 2010. “Introducing web archives as a new library service: The experience of the national library of France.” *Liber Quarterly* 20 (2). https://liberquarterly.eu/article/view/10584/11316.

Barrowcliffe, Rose. 2021. “Closing the narrative gap: Social media as a tool to reconcile institutional archival narratives with Indigenous counter-narratives.” *Archives and Manuscripts* 49 (3), Article 3. https://doi.org/10.1080/01576895.2021.1883074.

Ben-David, Anat. 2021. “Critical web archive research.” In *The Past Web: Exploring Web Archives*, 181–188. Springer. https://doi.org/10.1007/978-3-030-63291-5.

Bergis, J., Summers, E., and Mitchell, V. J. 2018. “Documenting the Now White Paper: Ethical Considerations for Archiving Social Media Content Generated by Contemporary Social Movements: Challenges, Opportunities, and Recommendations. Documenting the Now, Documenting the Now.”

Bingham, Nicola, and Helena Byrne. 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.

Brügger, Niels. 2018. *The Archived Web: Doing History in the Digital Age*. Cambridge: MIT Press.

Bruns, Axel, and Katrin, Weller. 2016. “Twitter as a first draft of the present: And the challenges of preserving it for the future.” *Proceedings of the 8th ACM Conference on Web Science*: 183–189. https://doi.org/10.1145/2908131.2908174

Cannelli, Beatrice. 2022. “Mapping social media archiving initiatives: State of the art, trends, and future perspectives.” IIPC Net Preserve Blog. https://netpreserveblog.wordpress.com/2022/11/30/mapping-social-media-archiving-initiatives-state-of-the-art-trends-and-future-perspectives/

Caswell, M., Migoni, A. A., Geraci, N., and Cifor, M. 2017. “‘To be able to imagine otherwise’: Community archives and the importance of representation.” *Archives and Records* 38 (1). https://doi.org/10.1080/23257962.2016.1260445.

Chambers, S., Birkholz, J., Geeraert, F., Pranger, J., Messens, F., Lieber, S., Mechant, P., Michel, A., and Vlassenroot, E. 2021. “BESOCIAL: final report WorkPackage1 an international review of social media archiving initiatives.” 91. https://www.kbr.be/wp-content/uploads/2020/07/202012_BESOCIAL_Report_WP1_Review_of_existing_social_media_archiving_projects.pdf

Clayton, J., & Hoskins, P. 2022. “Elon Musk takes control of Twitter in $44bn deal.” BBC News. October 28. https://www.bbc.com/news/technology-63402338

Colin-Arce, A., Fernández-Quintanilla, S., Benítez-Pérez, V., & García-Monroy, A. 2023. “Web Archiving en español: Barriers to Accessing and Using Web Archives in Latin America.” https://www.youtube.com/watch?v=plQURfARGBc

Cui, C., Pinfield, S., Cox, A., & Hopfgartner, F. 2023. “Participatory Web Archiving: Multifaceted Challenges.” In *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity,* edited by I. Sserwanga, A. Goulding, H. Moulaison-Sandy, J. T. Du, A. L. Soares, V. Hessami, and R. D. Frank, 79–87. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28035-1_7

Ferré-Pavia, C., Zabaleta, I., Gutierrez, A., Fernandez-Astobiza, I., & Xamardo, N. 2018. “Internet and social media in European minority languages: Analysis of the digitalization process.” *International Journal of Communication* 12 (22). Available at: https://ijoc.org/index.php/ijoc/article/view/7464

Fondren, E. & Menard McCune, M. 2018. “Archiving and Preserving Social Media at the Library of Congress: Institutional and Cultural Challenges to Build a Twitter Archive.” *Preservation, Digital Technology & Culture* 47(2). https://doi.org/10.1515/pdtc-2018-0011.

Gebeil, Sophie and Valérie Schafer. 2020. “Exploring special web archives collections related to COVID-19: The case of the French National Library (BnF).” *WARCnet Papers*.

Goldsmith, L. P., Rowland-Pomp, M., Hanson, K., Deal, A., Crawshaw, A. F., Hayward, S. E., Knights, F., Carter, J., Ahmad, A., Razai, M., Vandrevala, T., and Hargreaves, S. 2022. “Use of social media platforms by migrant and ethnic minority populations during the COVID-19 pandemic: A systematic review.” *BMJ Open* 12 (11). https://doi.org/10.1136/bmjopen-2022-061896.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., and Moreno, Y. 2014. “Assessing the bias in samples of large online networks.” *Social Networks* 38: 16–27. https://doi.org/10.1016/j.socnet.2014.01.004.

Good, K. D. 2012. “From scrapbook to Facebook: A history of personal media assemblage and archives.” *New Media & Society*: 15 (4). https://doi.org/10.1177/1461444812458432.

Graham, M., Hale, S. A., and Gaffney, D. 2014. “Where in the world are you? Geolocation and language identification in Twitter.” *The Professional Geographer* 66 (4): 568–578. https://doi.org/10.1080/00330124.2014.907699.

Harris, Verne. 2002. “The Archival Sliver: Power, Memory, and Archives in South Africa.” *Archival Science* 2, no. 1–2: 63–86. https://doi.org/10.1007/BF02435631.

Hegarty, Kieran. 2022. “The Invention of the Archived Web: Tracing the Influence of Library Frameworks on Web Archiving Infrastructure.” *Internet Histories* 6 (4): 432–51. https://doi.org/10.1080/24701475.2022.2103988.

Henninger, Maureen, and Paul Scifleet. 2016. “How Are the New Documents of Social Networks Shaping Our Cultural Memory.” *Journal of Documentation* 72 (2): 277–98. https://doi.org/10.1108/JD-06-2015-0069.

Hino, Airo, and Robert A. Fahey. 2019. “Representing the Twittersphere: Archiving a Representative Sample of Twitter Data under Resource Constraints.” *International Journal of Information Management* 48 (October): 175–84. https://doi.org/10.1016/j.ijinfomgt.2019.01.019.

Jackson, T. 2020. “‘I’ve never told anybody that before’” In *Communities, Archives and New Collaborative Practices*, edited by S. Popple, A. Prescott, and D. H. Mutibwa, (1st ed., 93–106). Bristol University Press. https://doi.org/10.2307/j.ctvx1hvvd.13.

Jimerson, Randall. 2006. “Embracing the Power of Archives.” *The American Archivist* 69

(1): 19–32. https://doi.org/10.17723/aarc.69.1.r0p75n2084055418.

Littman, Justin, Daniel Chudnov, Daniel Kerchner, Christie Peterson, Yecheng Tan, Rachel Trent, Rajat Vij, and Laura Wrubel. 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.

Lutz, C. 2022. "Inequalities in social media use and their implications for digital methods research." *The SAGE Handbook of Social Media Research Methods*: 679–690.

Maemura, Emily. 2023. "Sorting URLs out: Seeing the Web through Infrastructural Inversion of Archival Crawling." *Internet Histories* 7 (4): 386–401. https://doi.org/10.1080/24701475.2023.2258697.

Masanès, Julien, Daniela Major, and Daniel Gomes. 2021. "The Past Web: A Look into the Future." In *The Past Web: Exploring Web Archives*, edited by Daniel Gomes, Elena Demidova, Jane Winters, and Thomas Risse, 285–91. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-63291-5_22.

Messens, F., Birkholz, J. M., Chambers, S., Geeraert, F., Michel, A., Mechant, P., Vlassenroot, E., Lieber, S., Dimou, A., and Watrin, P. 2021. "BESOCIAL–Towards a sustainable strategy for archiving and preserving social media in Belgium." Digital Humanities Benelux 2021 Conference.

Pehlivan, Z., Thièvre, J., & Drugeon, T. 2021. "Archiving Social Media: The Case of Twitter." In *The Past Web: Exploring Web Archives,* edited by D. Gomes, E. Demidova, J. Winters, and T. Risse, 43–56. Springer International Publishing. https://doi.org/10.1007/978-3-030-63291-5_5.

Pendergrass, Keith L., Walker Sampson, Tim Walsh, and Laura Alagna. 2019. "Toward Environmentally Sustainable Digital Preservation." *The American Archivist* 82 (1): 165–206. https://doi.org/10.17723/0360-9081-82.1.165.

Pietrobruno, S. 2013. "YouTube and the social archiving of intangible heritage." *New Media & Society* 15 (8), Article 8. https://doi.org/10.1177/1461444812469598.

Richardson, Allissa V. 2020. "The Coming Archival Crisis: How Ephemeral Video Disappears Protest Journalism and Threatens Newsreels of Tomorrow." *Digital Journalism* 8 (10): 1338–46. https://doi.org/10.1080/21670811.2020.1841568.

Ringel, Sharon, and Roei Davidson. 2022. "Proactive Ephemerality: How Journalists Use Automated and Manual Tweet Deletion to Minimize Risk and Its Consequences for Social Media as a Public Archive." *New Media & Society* 24 (5): 1216–33. https://doi.org/10.1177/1461444820972389.

Schafer, V., and Els, B. 2020. "Exploring special web archive collections related to COVID-19: The case of the BnL." *WARCnet Papers*.

Schafer, Valérie, Gérôme Truc, Romain Badouard, Lucien Castex, and Francesca Musiani. 2019. "Paris and Nice Terrorist Attacks: Exploring Twitter and Web Archives." *Media, War & Conflict* 12 (2): 153–70. https://doi.org/10.1177/1750635219839382.

Schafer, Valérie, and Jane Winters. 2021. "The Values of Web Archives." *International Journal of Digital Humanities* 2 (1–3): 129–44. https://doi.org/10.1007/s42803-021-00037-0.

Schwartz, Joan M., and Terry Cook. 2002. "Archives, Records, and Power: The Making of Modern Memory." *Archival Science* 2 (1–2): 1–19. https://doi.org/10.1007/BF02435628.

Simon, R. I. 2012. "Remembering together." In *Heritage and Social Media: Understanding Heritage in a Participatory Culture*, 89–106. Routledge.

Sinn, Donghee, and Sue Yeon Syn. 2014. "Personal Documentation on a Social Network Site: Facebook, a Collection of Moments from Your Life?" *Archival Science* 14 (2): 95–124. https://doi.org/10.1007/s10502-013-9208-7.

Statista.com. 2023. "Monthly Active Users by Social Media Platform (in millions)."

https://web.archive.org/web/20231210153436/https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Storrar, T. 2014. "Archiving social media." May, 8. *The National Archives Blog*. https://blog.nationalarchives.gov.uk/archiving-social-media/

Thomson, S. D. 2016. "Preserving Social Media (16–01; DPC Technology Watch Report)." https://www.dpconline.org/docs/technology-watch-reports/1486-twr16-01/file

Tromble, Rebekah, Andreas Storz, and Daniela Stockmann. 2017. "We Don't Know What We Don't Know: When and How the Use of Twitter's Public APIs Biases Scientific Inference." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3079927.

Van Dijck, José. 2011. "Flickr and the Culture of Connectivity: Sharing Views, Experiences, Memories." *Memory Studies* 4 (4): 401–15. https://doi.org/10.1177/1750698010385215.

Wu, Siqi, Marian-Andrei Rizoiu, and Lexing Xie. 2020. "Variation across Scales: Measurement Fidelity under Twitter Data Sampling." *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May): 715–25. https://doi.org/10.1609/icwsm.v14i1.7337.

Yakel, Elizabeth. 2003. "Archival Representation." *Archival Science* 3 (1): 1–25. https://doi.org/10.1007/BF02438926.

SECTION 3
The web as heritage in the making: debates and challenges

Challenges in archiving the personalized web

Erwan Le Merrer, Camilla Penzo, Gilles Tredan, Lucas Verney

Abstract: The decision-making algorithms embedded within online platforms are determining content shown to users. This personalization steers the dissemination of information, in contrast with the idea of a universal World Wide Web. Personalization thus generates a combinatorial explosion of different versions of the web, rendering each user's experience distinct. This raises critical questions: what elements of a personalized web should be archived? How can the collected user journeys capture a representative picture of our times? Navigating personalization is essential to capture the contemporary web experience, yet it presents methodological and technical challenges. In this chapter, we identify key challenges in performing a representative sampling of personalization within online platforms.

Keywords: personalization, archival, YouTube, 2022 French presidential election.

1. Introduction

The web has evolved from its static origins to a dynamic landscape where each user encounters an ever-changing and algorithmically tailored version. A few years ago, studies showed that people were generally unaware of the existence of algorithmic personalization (Eslami et al. 2015; Powers 2017). More recent studies (Schmidt et al. 2019; Eg, Demirkol Tønnesen, and Tennfjord 2023), however, suggest that users may grasp the notion that online content is filtered or that recommendations are based on their profiles, even if they are not necessarily familiar with algorithmic processes. Nevertheless, details regarding the algorithmic personalization of each platform remain undisclosed to both users and regulators. Recommendation algorithms, despite their critical role in selecting and ranking information, can inadvertently reinforce popularity as self-fulfilling prophecies (Salganik and Watts 2008). Furthermore, these algorithms often overlook the verification of information sources, potentially leading to the propagation of disinformation and the creation of filter bubbles.

Given that personalization inherently renders each user experience unique, collecting the entirety of the internet might offer limited insights into user experiences and journeys on online platforms, see e.g. "The Christchurch Call to Action to Eliminate Terrorist and Violent Extremist Content Online". Consequently, archiving user journeys amid algorithmic decisions becomes essential to understand individual and group dynamics, addressing a salient need in multiple contexts, such as e-commerce, web search, and social media (Schafer, Truc, and Badouard 2019).

While some recent approaches proposed means for users to collect their personal web experience (Kiesel et al. 2018), this chapter focuses on the challenges arising from the need of global and systematic archival means, with a specific focus on a concrete use case, the

Erwan Le Merrer, CNRS, France, erwan.le-merrer@inria.fr, 0000-0001-8344-2135
Camilla Penzo, PEReN, France, camilla.penzo@finances.gouv.fr
Gilles Tredan, CNRS, France, gtredan@laas.fr, 0000-0003-4473-4332
Lucas Verney, PEReN, France, lucas.verney@finances.gouv.fr, 0000-0002-1361-1703

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Erwan Le Merrer, Camilla Penzo, Gilles Tredan, Lucas Verney, *Challenges in archiving the personalized web*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.10, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 79-94, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

YouTube recommender (Covington, Adams, and Sargin 2016). We highlight and contemplate the complex interplay of methodological and technical decisions required to collect a personalized web. Emphasizing the combinatorial explosion of different web versions—each tailored to a specific user profile—we underscore the unobservable nature of these variations. Dealing with the personalization of the web is necessary to accurately capture the user experience surfing the contemporary web, but it also raises several methodological and technical challenges.

1.1 A computer scientist take on archiving personalization

This chapter reflects our position as researchers actively engaged in the technical aspects of auditing online platforms. This nascent research field is at the crossroads of several computer science fields, such as information retrieval, data science and security by certain aspects. As such, our position inherently carries a technical bias that we humbly endeavor to overcome in the development of this chapter. We believe that the outcomes of our (technical) experience, navigating the intricacies of personalization layers omnipresent on major platforms, have implications that reach beyond the realms of auditing and our technical expertise.

Defining platform personalization is a straightforward task; it involves tailoring the content suggested to users based on their past behavior and (estimated) preferences. However, it is important to recognize that personalization encompasses various practices.

We can distinguish between coarse and fine-grained personalization. An example of coarse-grained personalization is the automatic selection of the user interface language based on their inferred location (e.g. displaying an interface in French to users with a French IP address). Coarse-grained personalization is a broad approach that uniformly impacts large sets of users. Primarily, such personalization influences how contents are displayed on the interface rather than the selection of displayed contents. In contrast, fine-grain personalization aims to predict which content will likely appeal to each user. An illustration of this is Twitter's algorithmic Timeline (Bandy and Diakopoulos 2021). Implementing this type of personalization requires a sophisticated platform mechanism. Firstly, the platform observes a user's reactions to specific content, such as monitoring where the user's mouse hovers or tracking which videos were watched entirely versus those that were quickly discarded. These observations are then stored and the platform transforms them into criteria for selecting the most relevant content to present next. Throughout this chapter, we will use the term *user profile* to denote the information the platform possesses about a given user.

A closely related, yet distinct concept is *contextual recommendation*. Contextual recommendation selects items to present to a user based on the item currently being 'consumed', rather than depending on the user's past item consumption. The conceptual difference is fundamental, and aligns with a valuable mathematical abstraction, as contextual recommendation adheres to a Markovian model where the future is independent of the past, given the present. In practice, however, observing this distinction proves challenging. Modern platforms typically generate a set of 'hybrid' recommendations that rely on both the context (the current item) and the user's past history (Le Merrer and Tredan 2018). The thin line separating the two becomes even more blurred when

considering that contextual recommendations are (often) computed using techniques like collaborative filtering (exemplified by phrases such as "users watching X also watch Y"), which rely on users' watch history to assess content similarity. Consequently, in platforms using hybrid recommendations, a user's history may contain items originating from both past contextual and personalized recommendations, establishing a mutual induction between the two recommendation types.

1.2 A combinatorial explosion of versions of the web

Consider a hypothetical platform offering 100 items. To collect contextual recommendations for each item by visiting its page, one would need 100 visits. Now consider the platform incorporating personalization, where visitors receive recommendations based on the two last visited items. In this scenario, observing recommendations associated with all items, one would need to make a staggering 10,000 visits. If the platform uses the last five item visits to compute recommendations for a given item, an exhaustive observation would require 10 billion visits. What was once a modest website transforms into an intricate personalization labyrinth.

This rough estimation underscores two fundamental and technical challenges inherent in archiving a personalized web. The first challenge is evident: the exponential number of visits required for exhaustive exploration renders such thoroughness *practically unattainable*. The second challenge, more nuanced, involves the need for *certain assumptions* about the internals of the recommendation system to conduct such analysis (e.g. the number of previously visited items influencing the user recommendation for the currently visited item).

While these challenges are technical in origin, we contend that their resolution cannot be solely technical. Archiving, and especially web archiving, grapples with the difficult questions of archive curation and selection (Milligan, Ruest, and Lin 2016).

1.3 Data collection setup and terminology

We consider the conventional operational framework for web archiving, wherein robots (hereafter *bots*) gather data from the public web pages of the targeted website. The term *platform* denotes an online website hosting one or more *algorithms* or models with which our bots interact. An example of such models is YouTube's recommendation algorithm, responsible for personalizing video content for users. In our scenario, we assume a lack of agreement with the observed platform, implying that no application programming interface (API) is accessible for data collection, nor for collecting users profiles or the recommendations they receive. While we will discuss alternative approaches, throughout this chapter we will refer to the use of bots for the extraction of data from online platforms. These bots are programs in the form of scripts designed to automate specific data extraction tasks, such as emulating a user on a platform to access and extract the personalization proposed to that user.

1.4 When personalization becomes profiling

While emphasizing the need to archive personalization in today's web, it becomes

imperative to discuss how personalization has now evolved into a much more intrusive practice referred to as *user profiling*. The ubiquity and popularity of mobile versions of online platforms have become increasingly pronounced in our daily lives. The vast majority of mobile users install apps aligned with their interests, needs, and daily routines, facilitating a highly refined personalization process. Companies are now capitalizing on their ability to accurately profile mobile users, see e.g. (Farseev et al. 2020), asserting that they enhance user experiences or make lifestyle improvements, when in reality, their primary objective is finely tailored advertisements. User profiling through mobile applications involves a chain of processes, starting with the analysis of user data collected through the application. This analysis exploits correlations between the application's usage patterns and the user's personality traits, reaching a point where the platform or mobile app producer can predict the user's most personal characteristics (Gustarini et al. 2016; Xu et al. 2016). Effective profiling can transcend demographics, personal interests and lifestyles can be inferred, delving into personality traits and psychological states (Zhao et al. 2019) as well. With psychological profiling data, influencing user behavior, whether through product sales or other actions, becomes very effective.

To complete the profiling paradigm, data is now considered the new gold, bought and sold by entities with novel business models (data brokers), e.g. (Andrés, Azcoitia, and Laoutaris 2022). These entities connect heterogeneous data from multiple sources to maximize their predictive power and, consequently, their economic value. Clearly, this emerging trend raises concerns about user privacy, but has for the moment only prompted a limited response from civil society (Exodus Privacy; NOYB European Center for Digital Rights 2023). We argue that, in the pursuit of archiving the contemporary web experience, both personalization and profiling should find a space in the records for historians.

2. Motivation: Example on YouTube

We now turn our attention to the study of YouTube's personalization, driven by the fact that 71% of U.S. teenagers reportedly consult YouTube daily ("Teens, Social Media and Technology", 2023).

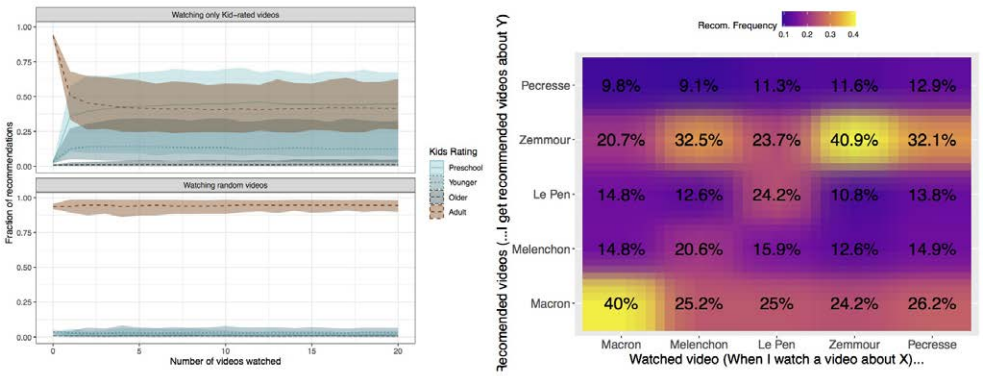
2.1 Measuring personalization using YouTube Kids

"YouTube is one of the largest scale and most sophisticated industrial recommendation systems" (Covington, Adams, and Sargin 2016), a system that has recently been at the center of recent controversies (Ledwich and Zaitsev 2020) due to its potential societal impact. We conducted a bot-driven study of personalization on YouTube (Le Merrer, Tredan, and Yesilkanat 2023), specifically focusing on measuring its consequences, with an application in the context of children recommendations. Notably, we discovered that the video identifiers were consistent across YouTube and YouTube Kids, the latter being a platform tailored for young users under the age of 13. We could thus automatically identify videos labeled as "for kids" on YouTube, i.e. those videos that appear on the platform YouTube Kids, providing a quantitative characterization of the effects of personalization. We used bots with two distinct behaviors: 'Control' bots that start with no profile and watch random videos from YouTube's personalized homepage, and 'kid' bots, that also

start with no history but exclusively watch ‘kid’ videos.

Figure 1 (left panel) presents the evolution of recommendations collected by each bot based on its behavior on the platform, as a function of the number of previously watched videos. Control bots consistently encounter a vast majority of adult (i.e. ‘non-kid’) videos (approximately 97%), whereas kid bots quickly trigger personalization, causing a shift in the recommended content mix towards a majority of kid videos. This significant change occurs within the first three watched videos and stabilizes after the fifth video.

Figure 1. (left) Composition of video recommendations (by age type) based on the number of previously watched videos, starting from empty profiles, for two video consumption profiles. (right) Candidate recommendation matrix for the 5 main candidates during the French 2022 presidential election.



In an archival context, we believe this experience carries two takeaways. Firstly, it is possible to generate and observe personalization using bots. Secondly, ‘control’ bots and ‘kid’ bots exhibit profoundly different encounters with the same YouTube platform after watching a few (< 5) videos. The contrasting experiences of kids and adults could be seen as a stark example of a filter bubble, a concept introduced by Eli Pariser (Pariser 2011). However, these filter bubbles likely lead every user into his or her own subjective journey on YouTube.

2.2 Collecting personalization during the French presidential campaign

Using the same setup as in (Le Merrer, Tredan, and Yesilkanat 2023), we gathered personalization data related to the French 2022 presidential election. While a comprehensive analysis of the collected material goes beyond scope of this chapter, we aim to present a perspective that sheds light on the methodological challenges posed by personalization.

Figure 1 (right panel) illustrates the response of personalization to bots with no prior history, and that randomly watch videos of the ‘French news’ YouTube page (from February 1 to April 10 during the first election round). Approximately 180 times a day, a bot with no profile watches five random videos consecutively from YouTube recommendations on the news page. We observe the titles of all the watched and subsequently suggested videos, considering a video to be about a specific candidate if their name or the name of their political party appears in the title. Titles lacking any mention of a candidate are disregarded. The figure demonstrates what happens when a bot inadvertently watches a video about candidate X: which videos are then recommended to the bot? To exemplify: when I watch a video about candidate Mélenchon, 32.5% of the recommendations (related to any considered candidate) are about candidate Zemmour.

We believe Figure 1 (right panel) vividly illustrates the challenges posed by personalization: after watching a single video about candidate Macron, users receive approximately twice as many recommendations about Macron compared to Zemmour, and vice-versa. Consequently, these users encounter different personalized recommendations, leading to a divergent perspective of the French political landscape on the platform.

Before presenting the conclusions drawn from this observation, it is important to acknowledge the limitations of our approach. Notably, we assign videos to a politician in a straightforward manner (based on name presence in the title) and our quantitative analysis does not delve into the semantics within each mention of a candidate. Hence, a video criticizing a candidate is treated equivalently as a video endorsing them, despite the potential substantial differences in their impact on personalization. Producing such an aggregate figure necessitates compressing millions of recommendations—referring to complex media objects— collected over more than two months into a 5×5 color matrix: it is necessarily partial (incomplete) and potentially biased. Moreover, one may argue that our bot behaviors are overly simplistic and fail to represent actual user experiences (long watch histories and diverse interests, etc.). This objection represents a central challenge that we will address in the subsequent discussion.

In the context of the French election, a mandate regulates the equal division of speaking times among the candidates in the traditional broadcasting media, starting 15 days before the election. This rule, overseen by an independent institution (Arcom), aims to foster fair competition among candidates, implementing equality at the producer level. However, transposing this rule to a personalized platform such as YouTube presents two significant challenges. Firstly, while media operate within well-defined categories requiring licenses, anyone can be a content producer on YouTube. Consequently, binding every content producer to a national rule appears difficult. The second challenge arises when aiming for equality at the receiver level. Since personalization tailors the experience on YouTube for each individual user, assessing an “average" speaking time is nearly impossible.

Elections hold great significance in the political life of democratic countries and arguably possess considerable historical value. However, a clear rule like ‘equal speaking time’ becomes nebulous when applied to personalized platforms. We contend that the same complexity applies to archival policies during elections: selecting content for archiving to provide an accurate retrospective view for historians in our contemporary times requires handling the personalization layer through which we observe online platforms.

3. Challenges

In this section, we present a structured overview of the main challenges encountered in the archiving of journeys on a personalized platform. At a broad level, three classes of challenges emerge: technical challenges arising from the algorithmic nature of the media platform, methodological challenges pertaining to the archivist’s selection of methods for constructing the archival fonds, and usability challenges focused on strategies enabling effective exploration of the archived fonds by future users. In essence, to archive personalization successfully, three fundamental questions must be addressed: how to collect personalization (technical), which aspects of personalization to collect (methodological), and how to present the collected personalization (usability). These questions are interdependent and mutually influence each other. For instance, technical limitations in observing all individual personalizations necessitate methodological choices, and these choices subsequently impact how the archive is then presented to users (usability). Given the interdependence of these issues, we advocate for the integration and juxtaposition of diverse disciplinary perspectives, such as computer science, history, and usability, to construct coherent solutions that facilitate accurate future analyses of our contemporary personalized experiences.

3.1 Technical Challenges

3.1.1 Platform opacity

Modern recommenders leverage a multitude of features, often numbering in the hundreds, to personalize users' experiences (Covington, Adams, and Sargin 2016). These include user-related data, such as demographics and consumption habits. While general techniques for implementing recommenders are publicly available (Gupta et al. 2020), the specific features employed by a corporate recommender in production are typically kept secret. Consequently, understanding the bot-simulated features that influence

personalization becomes a speculative endeavor for the programmer/archivist.

The inability to know and interact with every possible user feature used in the recommendation algorithm places the archivist in a *black-box* interaction scenario with the algorithm. Judging the impact of a particular feature on the resulting personalization necessitates tedious trial and error as illustrated in the previous section.

Despite the technical impossibility of representing every detail of a real user, we believe that the coarse-grain traits of simulated user profiles, precisely defined in the subsequent sections, can yield valuable insights. Simulated user profiles aim to represent user profiles of interest in a coarse manner, rather than ultra specific ones as exemplified by Mozilla's approach to highlight the existence of online personalization (Mozilla 2020): with 'TheirTube', they showcases ultra-coarse profiles such as 'liberal' or 'climate denier', asserting that their watch history encapsulates these personas and thus large user categories.

In order to craft synthetic yet more relevant profiles, discussions with sociologists and statisticians become crucial in crafting representative sets of personas, which can then be presented to algorithms through bots. Identified classes of people experiencing discrimination are also vital to extract personalization for further research into potential bias.

We note that, following the *Digital Services Act, 2022*, research endeavors have begun questioning the possibility of inferring which features impact algorithmic decisions (Rastegarpanah, Gummadi, and Crovella 2021), with the aim of exposing objectionable behaviors.

3.1.2 Frugality in load-responsible and non-interfering data extraction

When crawling a static website for archival purposes, the resulting server load is proportional to the disk space required to implement the website. The scenario shifts significantly, however, when dealing with dynamic websites using personalization, as they are designed to generate, filter, or sort vast amounts of content tailored to users with the aim of prolonging their stay on the website (Covington, Adams, and Sargin 2016). Consequently, crawls and data collections can be virtually endless, allowing bots to navigate through an intricate maze of personalized content. For this reason, frugality becomes a crucial practical consideration in the collection process, ensuring not only a respectful interaction with the platform infrastructure, but also for extracting a manageable amount of data for archival purposes.

Drawing a parallel with the general principle of minimal interaction with an object of study *in vivo*, we emphasize the necessity for frugal extraction.

Avoid loading platform infrastructures

Personalization on platforms has evolved to rely predominantly on complex machine learning models (Covington, Adams, and Sargin 2016). Consequently, engaging with these platforms entails compute-intensive processes in comparison to their static counterparts. When using bots for measurements and data extractions, it is imperative to consider the resulting load on the platforms to ensure responsible operation. Specifically, interactions should not disrupt the platform's service by employing overly heavy machinery to achieve collection objectives.

To avoid such disruptions, platforms commonly adopt defensive measures as rate-limiting mechanisms (Cloudfare 2023). Data extraction must accordingly account for these considerations by estimating what is tolerable for the platform.

Avoid bias in observed recommenders

Data extraction should ideally be conducted without interfering significantly with the recommender. Unlike platforms serving static websites, modern algorithms and models continuously track and adapt for up-to-date personalization, introducing the likelihood that bot actions become integrated into the functioning of the recommender, through such mechanisms as re-training (fine-tuning) based on user activity logs.

The degree of bias introduced is directly proportional to the similarity between bot actions and user actions. To exemplify the point, and at the other extreme, offensive bots may engage in *poisoning* attacks (Fang, Gong, and Liu 2020), interacting with specific items, to prompt the recommender to promote them to a larger audience. It is worth noting that this philosophy of 'just enough' interaction aligns with legal considerations, such as in the European legal system, where the data collection infringement (breaching terms of service for instance) by an auditor to collect evidence must be proportionate to support a given claim (Le Merrer, Pons, and Tredan 2023).

Avoid being sand-boxed

In a tactic infamously illustrated by the 'dieselgate' scandal, certain operators may be inclined to detect and create specific favorable versions of their systems during regulatory audits, a practice known as 'deceptive manipulation' (Siano et al. 2017). This behavior could extend to archival initiatives.

While it is essential for bots to behave in a manner indistinguishable from legitimate human-operated accounts (Cresci 2020) to avoid being detected, the archival context introduces unique challenges. Simulating a user with a bot requires obfuscation to effectively trigger and collect accurate personalization. Consequently, and depending on the targeted platform, bot actions may extend beyond mere metadata collection. For example, on platforms like YouTube, bots might emulate video visualization to conceal their true nature. Although this incurs significant traffic generation, it may be deemed unavoidable to achieve effective personalization thus the data extraction goal.

3.2 Methodological Challenges

3.2.1 Realism and representativity

Data collection from users and associated limitations

Common practices for collecting data on how online platforms personalize the user's experience include data acquisitions (Hosseinmardi et al. 2021) and data donations (Ohme and Araujo 2022), where users willingly share or sell data related to their personalized experiences on specific platforms. This can happen through the use of a dedicated plugin in their web browser, see for example the 2017 ProPublica article. While this approach provides valuable information for archivists, it unfortunately introduces significant

problems.

The first challenge arises from the widespread (and still growing) use of mobile applications to access platforms, replacing the conventional web browser access. These applications, tightly controlled by platform providers, conveniently prevent data extraction, and mobile operating systems do not support the use of plugins. Additionally, the scale of reaching and persuading a large audience to participate in a common archival objective proves complex and often costly. Consequently, the data obtained may not be sufficiently representative for upstream analysis by researchers, leading to potential biases since those willing to install plugins are likely tech enthusiasts, representing only a specific subset of society.

Personalization relies on platform algorithms applied to user profiles, containing the history of user actions. However, gathering data from users does not ensure the completeness of data in this intricate relationship presenting challenges akin to any data collection in a vast array of possibilities. This completeness is essential for performing unbiased and meaningful analyses of collected data donations.

Lastly, as personalization exposes users' tastes and habits, raising concerns about privacy, compliance with legal requirements becomes a critical consideration. For a detailed discussion on the impact of the nature of the collected data on legal obligations, please refer to Le Merrer, Pons, and Tredan (2023).

Personas from simulated users

Personas, in the context of simulated users, refer to users simulated by bots with a well-defined agenda: persona x might simulate on YouTube a video game enthusiast residing in the USA, while persona y might simulate a French individual using YouTube as a news source. Scripting allows these bots to exhibit various behaviors, employ geographically distributed IP addresses, and interact with the platform incorporating daily habits, for example. The art of crafting advanced bots lies in constructing the most realistic interactions to convincingly impersonate specific user types (Cresci 2020). Control conditions can be established to ensure that programmers accurately trigger personalization with their bots (Le Merrer, Tredan, and Yesilkanat 2023).

These bots address some of the challenges associated with obtaining personalization data from real users. They offer full control over the actions taken, directly linked to systematically collected personalization. Bias is minimized, as programmers control the history of actions and metadata associated with all their bots.

Conveniently, bots are more straightforward to set up than recruiting real users. Bringing the analysis to a larger scale relates only to the cost of hosting these scripts and the data they generate. Furthermore, there are no legal issues concerning personal data (at least in the European Union, as exposed by Le Merrer, Pons, and Tredan (2023), as the personalization of the collected data does not involve real individuals. However, a drawback is the inability for programmers/archivists to ascertain whether their bots have been detected by the platform. This introduces the possibility that the platform might willingly treat these bots differently, potentially offering similar personalization as real users with comparable profiles or occasionally biasing their personalization as it sees fit.

In the following section, we delve into the challenges associated with these personas,

and their role in extracting personalization.

Data collection from simple and specific actions

In an alternative scenario, an archivist may find the need to extract personalization data from profiles characterized by clear histories and routine actions. These actions are clearly not aimed at approaching user behavior, but rather focus on extracting consistent data across time. Consider the recommendations made to a profile diligently visiting YouTube's news page every day at noon, or those made in response to a profile limited to entering a set of predefined words of interest in the search bar. Despite the simplicity of these scenarios, they allow for precise tracking of the recommender system's evolution on the platform.

In this simplified case, the archivist might opt for a blank user, i.e. a user profile devoid of any history or prior interactions with the platform. The recommendation algorithm would not be influenced by previous choices, with the aim of having recommendations from the platform in the most neutral as possible scenario.

Another synthetic data extraction approach involves a one-shot, yet potentially comprehensive, gathering of personalization data in response to well-defined sequential actions on the platform. This approach can serve as a basis for auditing a specific aspect of the recommender at a given point in time.

3.2.2 Mainstream vs. fringe profiles

Personalization can be envisioned as a vast space, like a country, where each potential user profile corresponds to an address. Given the impossibility of exhaustively exploring this space, a deliberate selection, or sampling, must be made: where should the focus of observation lie? While virtually any focusing strategy is possible, we briefly introduce two paradigmatic ones.

The first, which we term mainstream, involves focusing the observation of personalization on the most prevalent profiles, those with the most common tastes and behaviors among the user population. In our metaphorical country of personalization, this corresponds to directing sampling towards densely populated areas, such as the capital city. The primary advantage of this strategy lies in its efficiency: each personalization is likely to capture the experience of a substantial user base. Randomly sampling inhabitants of Greece would yield roughly one third residing in the Athens region. Likewise, programming bots to watch videos suggested to an empty profile at random would likely result in mainstream tastes.

The clear drawback of this strategy is its potential to overlook what is not mainstream and which could hereafter be referred to as 'fringe' personalization. This pertains to how personalization influences users who are not representative of the overall user population. For example, in the Facebook-Cambridge Analytical data scandal (Insider 2019), Cambridge Analytical targeted highly specific profiles that diverged from the mainstream. Similarly, the rabbit-hole phenomenon, often studied as a fringe personalization regime, focuses on particular sub-populations, such as anti-vaccine advocates, conspiracy theorists, and far-right movements. A mainstream-only archival strategy would not support such studies, even though they hold value for archiving.

3.3 Usability Challenges

Once technical and methodological solutions have been developed, a final challenge lies in determining the appropriate methods for exploring the collected personalized data (as highlighted in Kelly et al. 2018, and put in relation with the Wayback Machine). A general approach would be to target the most accurate browsing experience, allowing future archive users to closely experience the mechanics of contemporary systems. However, implementing such a system would require significant efforts in emulating the logic of each target website. For example, TikTok and YouTube obey different browsing mechanisms and each requires recreation. Moreover, the archive is destined to be an imperfect copy of the original platform, capturing only a fraction of the website realistically, and unable to reproduce the complete dynamics of these social networks.

An opposing approach could aim for a unified presentation enabling future archive users to compare media platforms on an identical basis, with an implicit emphasis on content rather than presentation.

A central ergonomic challenge lies in navigating personalization in itself. While the Wayback Machine provides a suitable slider for exploring the (continuous though discretely sampled) temporal dimension of a web archive, envisioning an interface for exploring personalization poses a unique question. For archives based on synthetic profiles, TheirTube[1] prompts visitors to select one of the personas used for collection. However, no such solution exists for archives based on (real) data collections directly from users. This distinction illustrates how a usability approach is contingent on the technical and methodological decisions that shaped the personalization data collection.

4. Conclusions and open questions

In this chapter, we assert, from our technical perspectives, that personalization poses a challenge to traditional web archiving methods. We demonstrate how personalization impacts data collection on YouTube and the technical challenges associated with its analysis. Our data collection on YouTube emphasizes that the notion of a universal web no longer holds. There is no singular version of YouTube, as each user is presented with content tailored to their past actions or user profile characteristics. We contend that an effective archival strategy must include the archiving of contemporary personalization, in a consistent manner, and in addition with proposals to leverage ‘emergency’ and focused archiving of some platforms during important events for instance (Schafer, Truc, and Badouard 2019). Addressing this challenge raises several technical, methodological, and usability issues, such as how to manipulate personalization, which personalized versions to archive, and how to present this personalization to future archive users. The interconnected nature of these problems underscores the conclusion of this chapter: the need to integrate and reconcile diverse disciplinary perspectives (computer science, history, usability) to construct coherent solutions facilitating accurate future analyses of our contemporary, personalized times. Although all possible approaches may come with advantages and

[1] https://www.their.tube

drawbacks, we believe it is crucial to define a set of good practices facing the necessary archiving of personalized content.

While our primary focus has been on web archiving, personalization also impacts the information disseminated to users through mobile applications (apps). The technical opacity inherent in mobile apps compared to web pages adds an extra layer of complexity. Due to the increased collection of personal information through these mobile apps, personalization becomes extremely efficient and is referred to as user profiling.

We argue that, with the goal of archiving the user experience in today's web interactions, both personalization and profiling should be integral to the archival process, for they are deeply embedded in our digital lives and should be preserved for the benefit of historians.

Finally, while our current focus lies in understanding how personalization impacts our methods for documenting the history of the web, we believe that the very process of personalization (along with profiling) ought to be studied as a historical phenomenon, thereby recognizing its central role as a contemporary opinion-maker.

References

Azcoitia, Santiago Andrés, and Nikolaos Laoutaris. 2022. “A Survey of Data Marketplaces and Their Business Models.” *arXiv*. https://doi.org/10.48550/ARXIV.2201.04561.

Bandy, Jack, and Nicholas Diakopoulos. 2021. “Curating Quality? How Twitter’s Timeline Algorithm Treats Different Types of News.” *Social Media + Society* 7 (3).

Cloudfare. 2023. “What is rate limiting? | Rate limiting and bots.” https://web.archive.org/web/20240424000000*/https://www.cloudflare.com/learning/bots/what-is-rate-limiting/.

Covington, Paul, Jay Adams, and Emre Sargin. 2016. “Deep Neural Networks for Youtube Recommendations.” In *Proceedings of the 10th Acm Conference on Recommender Systems*, 191–98.

Cresci, Stefano. 2020. “A Decade of Social Bot Detection.” *Commun. ACM* 63 (10): 72–83.

Eg, Ragnhild, Özlem Demirkol Tønnesen, and Merete Kolberg Tennfjord. 2023. “A Scoping Review of Personalized User Experiences on Social Media: The Interplay Between Algorithms and Human Factors.” *Computers in Human Behavior Reports* 9: 100253.

Eslami, Motahhare, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. ““I Always Assumed That I Wasn’t Really That Close to [Her]’: Reasoning About Invisible Algorithms in News Feeds.” In *Proceedings of the 33rd Annual Acm Conference on Human Factors in Computing Systems*, 153–62. CHI ’15. New York, NY, USA: Association for Computing Machinery.

NOYB European Center for Digital Rights. 2023. “How Mobile Apps Illigally Share Your Personal Data.” https://web.archive.org/web/20240424000000*/https://noyb.eu/en/how-mobile-apps-illegally-share-your-personal-data.

Fang, Minghong, Neil Zhenqiang Gong, and Jia Liu. 2020. “Influence Function Based Data Poisoning Attacks to Top-N Recommender Systems.” In *Proceedings of the Web Conference 2020*, 3019–25.

Farseev, Aleksandr, Qi Yang, Andrey Filchenkov, Kirill Lepikhin, Yu-Yi Chu-Farseeva, and Daron-Benjamin Loo. 2020. “SoMin.ai: Personality-Driven Content Generation Platform.” *arXiv E-Prints*, November, arXiv: 2011.14615.

Gupta, Udit, Carole-Jean Wu, Xiaodong Wang, Maxim Naumov, Brandon Reagen, David Brooks, Bradford Cottel, et al. 2020. “The Architectural Implications of Facebook’s Dnn-Based Personalized Recommendation.” In *2020 Ieee International Symposium on*

High Performance Computer Architecture (Hpca), 488–501. IEEE.
Gustarini, Mattia, Marcello Paolo Scipioni, Marios Fanourakis, and Katarzyna Wac. 2016. "Differences in Smartphone Usage: Validating, Evaluating, and Predicting Mobile User Intimacy." *Pervasive and Mobile Computing* 33: 50–72.
Hosseinmardi, Homa, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M Rothschild, and Duncan J Watts. 2021. "Examining the Consumption of Radical Content on Youtube." *Proceedings of the National Academy of Sciences* 118 (32): e2101967118.
Business Insider. 2019. "The Cambridge Analytica Whistleblower Explains How the Firm Used Facebook Data to Sway Elections." https://web.archive.org/web/20240424000000*/https://www.businessinsider.com/cambridge-analytica-whistleblower-christopher-wylie-facebook-data-2019-10?r=US&IR=T.
Kelly, Mat, Justin F Brunelle, Michele C Weigle, and Michael L Nelson. 2013. "A Method for Identifying Personalized Representations in Web Archives." *D-Lib Magazine* 19 (11–12).
Kiesel, Johannes, Arjen P de Vries, Matthias Hagen, Benno Stein, and Martin Potthast. 2018. "WASP: Web Archiving and Search Personalized." https://ceur-ws.org/Vol-2167/paper6.pdf
Ledwich, Mark, and Anna Zaitsev. 2020. "Algorithmic Extremism: Examining Youtube's Rabbit Hole of Radicalization." First Monday.
Le Merrer, Erwan, Ronan Pons, and Gilles Tredan. 2023. "Algorithmic Audits of Algorithms, and the Law." *AI and Ethics*, 1–11.
Le Merrer, Erwan, and Gilles Tredan. 2018. "The Topological Face of Recommendation." In Complex Networks & Their Applications Vi: *Proceedings of Complex Networks 2017 (the Sixth International Conference on Complex Networks and Their Applications),* 897–908. Springer.
Le Merrer, Erwan, Gilles Tredan, and Ali Yesilkanat. 2023. "Modeling Rabbit-Holes on Youtube." *Social Network Analysis and Mining* 13 (1): 100.
Milligan, Ian, Nick Ruest, and Jimmy Lin. 2016. "Content Selection and Curation for Web Archiving: The Gatekeepers Vs. The Masses." In *Proceedings of the 16th Acm/Ieee-Cs on Joint Conference on Digital Libraries*, 107–10.
Mozilla. 2020. "Political Advertisements from Facebook." https://web.archive.org/web/20240424000000*/https://foundation.mozilla.org/en/blog/step-inside-someone-elses-youtube-bubble.
Ohme, Jakob, and Theo Araujo. 2022. "Digital Data Donations: A Quest for Best Practices." *Patterns* 3 (4).
Pariser, Eli. 2012. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Books.
Powers, Elia. 2017. "My News Feed Is Filtered?" *Digital Journalism* 5 (10): 1315–35.
Exodus Privacy. "Exodus Privacy Analyzes Privacy Concerns in Android Applications." https://web.archive.org/web/20240424000000*/http://https://exodus-privacy.eu.org/.
ProPublica. 2017. "Political Advertisements from Facebook." https://web.archive.org/web/20240424000000*/https://www.propublica.org/article/help-us-monitor-political-ads-online.
Rastegarpanah, Bashir, Krishna Gummadi, and Mark Crovella. 2021. "Auditing Black-Box Prediction Models for Data Minimization Compliance." *Advances in Neural Information Processing Systems* 34: 20621–32. https://proceedings.neurips.cc/paper_files/paper/2021/file/ac6b3cce8c74b2e23688c3e45532e2a7-Paper.pdf
Digital Services Act. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending

Directive 2000/31/EC (Text with EEA Relevance). OJ L. https://web.archive.org/web/20240424000000*/http://data.europa.eu/eli/reg/2022/2065/oj/eng.

Salganik, Matthew J., and Duncan J. Watts. 2008. “Leading the Herd Astray: An Experimental Study of Self-Fulfilling Prophecies in an Artificial Cultural Market.” *Social Psychology Quarterly* 71 (4): 338–55.

Schafer, Valérie, Gérôme Truc, Romain Badouard, Lucien Castex, and Francesca Musiani. 2019. “Paris and Nice Terrorist Attacks: Exploring Twitter and Web Archives.” *Media, War & Conflict* 12 (2): 153–70.

Schmidt, Jan-Hinrik, Lisa Merten, Uwe Hasebrink, Isabelle Petrich, and Amelie Rolfs. 2019. “How Do Intermediaries Shape News-Related Media Repertoires and Practices? Findings from a Qualitative Study.” *International Journal of Communication* 13 (0). https://web.archive.org/web/20240424000000*/https://ijoc.org/index.php/ijoc/article/view/9080.

Siano, Alfonso, Agostino Vollero, Francesca Conte, and Sara Amabile. 2017. “‘More Than Words’: Expanding the Taxonomy of Greenwashing After the Volkswagen Scandal.” *Journal of Business Research* 71: 27–37.

“Teens, Social Media and Technology.” 2023. Pew Research Center. https://web.archive.org/web/20240424000000*/https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/.

“The Christchurch Call to Action to Eliminate Terrorist and Violent Extremist Content Online.” n.d. https://web.archive.org/web/20240424000000*/https://www.christchurchcall.com/assets/Documents/Christchurch-Call-full-text-English.pdf.

Xu, Runhua, Remo Manuel Frey, Elgar Fleisch, and Alexander Ilic. 2016. “Understanding the Impact of Personality Traits on Mobile App Adoption – Insights from a Large-Scale Field Study.” *Computers in Human Behavior* 62: 244–56.

Zhao, Sha, Shijian Li, Julian Ramos, Zhiling Luo, Ziwen Jiang, Anind K. Dey, and Gang Pan. 2019. “User Profiling from Their Use of Smartphone Applications: A Survey.” *Pervasive and Mobile Computing* 59: 101052.

Mapping the archival horizon: A comprehensive survey of COVID-19 web collections in European GLAM institutions

Nicola Bingham

Abstract: This chapter analyzes the COVID-19 web collections of Galleries, Libraries, Archives, and Museums (GLAM) across Europe. As the pandemic reshaped the digital landscape, these institutions have worked to archive web content from this unprecedented event. The study explores European GLAM web archiving efforts, focusing on the scope, methods, and challenges of curating COVID-19 content. It features a comparative analysis that highlights similarities and differences in collection strategies, emphasizing the implications for future digital preservation and access to pandemic-related content. This research offers insights valuable to information professionals and researchers studying the role of GLAM institutions in preserving key contemporary historical moments.

Keywords: COVID-19, web collections, digital preservation, GLAM institutions, web archiving

Introduction

Amid the COVID-19 pandemic, cultural heritage institutions embarked upon substantial endeavors to document this epochal period for posterity. In 2021, within the purview of the WARCNet project, a cohort of researchers delved into the dynamics of COVID-19 web archiving within the GLAM sector in Europe. In the summer of 2022, a comprehensive survey[1] was initiated to solicit participation from European counterparts in the GLAM sector, seeking insights into the breadth of their web collections and collection methodologies. The results of this study will be published in 2024. The focal point of the 2022 study encompassed three distinct respondent cohorts delineated by their respective countries of origin: the United Kingdom, Belgium, and other European nations. This segmentation was necessitated by the notably substantial response rates from the UK and Belgium, attributed to the distribution strategy employed for the survey. The current chapter integrates the survey's findings and extends its scope by assimilating an additional comprehensive literature review, thereby enriching the analysis of the gathered data.

The Survey

[1] A copy of the survey questions is held in the British Library Research Repository https://doi.org/10.23636/g5n0-pp30

Nicola Bingham, British Library, United Kingdom, nicola.bingham@bl.uk, 0000-0002-5510-9869

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Nicola Bingham, *Mapping the archival horizon: A Comprehensive survey of COVID-19 web collections in European GLAM institutions*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.11, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 95-112, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

The questions included in the survey were informed by a hypothesis about the decisive factors that form web archives. In contrast to physical archives, archiving the web is a relatively novel endeavor characterized by technological advances and little international coherence. Local strategies are governed by specific national legislations and new tools for harvesting are continuously being developed as web archiving technology evolves. The survey aimed to probe the extent to which these differences have affected both rapid response and regular collection strategies regarding the COVID-19 crisis. By addressing questions on juridical and technological factors, we guide respondents' attention towards how these have affected collecting strategies.

The survey, conducted from June 13 to September 5, 2022, spanned approximately three months. It reached around 600 recipients via email addresses and distribution lists such as the Archives-NRA JISC mail and IIPC lists, compiled through the project team's existing contacts, network, and desk-based research. The team targeted individuals knowledgeable about collecting policy or involved in research within Museums, Libraries, Archives, and Galleries across Europe. Distribution favored the UK and Belgium due to team members' affiliations with National Libraries in those countries, resulting in contact with 99 organizations/individuals in the UK and 354 in Belgium. Additionally, efforts were made to engage European National Libraries and Archives affiliated with the International Internet Preservation Consortium (IIPC).

The survey was structured into sections focusing on respondent profiles, motivations behind collecting pandemic-related materials, collection scope, the contextualization of collection activities, access challenges, and creation hurdles. SNAP survey software was chosen to carry out the survey primarily because it was utilized by the British Library (BL) and offered secure data storage.

Respondents

The survey garnered 61 responses from 12 countries, resulting in a completion rate of roughly 10%. Factors contributing to the low response include the timing coinciding with university staff leave over the summer, ambiguity in identifying recipients in large organizations, and resource constraints in smaller ones. Additionally, the survey's title and focus may have deterred non-collecting organizations, leading some to perceive their input as redundant. Acknowledging the low response rate as a limitation underscores the need for a cautious interpretation of the results. To maintain respondent anonymity, countries with fewer than five responses were grouped under 'other European countries', with two unidentified

nationalities included in this category. Therefore, the analysis compares COVID-19 web archiving practices across three regions: the UK, Belgium, and other European countries, with 33% from the UK, 25% from Belgium, and 41% from other European countries among the respondents.

Motivation for collections

The motivation for initiating a special collection on COVID-19 was surveyed through the first two questions of the survey. Participants were first asked to report whether their institution had indeed curated a special collection of web materials about the COVID-19 crisis. The average percentage of respondents who initiated a COVID-19 collection across Europe was approximately 70.67%.

Respondents were asked to describe where the initiative to collect web materials related to the COVID-19 crisis had originated. Motivation was notably consistent across the UK, Belgium, and other European countries. In both the UK and Belgium, roughly half of the respondents indicated that the initiative came from the institution's staff (47% in the UK and 54% in Belgium). Approximately 30% of respondents from these countries mentioned that the impetus came from institutional management (32% in the UK and 31% in Belgium) and in other European countries, a majority (54%) also cited that the impetus predominantly originated from the staff. This similarity in responses across regions suggests a common trend where staff within institutions played a significant role in driving the initiative to collect web materials related to COVID-19.

Collaboration and partnerships

From the results of the survey and a review of additional literature, several insights emerge regarding collaborative partnerships between archiving institutions and external organizations in building COVID-19 web collections.

Collaboration for technical support was observed to varying degrees among institutions in Belgium, the UK, and other European countries in the survey. Belgian institutions showed relatively higher levels of technical collaboration. Chiara Zuanni, (2022) writing in the context of museums, identifies that the urgent need to collect digital items led to rapid developments in this area together with novel collaboration between memory organizations and digital specialists.

Several instances of participatory approaches have been observed, such as the National Library of Luxembourg's call for participation to integrate websites overlooked in the initial collecting phase (Schafer and Els, 2020). Furthermore, the Danish Web collection on coronavirus was part of a general project documenting coronavirus lockdowns in Denmark in 2020.

This effort was a cooperation between several cultural institutions, the National Archives (Rigsarkivet), the National Museum (Nationalmuseet), the Workers Museum (Arbejdermuseet), local archives, and the Royal Danish Library. Also in Denmark, the 'Days with Corona' initiative from the Danish Folklore Archives run by the National Museum and the Royal Library called on citizens to contribute photos, narratives, and websites for inclusion in the web archive, showcasing a broader engagement strategy involving cultural institutions, local archives, and libraries. Netarchive also asked the public for help by nominating URLs of web pages related to coronavirus, social media profiles, hashtags, memes, and any other relevant material (Schostag 2020).

The experience gained from COVID-19 collections has led to plans for future collaborations and thematic collections. For example, the National Library of Luxembourg aims to expand partnerships with the Government IT Centre, the CNA (National Audiovisual Centre) and the CNL on Luxembourg authors and publishers and other cultural institutions to deepen thematic collections. (Schafer and Els 2020)

Overall, these insights demonstrate the varying degrees of collaboration, outreach strategies, and participatory efforts employed across different countries in the context of COVID-19 collection initiatives. These approaches not only enrich the collections but also pave the way for future collaborations and more inclusive, engaging archival practices.

Platforms like the Internet Archive (IA) and the UK Web Archive offer features enabling users to nominate websites, acknowledging that many sites are missed simply because archivists are unaware of them. While participation enriches archives, it remains somewhat sporadic and mainly reaches those already digitally engaged. Efforts for broader engagement are expected to bring different participation forms and contribute to more inclusive collections.

Cui Cui et al. (2023) noted in their conference presentation that participatory web archiving redistributes power among stakeholders e.g., community partners, creators, and users, aiming to overcome limitations in traditional archiving. However, the participatory approach raises questions about effectiveness, mechanisms, and impacts.

Underrepresented groups

As Schafer and Winters (2021) point out, in terms of minorities, and marginalized populations affected by the pandemic, "Web archives include information for and about diverse groups in society and hold out the promise of preserving the voices of individuals who in previous centuries would only have featured in an archive if they interacted with the church, central government, or the law."

The survey highlighted several examples of the various archiving institutions' attempts to contact underrepresented groups, such as engaging mental health charities, using social media, contacting partner institutions, and utilizing internal networks (e.g., LGBTQ staff networks) to connect with underrepresented communities during the creation of COVID-19-related web archives (Geeraert and Bingham 2020).

Religious communities have been the focus of some collecting efforts, for example, the Marian Library, University of Dayton, Ohio built a collection which reflected how the pandemic affected underrepresented and underserved communities of Catholics. The collection gives a glimpse of how the pandemic has affected the faith and religious practice of 'ordinary' Catholics (Harris et al. 2023).

Another example was provided by Ben Els at the BnL, who initially looked at a Muslim community website (shoura.lu) which contained information and recommendations for its community about services in mosques, religious holidays, etc. BnL then expanded its collecting strategy to include other religious communities. Ben Els pointed out "It is important to capture the minority viewpoint as well and I have also included sites for border residents, for example, the site frontaliers.lu, because they face different problems from residents." (Schafer and Els 2020)

There was increased global attention on equality, diversity, and inclusion during 2020 due to the pandemic's spotlight on inequalities and the rise of movements like Black Lives Matter. Initiatives focused on the challenges faced by communities during events like racial injustice protests, struggles within the LGBTQ+ community, and instances of anti-Asian racism. Archivists prioritized capturing these narratives to ensure a more accurate representation of ongoing historical events and discussions (Greenwood 2022).

Meanwhile, Schafer and Winters (2021) noted controversies in web archiving, citing disputes over ethical concerns, such as the collection of sensitive content during events like the George Floyd protests. These debates prompted calls for more inclusive archiving policies, advocating for broader representation, particularly of Black perspectives, underscoring the growing complexity and ethical considerations in archival practices. Schafer and Winters highlight instances of disputes, including lawsuits and debates between entities like the Internet Archive (IA) and editors. One example involves clashes between the IA and Doc Now during the George Floyd protests, where concerns about risks to protesters' safety were raised against the collecting effort, reflecting a divergence of opinions on the ethical aspects of archiving sensitive content.

Scope

Regarding the level of archiving, the results from the GLAM survey indicate a correlation between the level at which institutions operate and the focus of their web archiving efforts. In the UK, a majority of respondents collected at a national scale, which aligns with a significant portion of these institutions operating at a national level. In Belgium, a higher percentage of institutions operated at a regional level, which coincided with more respondents indicating collection efforts at the regional level.

Smaller web archives, like those operated by universities or individual scholars, may lack resources and infrastructure for independent hosting and preservation. Consequently, they often link their content to larger initiatives like the Internet Archive to avoid issues with copyright, data storage, and preservation, enabling their content to be preserved and accessible through such extensive platforms (Priem and Grosvenor 2022).

In many cases specialist collections were built adhering to the institutional scope for example the University of Dayton's collection on Marian Devotion (Harris et al. 2021). However, some web archives are not under the control of libraries, such as arquivo.pt, the Portuguese Web Archive, which is linked to the Portuguese national research and education network.

There is evidence that web collections vary based on specific national or local events. For instance, the Royal Danish Library coronavirus web collection includes web activity related to Queen Margarethe II's 80th birthday celebration during the pandemic (Schostag 2020). This example underscores that local events unique to different nations significantly shape the content and focus of individual web collections. It prompts the question of whether similar examples from other countries, like the UK's 'eat out to help out' initiative, would also impact web collections in distinct ways.

In some cases, Web archives have structured their COVID-19 collections within a broader context of historical and potential future pandemics. However, the selection process has been somewhat subjective, often excluding recent pandemics like swine flu or HIV/AIDS (Priem and Grosvenor 2022). Some initiatives such as the UK Web Archive have organized COVID-19 collections within a larger category covering other pandemics, potentially indicating a placeholder for future pandemic-related collections (Geerart and Bingham 2020).

Similarly, the ISCHE Education & Pandemics Archive acknowledges the broader context of historical pandemics and epidemics, highlighting their significance in local, national, and transnational education histories. These initiatives aim to recognize the 'pandemic century' and the multitude of disruptions and challenges, reshaping historical perspectives that traditionally prioritize human-centric narratives and control (Priem and Grosvenor 2022).

Type of content collected

The survey investigated preferences in collecting social media and website content, revealing distinct tendencies between Belgium and the UK. In Belgium, Facebook tops the list as the most frequently collected social media platform, followed by Twitter, Instagram, and YouTube. Conversely, the UK prioritizes Twitter, blogs, university websites, and non-profit websites. These variations might be attributed to technological limitations, for example, Heritrix, a widely used crawler in UK web archives, faces challenges in capturing Facebook content whereas Belgium's preference for Facebook archiving could stem from the Flemish Institute for Archives research project, meemoo, on social media archiving practices: "Best practices for archiving social media in Flanders and Brussels" (meemoo 2023).

These differences in collection preferences highlight varying priorities and technological constraints between different web archiving institutions. Schostag explains that at the Royal Danish Library "More or less successfully, we tried to capture content from Facebook, Tik-Tok, twitter, youtube, instagram, reddit, imgur, soundcloud, and pinterest. Twitter is the platform we are able to crawl with Heritrix with rather good results. We collect Facebook profiles with an account at Archive-It, as they have a better set of tools for capturing Facebook. With frequent Quality Assurance and follow-ups, we also get rather good results from Instagram, TikTok, and Reddit. […]As Heritrix has problems with dynamic web content and streaming, we also used Webrecorder.io […]. However, captures with Webrecorder.io are only drops in the ocean." (Schostag 2020).

The National Library of Luxembourg (BnL) focused primarily on archiving websites due to the significant technical challenges and higher costs associated with archiving social media platforms. While they included some Facebook pages in their collection, they prioritized news media, websites, and Twitter due to their higher effectiveness in archiving compared to other social media platforms (Schafer and Els 2020).

Sub-categories

The survey asked respondents to comment on the criteria for selection in Covid-19 web collections from the following options; 'based on language(s)', 'top-level domain', 'subject/theme of web content', 'nationality of the creators/owners', 'postal addresses mentioned in the web content', 'specific hashtags', and 'specific social media profiles'.

Belgium and the UK primarily relied on the subject or theme of the web content, nationality of creators/owners, and specific social media profiles for collection scoping. In Belgium, no respondents indicated that the top-

level domain or the postal addresses mentioned in the web content were determining criteria and in the UK none of the respondents indicated that language was used as a selection criterion. This can be partly explained by the fact that there is no top-level domain crawl 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. Overall, the survey indicated that the most common way to scope COVID-19 collections among all the survey respondents was based on the subject/theme of the web content.

News content was a priority, however, it was handled differently according to the collecting institution. At the National Library of Luxembourg between mid-March and mid-July, for example, curators selected individual articles from Luxembourg news websites constituting a collection of just over 26,000 articles. From the beginning of June, they requested a daily capture of each news website from the Internet Archive, which had previously been undertaken only twice a year (Schafer and Els 2020).

In contrast, Schostag (2020) explains that the Danish Web Archive crawls all Danish news media from several times daily to once weekly, so felt there was no need to curate individual news articles in the pandemic event crawl. Rather, they focused on augmented activity on social media, blog articles, new sites emerging in connection to the event, and reactions in news media outside Denmark.

Curators building web archives about the COVID-19 pandemic responded dynamically to unfolding events, employing inclusive approaches in content selection. They focused on various pandemic-related topics such as medical care, government communications, lockdown effects, remote work, and cultural activities. Specific themes were identified in Luxembourg, Belgium, and the UK, encompassing liberalism, solidarity, inequality, and more (Schafer and Els 2020).

The collection scopes varied, including creator-based and local collections, and selection methodologies ranged from keyword searches to open calls for participation. The curated subtopics extended beyond health effects to societal concerns like changing political and economic landscapes, threats to food security, misinformation spread, and governmental exploitation of the crisis. However, some collections, like Cornell University's focus on international labor challenges, leaned more towards an economic perspective and lacked a direct human narrative (Greenwood 2022).

Temporal span

The survey revealed that the temporal span of COVID-19 web collections varied across institutions and countries. Most collections began

their gathering process in March 2020, aligning with the imposition of lockdown measures in Belgium and the UK. Other European countries started slightly earlier, primarily in the first half of March 2020. However, the Danish web archive commenced its efforts about six weeks before the virus reached Denmark, prompted by global reactions to a controversial cartoon. As stated by Schostag (2020), "In a sense, the story of Corona and the national Danish Web Archive (Netarchive) starts at the end of January 2020 – about 6 weeks before Corona came to Denmark. A cartoon by Niels Bo Bojesens in the Danish newspaper 'Jyllandsposten' (26 Jan 2020) showing the Chinese flag with a circle of yellow corona-viruses instead of the stars caused indignation in China and captured attention worldwide. We focused on collecting reactions on different social media and in the international news media. Particularly on Twitter, a seething discussion arose with vehement comments and memes about Denmark."

Regarding the end dates for collection efforts, by the time of the survey conducted between June and September 2022, a quarter of UK respondents and almost half of the Belgian respondents had ceased collecting. This cessation was due to various factors, such as diminished interest, staff reallocation, completion of projects, the end of pandemic-related measures, lack of relevant online content, and technical difficulties in capturing specific social media.

The decision to halt collections was not universally aligned and was often influenced by practical considerations, like budgetary constraints. At the National Library of Luxembourg, for example, Ben Els mentions having ample storage capacity (terabytes), but expresses reluctance to allocate an additional 5 terabytes due to limitations. He had hoped for a slowdown in June but observed continued activity, maintaining the pace of collection similar to that of March. He notes comparing the number of articles collected from media websites to the recorded number of COVID-19 cases as part of their collection analysis (Schafer and Els 2020). The unpredictability of the pandemic's trajectory made it challenging for institutions to definitively determine when to conclude their collection activities, leading to varied responses among institutions.

Some institutions expressed intentions to cease collection efforts once the World Health Organization declared an end to the pandemic. The relationship between these intentions and actual collection cessation could provide valuable insights into the dynamics between collecting practices and the perceived trajectory of the pandemic.

The Danish Royal Library emphasizes the importance of broad, domain-scale archiving to capture a comprehensive view. Their timing for domain crawling during the Danish Corona lockdown, starting on March 14, proved fortuitous. These broad crawls, conducted up to four times a year, capture all Danish-related content across domains. This approach allows for the

inclusion of Corona-related content that curators might not identify using their usual methods, utilizing keyword searches and link scraping tools (Schostag 2020).

Technology and quality assurance

The survey focused on the technologies and quality assurance methods used in web content harvesting. Different tools were employed across countries, with Belgium favoring Webrecorder and Chrome SingleFile, while the UK used Webrecorder and Heritrix. Larger institutions preferred Heritrix for scalability, while smaller entities chose user-friendly tools like Webrecorder, aligning with their specific archiving needs. Quality control primarily involved visual inspection, with some using patch crawling or specific methods.

Webrecorder emerged as the most popular tool across surveyed countries. Organizational tool choices vary due to complex factors, leading some to utilize combinations of tools. Large-scale initiatives opt for robust, scalable tools like Heritrix, capable of capturing extensive content, while smaller projects prefer simpler tools like Webrecorder, requiring less technical expertise.

The survey identified a long tail of less frequently mentioned tools, often open-source and straightforward. For instance, Chrome SingleFile, a browser extension for archiving complete web pages in a single HTML file, appeals to smaller organizations for its offline viewing capabilities. Some smaller entities archive web content without specific legal mandates, utilizing tools like Instaloader for Instagram, operating within platform terms of service and legal compliance.

Access

The landscape of access to web archive collections varies according to country and institution. For example, the access restrictions imposed due to personal data protection laws, as seen in the Danish case, limit public access to COVID-19 web archives. In Denmark, only researchers affiliated with specific institutions can apply for access related to particular research projects. Although one project on values in COVID-19 communication has already been initiated (Schostag 2020).

A common challenge across Belgium, the UK, and other European countries is the underutilization of these special collections for research and teaching. Responses from institutions in these countries indicated that the collections were rarely used or that there was uncertainty about their usage. This lack of insight into collection usage suggests that either these

collections are underused or that there is a gap in understanding their potential.

Few institutions in Belgium or the UK provided data-level access or offsite access, whereas some other European countries did so. Additionally, sharing derived datasets or seed lists with the public was rare in the UK and Belgium, but more common in other European countries. Restrictions on access were often linked to access being limited to the institution's reading room, with additional limitations surrounding intellectual property rights, personal data protection, and legal deposit legislation.

Despite accessibility challenges, efforts were made to promote access. For instance, the Luxembourg National Library (BnL) provided information and starting points on its website and intended to offer interactive statistics and metadata, such as title and keywords, to aid users in navigating the collection (Schafer and Els 2020).

Ben Els (2020) stated that the archive is accessible at the BnL, but on the webarchive.lu website they also try to give information and starting points, to outline the background of the collection and explain the processes to interested people. This implies that interest from academic partnerships in accessing and analyzing COVID-19 collections exists, with the mention of WARCnet and an academic team being the first on the horizon.

Schafer and Winters (2021) highlight the complex interplay between ethical practices, legal frameworks, and protection in web archiving. In many countries, web archiving is regulated by legal deposit and copyright laws. Some countries have explicit laws addressing the archiving of digital content, enabling collection and preservation but potentially restricting access. However, legal frameworks often lag behind technological advancements, posing challenges to good governance criteria. Restrictions on access might lead researchers and the public to rely solely on platforms like the Internet Archive (IA), missing out on more comprehensive collections held by other archival institutions. The IA's copyright policy, allowing unrestricted access unless there are take-down requests, contrasts with legal limitations in other archives. Balancing citizen safeguards with increased access could foster innovation and experimentation in web archiving. This underscores the need for a balance between protecting individual rights and enhancing accessibility and innovation in web archiving practices.

In her conference presentation, Alice Austin (2023) delineated the strategic approach adopted by the Archive of Tomorrow project in addressing the complexities associated with the collection, organization, and accessibility of disputed or obsolete medical information for researchers. The deliberate inclusion of data sourced from dissenting or contentious entities, juxtaposed with content from established authoritative sources, is poised to facilitate forthcoming investigations into the dynamic interplay

between health discourse and the online sphere. Moreover, there is an anticipation that this curated collection will serve as a pivotal experimental ground for scrutinizing the methodologies involved in crafting, administering, and leveraging archived web resources for scholarly inquiry.

Research

According to Priem and Grosvenor (2022), the COVID-19 pandemic has not only sparked a renewed interest in history, it has also focused attention on how the present can be historically preserved. Therefore, it is safe to predict that the COVID-19 crisis and its documentation will be analyzed by future historians, and it will bring about methodological and technological changes that affect our ways of working as historians of education.

Piguet and Montebello (2020) point out that in response to the global Covid-19 pandemic and its impact on societal norms and individual freedoms due to widespread lockdowns, historians, sociologists, and archivists are urging individuals, institutions, and public authorities to document this unprecedented event. They emphasize the importance of archiving this exceptional period, aiming to uncover the overlooked aspects of daily life, often invisible but integral to human societies. The analysis of these archives is expected to illuminate lesser-known lives and contribute to various historiographical fields, the history of solidarities and social policies, the history of public policy, or the history of epidemics. The ultimate goal is to create an inclusive and civic memory of the pandemic, fostering collaboration between potential donors and archives globally.

There was a surge in content creation initiatives during the pandemic, as illustrated by the Made By Us project that mapped over 450 COVID-19 crowdsourcing projects globally (Priem and Grosvenor 2022). These initiatives encouraged the public to self-document and share their pandemic experiences, by posting images, audio files, texts, videos, and other content online. Marta Severo et al. (2023) describe how web archives and cultural institutions actively engaged in memorialization and historicization during the pandemic. Severo identifies two primary types of collections: one capturing born-digital data like tweets and online videos, and another seeking contributions and observations from the public, like photos and stories, to build a memory of this unprecedented time. Initiatives like ‘Vitrines En Confinement’ (‘Windows on Confinement’) aim to systematically collect and document lockdown experiences from public spaces.

In some respects, content created during the pandemic serves to project an awareness of history in the making. This content portrays individuals as protagonists in future narratives, akin to the way World War narratives are constructed, mythologized, and remembered. The pandemic is framed as a

critical historical moment, prompting GLAM organizations and various institutions worldwide to create and collect histories related to individual and community responses to COVID-19 (Priem and Grosvenor 2022).

Acker and Chaiet (2020) investigated how web archives were being "weaponized to propagate and preserve health misinformation circulating on platforms like Facebook and Twitter" as a result of the pandemic. They explain that typically most social media content has been gathered in the aggregate on a macro-level basis allowing researchers to understand the pandemic on a large scale, but that the individual's experience, or the micro-level analysis, should be considered as well.

Challenges

Survey respondents were asked what they deemed to be the most important challenges for the COVID-19 special collection. With regard to working from home, this was noted as a challenge for some web archiving institutions, however there is evidence that for others remote working provided an opportunity to devote more time to web archiving projects. Employees at the University of Dayton, for example, began remote working in March 2020 and were able to capture websites related to their library's collection development policy (Harris et al. 2021).

Technological adaptations facilitated work from home, with platforms like Archive-It supporting remote access. For example, Ben Els (2020) stated: "Video conferencing worked well. Our team is very small and for crawls the Archive-It platform is accessible from home. A lot of things that seemed impossible before have now become possible."

Quick policy changes and adaptations were crucial in response to the pandemic's onset, altering appraisal methods, prioritizing digital records, and emphasizing metadata creation and web archiving tools. Despite such shifts, managing increased workloads, securing permissions, avoiding duplication, and handling the enormity of collected material proved challenging. Respondents highlighted the emotional and time-consuming nature of the project, especially in capturing transient websites, requiring a balance between project demands and existing responsibilities. Additionally, while acknowledging the project's importance, respondents emphasized the need for increased investment in web archives to enhance responsiveness, especially considering limited resources and small team sizes in national memory institutions. The evolving situation led to reflections on the effectiveness of adding COVID-19-related content to existing collections, considering user experience.

Other challenges cited included; ensuring that work is not duplicated, making the right choices under the stress of time, data cleaning, training and coding in the midst of a pandemic, not having a platform to share

methodologies with other researchers, the long duration of the pandemic, the sheer size of the material, and ensuring the needs of the COVID-19 collections did not supersede the institutions' other existing web archiving priorities (Greenwood 2022).

Schafer and Winters highlight that the level of responsiveness demonstrated during the pandemic was impressive, but even more remarkable given the small size of the teams involved and the relatively low level of funding provided to web archives of all kinds. The web archives based in national memory institutions are usually staffed by a handful of people, who are competing internally for limited resources. Increased investment in web archives, as custodians of the world's digital cultural heritage, is essential if greater responsiveness is to be delivered.

Finally, rapid response archiving is time-consuming, and can also be emotionally consuming, especially considering the feelings of urgency that come when trying to capture and preserve very ephemeral websites. Archivists must balance the demands of this project with their other work responsibilities at a time when there is already added stress in dealing with the pandemic, personally and professionally (Harris et al. 2021).

Conclusion

This chapter undertook a comprehensive examination of the COVID-19 web collections curated by Galleries, Libraries, Archives, and Museums (GLAM) institutions in Europe. The research drew upon a comprehensive survey initiated by WARCnet researchers, aiming to engage European GLAM counterparts to gather insights into the breadth and methodologies employed in their COVID-19 web collections. Additionally, a literature review complemented the survey results to offer a more comprehensive understanding of the impact of COVID-19-related collecting among GLAM organizations in Europe and beyond.

The survey delved into the initiatives of European GLAM institutions in web archiving, scrutinizing the extent, methodologies, and challenges inherent in curating COVID-19 web collections. Through comparative analysis, this chapter illuminated both the commonalities and disparities in collection development strategies, thereby underscoring broader implications for the future of preserving and accessing pandemic-related web content. This research may provide insights for information professionals and researchers interested in the evolving role of GLAM institutions in documenting and preserving significant contemporary historical events.

However, it should be borne in mind that the survey data may carry potential biases due to specific factors in its distribution. The survey was predominantly circulated among individuals/organizations in the UK and

Belgium, influenced by the project team's proximity and networks. This skewed distribution, particularly with a higher representation of institutions in Belgium and the UK compared to other European nations, has limited the generalizability of findings.

Moreover, it is crucial to note that the survey was conducted during the pandemic, potentially capturing evolving collections in 2022, requiring organizations to adapt to unfolding events and changing priorities. Therefore, this analysis should be considered as reflecting collections in progress rather than finalized.

Despite these limitations, the survey in conjunction with a wider review of the available literature enables us to make several useful conclusions.

The insights gained underscore the diverse approaches, outreach strategies, and participatory efforts implemented across various countries during COVID-19 collection initiatives. These endeavors not only enriched the collections but also laid the groundwork for future collaborations, fostering more inclusive and engaging archival practices. While participatory web archiving aims to redistribute power among stakeholders and overcome traditional archiving limitations, questions remain about its effectiveness, mechanisms, and impacts. Despite efforts to engage users in nominating websites and contributing to archives, participation tends to reach digitally engaged individuals primarily. Broader engagement strategies are anticipated to diversify participation forms and contribute to more comprehensive and inclusive collections.

The research highlights the significance of web archives in preserving voices historically underrepresented in traditional archives. During the COVID-19 pandemic, archiving institutions made efforts to engage marginalized communities, utilizing diverse methods like social media, partner contacts, and internal networks to ensure representation in COVID-19-related web archives.

The pandemic's focus on inequalities led to increased attention on equality, diversity, and inclusion globally. Archivists prioritized documenting events like racial injustice protests, LGBTQ+ struggles, and instances of discrimination, aiming for a more accurate historical representation. However, these discussions emphasized the need for more inclusive archiving policies, particularly to represent diverse perspectives and navigate ethical dilemmas in archival practices.

The access landscape for web archive collections varies across countries and institutions, evident in cases like Denmark, where personal data protection laws restrict public access, limiting it to researchers affiliated with specific institutions and particular research projects. Similar challenges exist across Belgium, the UK, and other European nations, with collections being underutilized for research and teaching purposes, indicating a potential gap in understanding their value. Despite challenges, interest from

academic partnerships in accessing and analyzing COVID-19 collections exists, exemplified by initiatives like WARCnet. However, legal frameworks governing web archiving, often tied to legal deposit and copyright laws, present complexities.

This survey's results might inspire researchers to explore the extensive web archive collections available throughout Europe, particularly those related to COVID-19. By acknowledging biases, understanding diverse approaches, and recognizing limitations, archiving institutions and researchers can leverage and contribute to these valuable collections, enriching the knowledge surrounding COVID-19 and its societal impacts.

References

Acker, Amelia, and Chaiet, Mitch. 2020. “The Weaponization of Web Archives: Data Craft and COVID-19 Publics.” *Harvard Kennedy School (HKS) Misinformation Review* 1, no. 3: 2. Accessed December 31, 2020. https://doi.org/10.37016/mr-2020-41.

Austin, Alice. 2023. “Bridging the Gap: Capturing UK trans health discourse in the Archive of Tomorrow.” Abstract for the RESAW Conference 2023. https://web.archive.org/web/20240503135419/https://resaw2023.sciencesconf.org/434751/

Bingham, Nicola, Karin de Wild, Friedel Geeraert, and Caroline Nyvang. “Surveying the Landscape of COVID-19 Web Collections in European GLAM Institutions.” In *The Routledge Companion to Transnational Web Archive Studies*, edited by Niels Brügger. Publisher: Taylor & Francis (Forthcoming).

Cui, Cui, Stephen Pienfield, Andrew Cox, Frank Hopfgartner. 2023. “Participatory Web Archiving: The path towards more inclusive web archives?” Abstract for the RESAW Conference 2023. https://web.archive.org/web/20240503135603/https://resaw2023.sciencesconf.org/433545/

Geeraert, Friedel, and Nicola Bingham. 2020. “Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection, An interview with Nicola Bingham (British Library) conducted by Friedel Geeraert (KBR).” *WARCnet Papers*. Aarhus: WARCnet. https://web.archive.org/web/20240503135745/https://cc.au.dk/fileadmin/user_upload/WARCnet/Geeraert_et_al_COVID-19_IIPC__1_.pdf

Geeraert, Friedel, and Nicola Bingham. 2020. “Exploring special web archives collections related to COVID-19: The case of the UK Web Archive, An interview with Nicola Bingham (British Library) conducted by Friedel Geeraert (KBR).” *WARCnet Papers*. Aarhus: WARCnet. https://web.archive.org/web/20240503140007/https://cc.au.dk/fileadmin/user_upload/WARCnet/Geeraert_et_al_COVID-19_UKWA__1_.pdf

Greenwood, Amanda. 2022. “Archiving COVID-19: A Historical Literature Review.” *The American Archivist* 85, no. 1: 288–311. Accessed October 12, 2023. URL: https://web.archive.org/web/20240503140409/https://meridian.allenpress.com/american-archivist/article/85/1/288/486596/Archiving-COVID-19-A-Historical-Literature-Review.

Harris, Kayla, and Stephanie Shreffler. 2021. “Archiving Catholic Faith on the Web During

the COVID-19 Pandemic." University of Dayton eCommons Marian Library Faculty Publications, March 2021. Accessed November 5, 2023. https://web.archive.org/web/20240503140756/https://ecommons.udayton.edu/imri_fac_pub/206.

KBR. "PROMISE Project." Accessed September 12, 2023. https://web.archive.org/web/20240503141227/https://www.kbr.be/en/projects/promise-project/.

KBR. "BESOCIAL." Accessed September 12, 2023. URL: https://web.archive.org/web/20240503141538/https://www.kbr.be/en/projects/besocial/.

meemoo. "Best Practices for Social Media Archiving in Flanders and Brussels." 2023. Accessed October 14, 2023. URL: https://web.archive.org/web/20240503141733/https://meemoo.be/nl/projecten/best-practices-voor-de-archivering-van-sociale-media-in-vlaanderen-en-brussel.

Piguet, Myriam, and Caroline Montebello. 2020. "Covid-19 : pour une mémoire ordinaire de l'extraordinaire." *Libération*. Accessed October 01, 2023. URL: https://web.archive.org/web/20240503141951/https://www.liberation.fr/debats/2020/04/25/covid-19-pour-une-memoire-ordinaire-de-l-extraordinaire_1786299/

Priem, Karin, and Ian Grosvenor. 2022. "Future Pasts: Web Archives and Public History as Challenges for Historians of Education in Times of COVID-19." In *Exhibiting the Past. Public Histories of Education*, edited by Frederik Herman, Sjaak Braster, and Maria del Mar del Pozo Andrés, 177–96. Berlin, Boston: De Gruyter Oldenbourg. https://doi.org/10.1515/9783110719871-009

Schafer, Valérie, and Ben Els. 2020. "Exploring special web archive collections related to COVID-19: The case of the BnL An interview with Ben Els (BnL) conducted by Valérie Schafer (C2DH, University of Luxembourg)." *WARCnet Papers*. Aarhus: WARCnet https://web.archive.org/web/20240503142044/https://cc.au.dk/fileadmin/user_upload/WARCnet/Schafer_et_al_COVID-19_BnL.pdf

Schafer, Valérie, and Jane Winters. 2021. "The values of web archives." *International Journal of Digital Humanities*, 2, 129–44.

Schostag, Sabine. 2020. "The Danish Coronavirus web collection – Coronavirus on the curators' minds." *International Internet Preservation Consortium Blog*, July 29, 2020. https://web.archive.org/web/20240503142151/https://netpreserveblog.wordpress.com/2020/07/29/the-danish-coronavirus-web-collection/

Severo, Marta, Sarah Gensburger, and Louis Gabrysiak. 2023. "Lockdown collections and web archives: cross-exploration of BNF and INA archives." Abstract for the RESAW Conference 2023. https://web.archive.org/web/20240503142430/https://resaw2023.sciencesconf.org/435367

Zuanni, Chiara. 2022. "Contemporary Collecting in a Pandemic: Challenges and Solutions for Documenting the COVID-19 Pandemic in Memory Organizations." *Heritage* 5, no. 4: 3616–3627.

A Network to develop the use of web archives: Three outcomes of the ResPaDon project

Sara Aubry, Audrey Baneyx, Emmanuelle Bermès, Laurence Favier, Alexandre Faye, Marie-Madeleine Géroudet, Benjamin Ooghe-Tabanou

Abstract: Web archives represent a huge opportunity for new types of research, offering possibilities for mining and analysis in many scientific disciplines. However, technical, legal, and methodological barriers can prevent researchers from using web archives in their work. The ResPaDon project (Network of partners for the analysis and exploration of digital data) aims to reduce the initial effort required from researchers to get access to web archive collections and understand them. It brings libraries and research teams together to think, experiment, and share practices, in order to analyze the current and potential uses of web archives, to experiment new ways of accessing and exploring corpora, and to issue recommendations about services, roles, skills, and tools..

Keywords: experimentation, services, datasprint, usage.

A highly transformative age requires new ways of doing research. In this context, web archives represent a huge opportunity for new types of research, offering possibilities for mining and analysis in many scientific disciplines. However, technical, legal, and methodological barriers can prevent researchers from using web archives in their research work. Among these barriers, the methodological cost of entry is the initial effort required from the researcher to get access to the collection and to understand the data available in the web archives. In France, the ResPaDon project (Network of partners for the analysis and exploration of digital data) works towards reducing this methodological cost of entry, by associating academic and national libraries, researchers, and librarians in a network of partners. This chapter presents three outcomes of the project, thus showing how our methodology leads to a better understanding of the place of web archives in the research process, new ways of exploring web archives, and new ideas of services, including teaching and learning activities.

Sara Aubry, Bibliothèque nationale de France (BnF), France, sara.aubry@bnf.fr, 0009-0009-6601-7546
Audrey Baneyx, Sciences Po Paris, France, audrey.baneyx@sciencespo.fr, 0000-0001-8500-9343
Emmanuelle Bermès, Ecole nationale des École nationale des chartes, France, emmanuelle.bermes@chartes.psl.eu, 0000-0002-7926-3027
Laurence Favier, Université de Lille, France, laurence.favier@univ-lille.fr, 0000-0002-5809-0788
Alexandre Faye, Bibliothèque nationale de France (BnF), France, alexandre.faye@bnf.fr, 0009-0007-8508-8492
Marie-Madeleine Géroudet, Université de Lille, France, marie-madeleine.geroudet@univ-lille.fr, 0000-0001-7816-0723
Benjamin Ooghe-Tabanou, Sciences Po Paris, France, benjamin.ooghe@sciencespo.fr, 0000-0001-7698-3507

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Sara Aubry, Audrey Baneyx, Emmanuelle Bermès, Laurence Favier, Alexandre Faye, Marie-Madeleine Géroudet, Benjamin Ooghe-Tabanou, *A Network to develop the use of web archives: Three outcomes of the ResPaDon project*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.12, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 113-126, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

1. About the ResPaDon project

The ResPaDon project, undertaken by the University of Lille and the National Library of France (BnF), received funding for two years from the GIS[1] CollEx-Persée. Initiated in early 2021, the ResPaDon partnership involved four libraries—the BnF, along with the three university libraries from Lille, Sciences Po, and Campus Condorcet—and two research laboratories: GERiiCOo (Groupe d'Études et de Recherche Interdisciplinaire en Information et COmmunication), an information science lab from the University of Lille, and the médialab at Sciences Po. The objective of bringing libraries and research teams together was to foster reflection, experimentation, and the sharing of practices related to web archives. The aim was to bridge the gap between the producers and users of the web archive collection, facilitated by academic libraries. The project reached its conclusion in April 2023, culminating with an international conference organized by the University of Lille.

From its inception, ResPaDon aimed to develop the analysis and exploration of collections as data (Padilla 2019), to equip researchers with computational tools for leveraging the digital collections of libraries, such as text and data mining capabilities. Following decades of digitization and digital legal deposit, the BnF and its partners successfully curated extensive digital collections that held significant research potential, especially in the humanities. Initiatives such as the CORPUS project (Moiraghi 2018; Stirling 2022) and the creation of the BnF DataLab, inaugurated in 2021 (Bermès 2019; Carlin 2021), gave rise to new services and collaborative approaches between librarians and researchers within the French national library. However, it became evident that their impact would remain limited unless they were distributed to a nationwide audience, far beyond the BnF premises, and in collaboration with other research libraries across France.

In our pursuit, amidst the massive amount of available materials, we decided to focus on web archives. Inherently plural and complex, web archives exemplify the challenges we encounter in curating digital collections. Due to the global nature of the web, these archives interconnect with all types of collections in libraries and across research disciplines. Their technical characteristics necessitate heuristic considerations, both in their construction and understanding. Finally, they bring to light the apparent contradictions of a legal framework that imposes access restrictions on content that was initially freely accessible (Stirling 2012), a situation that prompted prior initiatives at the BnF to create a national network of partners among legal deposit libraries in France (Aniesa and Bouchard 2017).

[1] Groupement d'Intérêt Scientifique, or Scientific Interest Group

The ResPaDon project emerged in response to these questions: its founding partners sought to dismantle the organizational, technical, legal, and methodological barriers hindering the use of web archives as a source in French laboratories and research teams.

To achieve this goal, the idea of a network was the cornerstone of ResPaDon. At the heart of the partnership, the collaboration between university libraries and a national library demonstrated the potential to build on a common professional culture, while maintaining different relationships with patrons and collections. One of the main tasks of the project was organizing a series of eight workshops to foster collaboration among information professionals and researchers, providing a platform to envision how web archives could be made more readily accessible throughout the country.

The project was organized into five work packages:

- WP1 Strategy: primarily focused on the workshop cycle,
- WP2 Understanding usage: dedicated to studying past and current uses of web archives and web materials,
- WP3 Capsule experimentation: an organizational and technical prototype of a web archive capsule at the University of Lille to provide secure access to the BnF's web archives,
- WP4 Live web and archived web: an experimental adaptation of a research web crawling tool to web archives, culminating in a week-long DataSprint co-organized by the Sciences Po médialab and the BnF DataLab,
- WP5 Planning, training and communication: a work package resulting in a series of events including the opening and concluding conferences.

The overarching goals of the project can be summarized as follows: to analyze the current and potential uses of web archives, to experiment with new ways of accessing and exploring web archives, and to provide recommendations about services, roles, skills, and tools.

2. The use of web archives for scientific research

Work package 2 of the ResPaDon project aimed at summarizing real and potential uses of web archives, with particular emphasis on researchers' practices. It was led by the GERiiCOlab at the University of Lille. The group approached its tasks from two key perspectives: firstly, feedback from the BnF, which analyzed the evolution of research projects associated with the digital legal deposit over the last 20 years, and secondly, an interview survey of researchers who built their relevant corpora without specific assistance from library professionals. This research aimed to identify and characterize the type of web sources that interest scholars, the collection methods they use, their expectations with regard to the corpora

they build, and their expressed requirements for processing tools essential to scientific or teaching contexts.

Based on feedback drawn from 20 projects spanning the last 20 years, a typology of projects involving the BnF's digital legal deposit was proposed. This study was built upon a previous survey from 2011 that comprised 15 interviews with researchers. This initial study had highlighted the necessity of web preservation, although the use of web archives was not significantly pronounced, and also revealed a growing interest in collaborative collection.

In 2022, the objective shifted to providing an extensive overview. The adopted approach involved identifying project descriptors then grouping them based on common characteristics, thereby defining 'ideal-types'. This method did not aim to uncover specific project mechanisms or links, but to create a broader inventory to characterize distinctive elements. Ultimately, 20 projects were analyzed using this approach and projected onto a timeline. We observed an increasing number of projects over time. Some projects were undertaken by individuals, others by organized research teams, but all shared a commonality: a multidisciplinary orientation towards humanities. In terms of the collections used, different categories of research emerged: some researchers were solely using the BnF's collection, others collaborated with the library on selection, and some mixed materials and resources from different origins.

Ultimately, five types of projects emerged, each detailed as ideal-types according to three criteria (a. public and objectives, b. service, organization, and duration, c. evaluation and improvement):

- **Punctual research:** carried out by an individual over a relatively short time span. The researcher freely accesses the web archive application to identify and evaluate content quality. The researcher works mostly autonomously.
- **Archiving and enrichment**: involves PhD students or scholars selected through a call process. It leads to long-term collaboration and results in valorization of the research through activities such as virtual guided tours ('parcours guidés').
- **Mining and exploitation**: teams of researchers engage in collective work, testing queries, and documenting datasets. Project selection includes a feasibility assessment.
- **Reference collection**: a laboratory aims to create a consistent collection related to their research topic. The produced corpus is indexed in full-text. The duration of the collaboration with the library can vary and there is often a need for increased visibility.
- **Production process**: a fully organized project with a team including IT skills. The project can lead to the production of a corpus and/or tools that may be standardized and distributed beyond the project stakeholders.

The GERiiCOlab complemented this typology through an interview survey with scholars in political science, sociology, literature, and history of the web who had not received assistance from library professionals.

An educational experiment was also conducted with a group of information science master students from the University of Lille. A teaching experience immersed them in research sessions particularly focusing on e-voting (with queries based on the French expression 'vote électronique' in the BnF web archive collection on 2002 elections) and femicides (using the term 'féminicide' in the BnF web archive news and media collection—collecte 'Actualité'). Students encountered challenges such as delineating the boundaries of web archives and distinguishing them from the live web. They also had difficulty understanding the typological difference between the web archives collected by the BnF and the 'archive' sections on certain websites, particularly those of news organizations. Choosing rigorous elements for analyzing web archives also posed problems due to the varying typology from one document to another, and their intersectionality with fields such as archives, documentation, and librarianship. Analyzing a website archived only on the surface (archived home page, but few—if any—pages available beyond the home page) raised issues in building web archive corpora. Moreover, students also faced challenges with the search interface, including searching by URL and understanding the outcomes of proposed tools.

Despite these difficulties, however, there were opportunities. Laurence Favier's focus on the BnF's News collection (collecte 'Actualité'), encompassing national and local press websites (pure player or cross media), aimed to identify researchers' needs. Capturing the full web environment of newspapers (including social networks) remains a challenge for web archive collections: from the production of new types of content (such as press blogs) to comments and, of course, links between other articles and new content. Web archives dedicated to newspaper websites actually compete with commercial news databases such as Europresse, for example. However, they offer researchers the ability to critique the live web and assess the current version of a website. It is a matter of authenticity: the web archive provides evidence when institutions decide what is outdated or or reliable on their websites.

Finally, the close interviews report touches upon the challenge of mixing research on the live web and on various propositions of archives/archiving. The ongoing evolution poses challenges to web archives processed within the legal deposit framework, yet it presents a real opportunity to share knowledge, tools, and methods for searching and building corpora. 'Web' and 'archives' actually interact in a variety of ways. The use of past and outdated content available online is not limited to institutional web archives such as those provided by the BnF, it may also include the online archives

section of a website, platforms managed by the community, or a private actor (including Internet Archive). Furthermore, researchers also conduct their own preservation actions, either in partnership with a heritage institution or by using online services to quickly save a page, such as Save Page Now or Archive Today.

To conclude, this segment of our project highlights several facets of long-term research primarily based on web materials. It emphasizes the simultaneous need for archives of the disappeared web and those of the live web, the successive investigations required to identify the source of content beyond the website, the need to collect materials from the entire web ecosystem (including social networks), the involvement of private companies in collecting the materials, the need to 'make an archive', i.e. to build reference corpora that can be consulted and updated over time. The difficulties when it comes to defining the contours of the web archive are accentuated, encompassing the temporal dimension of the materials and the tools for reading and collecting them. The shift from the collected 'source' to the constituted archive involves both epistemological and technical dimensions that turn web-based corpora into scientific objects whose methodology has yet to be constructed.

3. Building and studying corpora from the past web using Hyphe

One of the main focuses of the ResPaDon project was to provide the research community and information professionals with methods and tools designed for the creation, analysis, and dissemination of web corpora. In this perspective, the project's work package 4 involved a collaborative effort between the Sciences Po Library, médialab, and the National Library of France (BnF). Together, they organized, ran, and evaluated an experiment based on the use of the Hyphe web crawler on web archives.

Developed by Sciences Po médialab as open-source software, Hyphe[2] was designed to provide researchers and students with a research-oriented crawler for building and enriching website corpora (Ooghe-Tabanou 2018). The tool uses links between them to map web territories and allow for the study of community structures. Hyphe employs a step-by-step methodology that guides users in curating and defining 'webentities' in a granular and flexible manner by choosing single pages, subdomains, combinations of websites, and more. The pages beneath these entities are then crawled to extract outgoing links and part of the textual content. The webentities discovered through these links can then be manually selected and further explored to enrich the corpus in an iterative way. The corpus of webentities and interconnecting links can be viewed at all times in the form of a

[2] https://web.archive.org/web/20230120163329/https://github.com/medialab/hyphe

network and exported for cleaning and analysis in other tools such as Gephi. The outcome makes it possible to study online communities and social phenomena, identify and group web actors, and explore the links between them.

As part of the ResPaDon project, Hyphe underwent extensions to work with the Past Web. The 'Archives de l'internet'—the BnF web archive search application—and Hyphe are now compatible. Hyphe's code was also modified to enable its operation on the Internet Archive's Wayback Machine as well to facilitate the completion and/or comparison of past web corpora. The following developments were made:

- ability to crawl and explore a web archive using archival URLs or permalinks[3]
- ability to target a specific date and crawl documents within a defined time period, making sure not to drift over time[4] and target a specific date and time period around this date;
- support both the opened archival URL mode used by the Wayback Machine and the closed proxy mode used by the BnF's Archives de l'internet
- support the option of either building corpora from web archives only, or combining data from both the live web and the web archives within the same instance, even the same corpus. This flexibility enables the curation of both the live web and archived web, through the BnF's and the Internet Archive's web archives.

In April 2022, Sciences Po and the BnF co-organized a one-week event known as a 'DataSprint' (Venturini 2018), convening teams of researchers, engineers, designers, web archivists, and collection specialists within the BnF DataLab, a dedicated space and service for the development of Digital Humanities at the BnF. The purpose of a DataSprint is to gather complementary skills and expertise from a variety of disciplines (political and social sciences, web archiving, engineering, design, digital methods, etc.) and devote several full days to working collaboratively in small groups. The timeframe of a full week allowed participants to delve into data to explore hypotheses on specific research questions and experiment with different methodological approaches to gain preliminary insights. Through this experiment, participants aimed to assess whether the corpora building and curation software used on the live web could also operate on corpora from the archived web. Additionally, they sought to determine the feasibility of a comparative approach between the two.

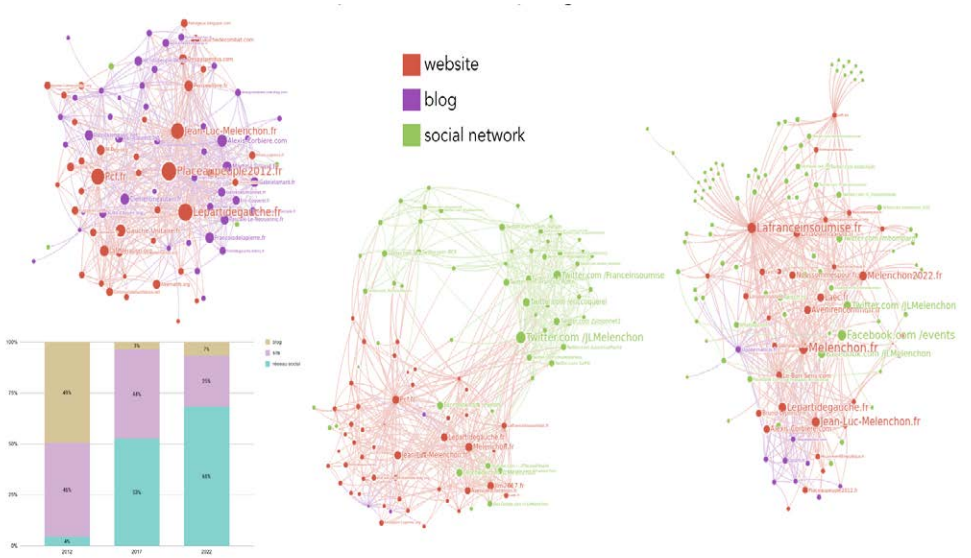
[3] Such as http://archivesinternet.bnf.fr/20170401090112/http://www.lours.org/

[4] Time-drifting happens when a user (or, in that case, the Hyphe crawler) navigating through a web archive follows a hyperlink towards a resource that has been captured at a different date. If the target of the link is not present in the archive for the same capture date, the Wayback Machine will try to replace it with a capture from another date, thus taking the user to a period of time possibly far from their initial request.

From April 4 to 8, 2022, twenty-five people worked within the BnF's DataLab reading rooms to explore and compare resources on the live web and web archives using Hyphe across four themes: political campaigns, COVID-19, theater arts critique, and genomics legal policies. The event resulted in the creation of a website[5] containing the DataSprint results with descriptions of the methodological tracks pursued, analyses conducted, and data visualizations generated.

The group working on the "Evolution of online political campaign structures" studied the structural transformations of online communities supporting candidates in the French presidential elections of 2012, 2017, and 2022, with a specific emphasis on Jean-Luc Mélenchon's campaigns. To facilitate a comparative approach, the group first defined the appropriate time windows based on archive availability. Coherent crawls were then initiated from persistent sources (Wikipedia pages), and common rules were established to select actors consistently across each snapshot. Finally, they devised a classification system to tag actors based on their editorial form and nature. With the corpora for the three consecutive presidential elections built, the challenge shifted to how to compare the compositions and structures of graphs. Digital methods were explored to visualize differences between two graphs, enabling the identification of disappeared actors, newcomers, and resilient ones. This allowed the group to compare the 2012, 2017, and 2022 web actors of Mélenchon's campaigns and their editorial forms, highlighting the subnetworks of websites involved in both the 2017 and 2022 campaigns and those exclusive to one of them. The compared graphs underscore the disappearance of blogs between 2012 and 2017, with the rise of social networks, initially forming as a separate network in 2017 and becoming deeply intertwined with campaign websites in 2022.

[5] https://web.archive.org/web/20231014053626/https://respadon.medialab.sciencespo.fr/



Comparison of the 2012, 2017 & 2022 web actors of Mélenchon's campaigns and their editorial forms

Another exploration during the DataSprint focused on the interactions between web officials and others on COVID-19. In this case, the comparative approach was based on three snapshots: 2020 (archives), 2021 (archives), and early 2022 (live web). The crawls were initiated from existing corpora built by collection curators and the group reused existing classifications to tag web entities by actor types (institutions, health experts, blogs, unions, NGOs, media, etc.). Their objective was to visualize the evolving position of an actor across the three graphs and the evolving proximity between actor types. They thus identified actors appearing, disappearing, or remaining over time, and explored how the positions of various actor types converge or diverge.

Methodological questions and future perspectives arising from this research include various uses for web archives in corpus building. These include longitudinal comparisons of time snapshots from the live web or archives, as well as the mixing of archives and live web to enhance corpora. Practical, technical, and systemic constraints have also surfaced such as the time-consuming nature of iterative corpus building and the complex reproduction process it entails. In this context, archive completeness and frequency is often an issue, and Hyphe has been considered a heuristic tool to evaluate it. Technical challenges include the permanent restructuring of the web (disappeared, repurposed, and redirected websites to new domain names), the resilience of links across time, and the complexities of modern full-JavaScript-based websites that are not yet crawlable with Hyphe. Building on these findings, Sciences Po médialab intends to pursue this work by connecting Hyphe to more existing web archives. Ongoing

collaboration with the INA (Institut national de l'audiovisuel) in France is already underway, and discussions with the national archives of other countries, such as Arquivo.pt in Portugal, have commenced following their expressed interest after the conference presentation.

4. Project results: main conclusions and recommendations

Circling back to the series of eight workshops organized within ResPaDon's work package 1, we can see how these ongoing research experiments on the use of web archives have fuelled a broader reflection on designing services tailored to academic needs. The workshops convened information professionals, researchers, and other stakeholders (engineers, lawyers, training organizations, etc.) to conceive how web archives could be made more accessible and usable for researchers. Each workshop session was organized around feedback presentations and discussions, covering eight topics throughout the series: access, usage, services, legal challenges, training, methods, the role of local and national players, and the creation and documentation of corpora. The outcome was a set of principles and recommendations regarding web archive access in higher education institutions, which were presented at the closing event in Lille in April 2023. The workshop findings were organized into a heuristic map and prioritized into potential actions based on five principles, as follows:

Principle 1. Given the unique nature of the web, the scientific study of its content necessitates the creation of an archive.

- Action 1. Support the definition and dissemination of methods for studying web sources to serve research.
- Action 2. Standardize the methodology for creating, documenting, and citing a web archive.
- Action 3. Integrate web sources into the development of digital literacy and culture for students, researchers, and professionals.

Principle 2. Web archives should serve as one of many sources for research.

- Action 4. Include web sources, archives, and the living web in the development of research practices and in opening up research processes and results.
- Action 5. Promote the exposure of metadata and the implementation of mechanisms for the discoverability of web archives.
- Action 6. Enable the discovery of web archives through browsing or exploring the living web.

Principle 3. Web archives should be usable autonomously for a variety of audiences.

- Action 7. Facilitate access to and reuse of web archives by changing current regulatory conditions.

- Action 8. Establish and sustain web archive access points in higher education and research establishments.
- Action 9. Facilitate the collective enrichment of an open-access sandbox for educational and research purposes.

Principle 4. A national network of researchers and information and library professionals is an essential catalyst for developing the use of web sources.

- Action 10. Unite actors interested in exploiting web sources and producing their archives around well-established nodes.
- Action 11. Develop various activities within these nodes to support access to and use of web sources, including awareness-raising and training, collaborative collections among partners, and remote access to web archives.
- Action 12. Implement co-leadership of the network by the nodes and national actors, involving shared documentation, regular meetings, and exchanges of practices.

Principle 5. The mediation of web sources by multiple actors requires the development of new skills.

- Action 13. Develop the skills of mediators within the nodes of the network.
- Action 14. Develop the skills of engineers supporting research projects based on web sources.
- Action 15. Develop collaborative collections by working with researchers and information and library professionals to encourage the acquisition of these new skills.

These recommendations represent just one facet of the project's deliverables. Notably, this chapter does not delve into the findings of work package 3, the experimentation capsule deployed in two locations within the university library in Lille. This real-life proof of concept, supporting the observations of researchers and students in work package 2, provided a comprehensive set of conclusions in terms of technical issues to secure access to web archives, organizational hurdles to overcome, as well as the need for documentation, training, communication tools, and support for these capsules to fulfill their role. Our recommendations also incorporate the results of this experiment as they seek to address the question: how can access to web archives be improved for researchers?

While ResPaDon's recommendations are meant to be realistic, we acknowledge that they are only drawing an ideal horizon, necessitating further consolidation in terms of resources. The partners have already convened to envision the next steps, including expanding the network to include new partners and addressing sustainability concerns for the service.

5. Conclusion

By compelling us to translate our concerns across the professions to identify our commonalities and our differences, the ResPaDon project raised questions so fundamental that they are often overlooked. Firstly, are web archives really ‘archives’? While the term is well established among librarians, it remains confusing for potential users who encounter a methodological gap when engaging with this reborn material (Brügger 2018). It is now more critical than ever to open up and extend this source and envision its use in conjunction with other forms of digital and analog material, particularly with the live web, which remains a vital and complementary entry point. For students and researchers, entering the realm of web archives is a step-by-step process: starting with an exploration of the content readily available on the web and progressing towards more complex tools such as text mining and link mapping. Seizing the low-hanging fruits involves providing clear and straightforward information on the library website, collaborating on seed selections, building sandboxes, or devising discreet ways of providing training.

Ultimately, this project has inspired us to envision the ideal training sessions for different audiences, fostering a vast and close-knit community, dismantling legal barriers, and positioning web archives as an element of digital culture like any other, a resource that is natural rather than mandatory for researchers in humanities. As we look ahead, we hope to pursue our efforts in the years to come and continue building this dream together.

References

Aniesa, Ange, and Ariane Bouchard. 2017. “Constituer un réseau d’accès aux archives de l’internet : l’exemple français.” https://library.ifla.org/id/eprint/1655/.

Bermès, Emmanuelle. 2019. “Quand le dépôt légal devient numérique : épistémologie d’un nouvel objet patrimonial.” *Quaderni. Communication, technologies, pouvoir*, no 98: 73–86. https://doi.org/10.4000/quaderni.1455.

Bonnel, Sylvie, and Clément Oury. 2014. “Selecting websites in an encyclopaedic national library: a shared collection policy for internet legal deposit at the BnF.” Paper presented at IFLA WLIC 2014 – Lyon – Libraries, Citizens, Societies: Confluence for Knowledge. Lyon, France, 2014. https://library.ifla.org/id/eprint/998/.

Brügger, Niels. 2018. *The Archived Web: Doing History in the Digital Age*. Cambridge (Mass.): MIT Press.

Brügger, Niels, and Ian Milligan. 2018. *The SAGE Handbook of Web History*. Newbury Park, California: Sage.

Carlin, Marie, and Arnaud Laborderie. 2021. “Le BnF DataLab, un service aux chercheurs en humanités numériques.” *Humanités numériques*, no 4. https://doi.org/10.4000/revuehn.2684.

Gebeil, Sophie. 2021. *Website story: histoire, mémoires et archives du web*. Bry-sur-Marne: INA.

Gomes, Daniel, Elena Demidova, and Jane Winters. 2021. *The Past Web: Exploring Web Archives*. Switzerland: Springer Nature.

Harris, Kayla, Christina A. Beis, and Stephanie Shreffler. 2021. “Citizen Web Archivists: Applying Web Archiving as a Pedagogical Tool.” *Journal of Electronic Resources Librarianship* 33, no 4: 262–72. https://doi.org/10.1080/1941126X.2021.1988463.

Milligan, Ian. 2020. “You shouldn’t Need to be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure.” *WARCNet papers*, Aarhus, Denmark. https://cc.au.dk/fileadmin/user_upload/WARCnet/Milligan_You_shouldn_t_Need_to_be__2_.pdf.

Moiraghi, Eleonora. 2018. “Le projet Corpus et ses publics potentiels.” Rapport public. Bibliothèque nationale de France. https://hal-bnf.archives-ouvertes.fr/hal-01739730.

Musiani, Francesca, Camille Paloque-Bergès, Valérie Schafer, and Benjamin G. Thierry. 2019. *Qu’est-ce qu’une archive du web ?* Marseille: OpenEdition.

Ooghe-Tabanou, Benjamin, Mathieu Jacomy, Paul Girard, and Guillaume Plique. 2018.

"Hyperlink is not dead!" *Proceedings of the 2nd International Conference on Web Studies*, 12–18. https://doi.org/10.1145/3240431.3240434.

Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner. 2019. "Always Already Computational: Collections as Data." https://zenodo.org/records/3152935.

Stirling, Peter. 2022. "Le dépôt légal de l'internet dans le projet CORPUS." Billet. *Web Corpora* (blog), 29 August 2022. https://webcorpora.hypotheses.org/111.

Stirling, Peter, Gildas Illien, Pascal Sanz, and Sophie Sepetjan. 2012. "The State of E-Legal Deposit in France: Looking Back at Five Years of Putting New Legislation into Practice and Envisioning the Future". *IFLA Journal* 38, no 1: 5–24. https://doi.org/10.1177/0340035211435323.

Venturini, Tommaso, Anders Munk, and Axel Meunier. 2018. "Data-sprinting: A public approach to digital research." In *Routledge Handbook of Interdisciplinary Research Methods*. Routledge. https://doi.org/10.4324/9781315714523-24.

SECTION 4
Web archive as a material for uncovering web history

Time, bits, and nickel: Managing digital and analog continuity

Julie Momméja

Abstract: In 1998, the Getty Center hosted the "Time and Bits: Managing Digital Continuity" conference, gathering the founders and thinkers of two San Francisco non-profit organizations interested in long-term thinking and archiving: the Internet Archive and the Long Now Foundation. This chapter proposes to discuss two different ways of archiving through time, in digital and analog formats, for virtual web contents and physical paper-based ones. It explores various types of archiving methods and tools and the management challenges they raise, in terms of time and space, but also innovation, maintenance, and "continuity". It depicts two distinct visions of the future of archiving which nonetheless converge in their mission of safeguarding, sharing, and giving access to information and knowledge for the decades and centuries to come..

Keywords: archives, digital, analog, longue durée, future.

Introduction

"One of the peculiar things about the 'Net is it has no memory. (...) We've made our digital bet. Civilization now happens digitally. And it has no memory. This is no way to run a civilization. And the Web—its reach is great, but its depth is shallower than any other medium probably we've ever lived with." (Kahle 1998)

In February 1998, the Getty Center in Los Angeles hosted the Time and Bits: Managing Digital Continuity conference, organized by the founders and thinkers of two non-profit organizations established two years earlier in San Francisco: the Internet Archive and the Long Now Foundation, both dedicated to long-term thinking and archiving.

The Internet Archive has since become a global example of digital archiving and an open library that provides access to millions of digitized pages from the web and paper books on its website. The Long Now Foundation's central project involves the design and construction of a monumental clock intended to tick for the next 10,000 years, promoting long-term thinking alongside its lesser-known archival mission: the Rosetta Project.

The Time and Bits: Managing Digital Continuity conference brought together these organizations to discuss what the Internet Archive's founder, Brewster Kahle, referred to as "our digital bet": a digital-only novel form of civilization with no history, posing challenges to the preservation of its immaterial cultural memory. Discussions at the conference raised concerns about the longevity of digital formats and explored potential archival and

Julie Momméja, University of Lorraine, France, julie.mommeja@univ-lorraine.fr, 0000-0003-1148-2490

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

 DOI 10.36253/979-12-215-0413-2.14, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 129-139, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

transmission solutions for the future. This foresight considered the 'heritage' characteristics of the digital world, which were yet to be defined as such on a global level. It was not until 2003 that UNESCO published a charter advocating for "digital heritage conservation", distinguishing between "digital-born heritage" and digitized heritage (UNESCO 2003, Musiani et al. 2019). The Getty Center conference thus emerges as a precursor in the quest to preserve both digital data and analog information for future generations. This endeavor, the chapter argues, aligns with the *longue durée*, a conception of time and history developed by French historian Fernand Braudel during World War Two (Braudel 1958).

While the Internet Archive has envisioned such a mission through the continuous digital recording of web pages and the digitization of paper, sound, and video documents into bits format, the Long Now Foundation's Rosetta Project began to take shape during the 'Time and Bits' conversations, offering a different approach to data conservation in an analog microscopic format, engraved on nickel disks.

Taking the 1998 gathering of the Internet Archive and Long Now Foundation as a starting point, this chapter aims to examine the challenges and strategies of 'digital continuity management' (or maintenance). It proposes to analyze the different ways these two case studies envision archiving and transmission to future generations, in both digital and analog formats—bits and nickel, respectively—for virtual web content and physical paper-based materials.

Through a comparative analysis of these two non-profit organizations, this chapter seeks to explore various archiving methods and tools, and the challenges they present in terms of time, space, innovation, maintenance, and 'continuity'. By depicting two distinct visions of the future of archiving represented by these organizations, it highlights their shared mission of safeguarding, sharing, and providing universal access to information, despite their differing formats.

The method used for this analysis combines theoretical, comparative, and qualitative studies through an immersive research process spanning over three years in the San Francisco Bay Area. This process involved conducting interviews and participant observations during seminars, talks, and meetings held by both the Long Now Foundation and at the Internet Archive. Additionally, archival research was conducted online, using resources such as the Internet Archive's Wayback Machine and the Long Now Foundation's blog dating back to 1996, as well as on-site at the Long Now Foundation.

This chapter adopts a multidisciplinary approach conjoining history, media, maintenance, and American studies to analyze the challenges faced by these two organizations in transmitting both material and intangible cultural heritage (UNESCO 2003).

The first part of this chapter concentrates on the future of archives and their longevity, a topic that was discussed during the 1998 Time and Bits conference. It suggests a parallel with the Braudelian *longue durée* perspective, which offers a novel understanding of time and history.

The second part focuses on digital transmission in 'hard-drive form' using the example of the Internet Archive and its Wayback Machine, comprising thousands of hard drives. The final segment of this chapter discusses the analog archival format chosen by the Long Now Foundation, represented by the Rosetta Disk, a small nickel disk engraved with thousands of pages of selected texts. This format is likened to a modern iteration of the Egyptian Rosetta Stone for preservation into the long-term future.

1. "Time and Bits: Managing Digital Continuity"…and maintenance in *longue durée*

"How long can a digital document remain intelligible in an archive?" This question, asked by futurist Jaron Lanier in one of the hundreds of messages posted on the Time and Bits forum that ran from October 1997 until June 1998, underscores not only concerns about the future 'life' of digital documents at the end of the 1990s, but also their meaning and understanding in archives for future generations. These concerns about digital preservation were central to discussions at the subsequent Time and Bits conference organized a few months later in February 1998 at the Getty Center in Los Angeles by the Getty Conservation Institute and the now-defunct Getty Information Institute, in collaboration with the Long Now Foundation.

The Long Now Foundation, formed in 1996, emerged from discussions among thinkers and futurists who later became its board members This group included Stewart Brand, recipient of the 1971 National Book Award for his *Whole Earth Catalog* and co-founder of the pioneering virtual community, the WELL, created in the 1980s (Turner 2006), engineer Danny Hillis, British musician and artist Brian Eno, technologists Esther Dyson and Kevin Kelly, and futurist Peter Schwartz (Momméja 2021). Schwartz, in particular, is the one who articulated the concept of the 'long now' as a span of 20,000 years—10,000 years deep into the past and 10,000 years into the very distant future. This timeframe coincides with the envisioned lifespan of the monumental Clock being constructed by the foundation in West Texas. The choice of this specific duration marks the end of the last ice age about 10,000 years ago, a period that catalyzed the advent of agriculture and human civilization, with some scholars even identifying it as the onset of the Anthropocene epoch. Indeed, a group of scientists extends their analysis beyond the industrial era, which has generally been studied as the beginning of this human-induced transformation of our biosphere,

considering the origins of agriculture as "the time when large-scale transformation of land use and human-induced species and ecosystem loss extended the period of warming after the end of the Pleistocene" (Henke and Sims 2020). For the founders of the Long Now Foundation, this 10,000-year perspective must therefore be developed in the opposite direction, towards the future (hence the expected duration of the Clock) forming the 'Long Now'.

The paper argues that 'long now' promoted by the organization can be paralleled with the concept of *longue durée* put forth by Annales historian Fernand Braudel. Braudel began elaborating the idea of *longue durée* during his time as prisoner of war in Germany. For five years, he diligently worked on his PhD dissertation, *La Méditerranée et le monde méditerranéen à l'époque de Philippe II* (Braudel 1949). It was during his internment that Braudel developed the concept of the 'very long time', a temporal construction that provided him solace from the traumatic events he experienced in the 'short time' and helped him gain insight into his condition by situating them on a much broader time scale (Braudel 1958). With newly stratified temporalities—from the immediate to the medium to the very long term—Braudel succeeded in escaping the space-time of which he was a prisoner, a 'here' and 'now' devoid of meaningful perspectives when a longer 'now' would liberate him from the present moment. *Longue durée* was thus imagined as a novel long-term approach to history, diverging from traditional narratives that focused on brief periods and dramatic events, such as wars. This is what Braudel referred to as "a rushed, dramatic narrative" (Braudel 1958). A second, longer type of history, based on economic cycles and conjunctures, was described by Braudel as spanning several decades, while *longue durée* offered a novel type of history that transcended events and cycles, extending even further to encompass centuries—although the French historian refrained from specifying an exact timeframe.

Longue durée, alongside its modern Californian counterpart, the 'long now', prompts us to reconsider our understanding of history in time as a means to encapsulate events far beyond our lifetimes. Braudel insisted historians should incorporate *longue durée* into their work and rethink history as an 'infrastructure' composed of layers of 'slow history'.

Given this perspective, how can we archive and transmit fast traditional history within the context of *longue durée*? In his foreword to the Time & Bits report, Barry Munitz, president and CEO of the J. Paul Getty Trust, explained the initiative behind the conference:

> We take seriously the notion of long-term responsibility in the protection of important cultural information, which in many cases now is recorded only in digital formats. The technology that enables digital conversion and access is a marvel that is evolving at lightning speed. Lagging far behind, however, are the means by which the digital legacy will be

preserved over the long term (Munitz 1998).

The two organizations selected for this chapter offer two distinct, yet complementary, visions of how archiving and transmitting should be approached, now and for the *longue durée*, in digital and analog formats.

2. Digital transmission in 'hard-drive form': The Wayback Machine and the Internet Archive

Addressing the "problem of our vanishing memory" was a focal point of the Time & Bits conference encapsulated by Internet Archive founder Brewster Kahle's question: "I think the issue that we are grappling with here is now that our cultural artifacts are in digital form, what happens?" (Kahle 1998). As noted by Stewart Brand, Kahle also pointed out that "one of the peculiar things about the 'Net is it has no memory. (…) We've made our digital bet. Civilization now happens digitally. And it has no memory. This is no way to run a civilization. And the Web—its reach is great, but its depth is shallower than any other medium probably we've ever lived with" (Kahle 1998).

As a way to resolve this 'digital bet' and the pressing need for 'digital continuity', Brewster Kahle embarked on a mission to archive the web on a massive scale, giving rise to the Internet Archive and its Wayback Machine: an archive comprising 20,000 hard drives and containing 866 billion web pages as of March 2024.

Like the Long Now Foundation, the Internet Archive is a non-profit organization founded in 1996 in San Francisco. In fact, both entities once occupied adjacent offices in the Presidio. Their missions can also be put in parallel: whereas the Long Now Foundation promotes long-term thinking through projects like the construction of a Clock and the preservation of foundational languages and texts of our civilization in analog form through the Rosetta Disk, the Internet Archive digitizes and archives analog documents and records digital textual heritage through its Wayback Machine.

The Internet Archive embarked on its mission with an imperative to save internet pages, immaterial data composed of bits, which had not previously been archived: "We began in 1996 by archiving the internet itself, a medium that was just beginning to grow in use. Like newspapers, the content published on the web was ephemeral—but unlike newspapers, no one was saving it" (Internet Archive 2024). Despite the transient and intangible nature of web pages, the Internet Archive remains committed to this mission, continuing to archive internet pages in a digital format to this day, with the ambition to remain open and collaborative, "explicitly promoting bottom-up initiatives intended to revalue human intervention" (Musiani et al. 2019).

Brewster Kahle, who could be regarded as the first digital librarian in

history, promotes “Universal Access to All Knowledge” and “Building Libraries Together”. These missions, as explained during the Internet Archive's annual celebration on October 21, 2015, at its headquarters in San Francisco, highlight the organization’s commitment to a wide array of digital content, including internet pages, books, videos, music, and games. Therefore, the internet appears as a “heritage and museographic object” (Schafer 2012), with information worth saving and protecting for the future. While the Library of Congress recently acknowledged the significance of Twitter content as a form of heritage (Schafer 2012), the Internet Archive has been standing as an advocate for the preservation and transmission of digital heritage as early as the 1990s. UNESCO further validated this recognition in 2003 by acknowledging the existence of “digital heritage as a common heritage” through a charter on the conservation of digital heritage (Musiani et al. 2019) where resources are ‘born digital’, before being, or even without ever being, analog:

> Digital materials encompass a vast and growing range of formats, including texts, databases, still and moving images, audio, graphics, software, and web pages. Often ephemeral in nature, they require purposeful production, maintenance, and management to be retained. Many of these resources possess lasting value and significance, constituting a heritage that merits protection and preservation for current and future generations. This ever-growing heritage may exist in any language, in any part of the world, and in any area of human knowledge or expression (UNESCO 2003).

The Internet Archive’s mission aligns perfectly with this definition, providing open access to documents that are "protected and preserved for current and future generations”, echoing once again the Long Now Foundation’s own mission. However, the pursuit of "universal access to all knowledge" raises questions about the quality or "representativeness of the archive" (Musiani et al. 2019) in the face of the abundance and diversity of the sources and formats available.

For instance, the music section of the Internet Archive connects visitors to San Francisco’s local counterculture history with a vast collection of recordings from Grateful Dead shows (17,453 items) that fans contributed to the organization in analog formats for digitization. This exchange has not only allowed the band’s fan community to flourish but has also bolstered the group’s the popularity: “they started to record all those concerts and you know, there are I think 2,339 concerts that got played by the Grateful Dead (…) and all but 300 of those are here in the archive” (Barlow 2015). In this way, the Internet Archive confirms its role as a universal collaborative platform and effectively contributes to a “new era of cultural participation” (Severo and Thuillas 2020), one that is proper to Web 2.0 but which the non-profit has been championing since the 1990s.

However, for the Internet Archive, and digital technology in general, to truly guarantee the archiving of human heritage ‘for future generations’

over the years, whether initially analog or digital, it is imperative to continuously improve and update storage formats and units to combat obsolescence and adapt to evolving technologies:

> Of course, disk drives all eventually fail. So we have an active team that monitors drive health and replaces drives showing early signs for failure. We replaced 2,453 drives in 2015, and 1,963 year-to-date 2016… an average of 6.7 drives per day. Across all drives in the cluster the average 'age' (arithmetic mean of the time in-service) is 779 days. The median age is 730 days, and the most tenured drive in our cluster has been in continuous use for 6.85 years! (Gonzalez 2016)

If "all contributions produced on these platforms, whether amateur or professional, participate in the construction and appropriation of cultural and memorial heritage" (Severo and Thuillas 2020), reliance solely on digital technology poses a substantial challenge to the preservation of our cultures in *longue durée*. Aware of the inherent risks associated with archiving both analog and 'digital heritage' on storage mediums with limited lifespans, the Internet Archive must make the maintenance and replacement of the hard drives that comprise its Wayback Machine a constant priority.

3. From stone to disk: the Rosetta Project through time and space

To embody Braudel's notion of 'slow history' and foster long-term thinking among people, the Long Now Foundation envisioned not only a monumental Clock as a time relay for future generations, but also a library for the deep future, soon materializing as an engraved artifact: The Rosetta Disk.

As explained by technologist Kevin Kelly, the concept of a miniature storage system comprising 350,000 pages of text engraved on a nickel disk, measuring just under eight centimeters in diameter, was proposed by Kahle during the Time and Bits: Managing Digital Continuity conference, "as a solution for long-term digital storage (…) with an estimated lifespan of 2,000–10,000 years" (Kelly 2008). These meeting discussions thus led to the emergence of the Rosetta Project within the Long Now Foundation, drawing inspiration from the Rosetta Stone. The final version of the Rosetta Project's Disk was unveiled in 2008: 14,000 pages of information in 1,500 different languages (Welcher 2008). Crafted in analog format, it was conceived as the solution to the ever-changing landscape of digital technologies.

While the Internet Archive possesses infinite possibilities for archiving, the Long Now Foundation's analog choice demands a thoughtful selection of texts to be micro-engraved onto the disk. The foundation decided to focus on several texts, both symbolic and universalist, such as the 1948 Universal Declaration of Human Rights, along with Genesis, chosen for its numerous

translations. Materials with a linguistic or grammatical vocation, such as the Swadesh list—a compendium of words establishing a basic lexicon for each language—were included, as well as grammatical information including descriptions of phonetics, word formation, and broader linguistic structures like sentences.

Unlike Kahle's digital and digitized heritage project, the Foundation's language archive is exclusively engraved, accessible only through a microscope. Such an archive is thus a finite heritage, with no scope for future development beyond the creation of new disks displaying new texts. While the Internet Archive and its Wayback Machine are constantly evolving, updated through constant digitization and the preservation of new web pages, the format and size of the nickel disk remain immutable.

To ensure the long-term survival of this archive, the foundation has embraced the "LOCKS" principle—Lots of Copies Keep Stuff Safe—and has opted to duplicate its Rosetta Disk. By distributing these duplicates worldwide, the project stands a greater chance of lasting in *longue durée*: "this project in long-term thinking would do two things: it would showcase this new long-term storage technology, and it would give the world a minimal backup of human languages" (Kelly 2008).

The final version of the Rosetta Disk, containing 14,000 micro-engraved pages, was presented at the Foundation's headquarters in 2008. "Kept in its protective sphere to avoid scratches, it could easily last and be read 2,000 years into the future" (Welcher 2008). Beyond its resilience within the timeline of the Long Now, the analog Rosetta Disk aspires to endure across space as well. Remarkably, as the Foundation had been developing its project since 1999, they were contacted by the European Space Agency (ESA) and the Rosetta Mission team which, coincidentally, was working on the launch of an exploratory space probe aptly named Rosetta. The Rosetta probe was launched on March 2, 2004, aboard an Ariane 5G+ rocket from Kourou, with the mission of studying comet 67P/Churyumov-Gerasimenko ('Tchouri') located near Jupiter. On board the probe was the very first version of the Rosetta Disk, less comprehensive than the version unveiled in 2008, nevertheless containing six thousand pages of translated texts.

Conclusion

On November 12, 2014, over a decade after its departure from Earth, the Rosetta probe finally reached Comet Tchouri. Upon arrival, it deployed its Philae lander onto the comet's surface, where, despite unexpected rebounds, it eventually stabilized itself to conduct programmed analyses. Nearly two years later, on September 30, 2016, the Rosetta module, with the Rosetta Disk on board, joined Philae on Tchouri, thus marking the conclusion of the mission: "With Rosetta we are opening a door to the origin of planet Earth

and fostering a better understanding of our future. ESA and its Rosetta mission partners have achieved something extraordinary today" (ESA 2014). Through a space mission focused on the future with the aim of better understanding the Earth's past, the Rosetta Disk fulfilled its project to become an archive in *longue durée*, transcending temporal and spatial boundaries.

Almost ten years later, both the Rosetta Disk and the Internet Archive, through a selection of books and documents from its datasets, became part of an even larger spatial archive which also includes articles from Wikipedia and books from Project Gutenberg, all etched on thin sheets of nickel. The Arch Mission Foundation's Lunar Library successfully landed on the Moon on February 22, 2024, thus reuniting for the first time the two non-profits' archival materials in a cultural and civilizational preservation project, built to remain on the Moon surface throughout the *longue durée*.

The Time and Bits: Managing Digital Continuity conference did not present a single solution to the challenges of digital archives and data transmission. Instead, it offered a range of options and tools for web archives, digital data, and analog documents to address our 'digital bet'. The two cases presented appear as two faces of the same disk—digital and analog—with a shared conservation objective: providing different means to consider *longue durée* and ensure archival continuity and maintenance in the long term. This continuity extends not only through time, but also across space, placing "digitally-born heritage" (Musiani et al. 2019) and more traditional forms of heritage on equal footing.

From the "creative city" (Florida 2002) of San Francisco, both organizations have managed to extend the boundaries of the "creative Frontier" (Momméja 2001), not only physically and digitally, but also through *longue durée* and space. From hard drives to disks, they offer a new form of coevolution between humans and machines, a 'post-coevolution' aimed at transmitting our cultural heritage to future generations through bits and nickel.

References

Brand, Stewart. 1999. *The Clock of the Long Now: Time and Responsibility*. New York: BasicBooks.

The European Space Agency. 2002. "Rosetta Disk Goes Back to the Future." The European Space Agency. December 3. https://web.archive.org/web/20240423005130/https://sci.esa.int/web/rosetta/-/31242-rosetta-disk-goes-back-to-the-future.

———. n.d. "Enabling & Support – Rosetta." The European Space Agency. https://web.archive.org/web/20240423005544/https://www.esa.int/Enabling_Support/Operations/Rosetta.

———. n.d. "Rosetta – Summary." The European Space Agency. https://web.archive.org/web/20240423004357/https://sci.esa.int/web/rosetta/2279-summary.

———. n.d. "Where Is Rosetta?" The European Space Agency. https://web.archive.org/web/20240423003218/https://sci.esa.int/where_is_rosetta/.

Florida, Richard. 2002. *The Rise of the Creative Class: And How It's Transforming Work, Leisure, Community and Everyday Life*. New York, NY: Basic Books.

Gonzalez, John. 2016. "20,000 Hard Drives on a Mission." Internet Archive Blogs. October 25. https://web.archive.org/web/20240423002926/https://blog.archive.org/2016/10/25/20000-hard-drives-on-a-mission/.

Henke, Christopher R, and Benjamin Sims. 2020. *Repairing Infrastructures the Maintenance of Materiality and Power*. https://web.archive.org/web/20240423002248/https://direct.mit.edu/books/oa-monograph/4962/Repairing-InfrastructuresThe-Maintenance-of.

Internet Archive. 2015. "Building Libraries Together, Celebrating the Passionate People Building the Internet Archive." Internet Archive, San Francisco, October 21. https://archive.org/details/buildinglibrariestogether2015.

Internet Archive. 2024. "About the Internet Archive." Internet Archive. https://web.archive.org/web/20240423001744/https://archive.org/about/.

Kahle, Brewster. 2011. "Universal Access to All Knowledge." San Francisco, November 30. https://web.archive.org/web/20240423001555/https://longnow.org/seminars/02011/nov/30/universal-access-all-knowledge/.

Kahle, Brewster. 2016. "Library of the Future." University of California Berkeley,

Morrison Library, March 3. https://web.archive.org/web/20240423001315/https://bcnm.berkeley.edu/events/109/special-events/1004/library-of-the-future.

Kelly, Kevin, Alexander Rose, and Laura Welcher. “Disk Essays.” The Rosetta Project. https://web.archive.org/web/20240422235700/https://rosettaproject.org/disk/essays/.

Kelly, Kevin. 2008. “Very Long-Term Backup.” The Long Now Foundation. August 20. https://web.archive.org/web/20240423000131/https://longnow.org/ideas/very-long-term-backup/.

The Long Now Foundation. “Time and Bits: Managing Digital Continuity.” 1998. February 8. https://web.archive.org/web/20240423001231/https://longnow.org/events/01998/feb/08/time-and-bits/.

MacLean, Margaret G. H., Ben H. Davis, Getty Conservation Institute, Getty Information Institute, and Long Now Foundation, eds. 1998. “Time & Bits: Managing Digital Continuity.” [Los Angeles: J. Paul Getty Trust].

Momméja, Julie. 2021. “Du Whole Earth Catalog à la Long Now Foundation dans la Baie de San Francisco : Co-Évolution sur la “Frontière” Créative (1955–2020).” Paris: Paris 3 – Sorbonne Nouvelle. https://theses.fr/2021PA030027.

Musiani, Francesca, Camille Paloque-Bergès, Valérie Schafer, and Benjamin Thierry. 2019. “Qu’est-ce qu’une archive du Web?” https://books.openedition.org/oep/8713/.

The Rosetta Project. n.d. “Disk – Concept.” The Rosetta Project. https://web.archive.org/web/20240423002348/https://rosettaproject.org/disk/concept/.

———. n.d. “The Rosetta Blog.” The Rosetta Project. https://web.archive.org/web/20240423002731/https://rosettaproject.org/blog/.

———. n.d. “The Rosetta Project, A Long Now Foundation Library of Human Language.” The Rosetta Project. https://web.archive.org/web/20240423003014/https://rosettaproject.org/.

Schafer, Valérie. 2012. “Internet, Un Objet Patrimonial et Muséographique.” Colloque *Projet pour un musée informatique et de la société numérique*, Musée des arts et métiers, Paris. https://web.archive.org/web/20240423012521/http://minf.cnam.fr/Papiers-Verifies/7.3_internet_objet_patrimonial_Schafer.pdf.

Severo, Marta, and Olivier Thuillas. 2020. “Plates-formes collaboratives : la nouvelle ère de la participation culturelle ?” *Nectart* 11 (2). Toulouse: Éditions de l’Attribut: 120–31. https://web.archive.org/web/20240423003238/https://www.cairn.info/revue-nectart-2020-2-page-120.htm.

Turner, Fred. 2006. *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. Chicago: University of Chicago Press.

UNESCO. 2004. “Records of the General Conference, 32nd Session, Paris, 29 September to 17 October 2003, v. 1: Resolutions.” UNESCO. General Conference, 32nd, 2003 [36221]. https://web.archive.org/web/20240423004242/https://unesdoc.unesco.org/ark:/48223/pf0000133171.page=81.

A Decade of transformation discourse: Sociotechnical imaginaries of the Dutch web between 1994–2004

Nathalie Fridzema, Susan Aasman, Tom Slootweg, Rik Smit

Abstract: Web archives enjoy an increasing awareness and usefulness across a range of fields and disciplines, contributing to historical studies with archived web material, but also about the internet. Knowing how the web is imagined is essential for interpreting archived web material, especially during a highly transformative age in which technology advances at a rapid pace. This chapter investigates such a time of studying the Dutch web between 1994 and 2004, confirming that discourse surrounding the internet is appropriated differently. By means of a thematic analysis of a purposive sample of public media, the chapter presents 5 discursive themes that each reflect specific understandings of the Dutch web situated in a particular context, specified through social structure, scope, and time.

Keywords: public web, The Netherlands, sociotechnical imaginaries, internet history, thematic historical analysis.

Web archives enjoy an increasing awareness and usefulness across a range of fields and disciplines, contributing to historical studies *with* archived web material, but also *about* the internet (Brügger 2018, 140). Conducting research into these areas is not merely a discussion of material developments; how technology is understood and used is influenced by distinct discourses that shape social practices, norms, and values (Smit 2018, 47). Knowing how the web is imagined is essential for interpreting archived web material, especially during a highly transformative age in which technology advances at a rapid pace.

This chapter investigates such a time of technological, high transformability by studying the Dutch web between 1994 and 2004. This period marks the beginning of the public availability of the web until the rise of social media. It is furthermore characterized by the rise and fall of influential Dutch initiatives that adhered to ideals such as freedom, openness, and creativity. The various interpretations of these concepts are actualized in the usage of common metaphors that define the web as, for example, *global village*, *information highway*, or *commercial paradise*. Each of these phrases comes with specific meanings that are highly context-dependent; while the United States' cultural hegemony drove some interpretations of the web, others are best situated in Europe's or, more specifically, Amsterdam's particular creative culture in the 1990s. This case

Nathalie Fridzema, University of Groningen, The Netherlands, n.fridzema@rug.nl, 0009-0006-3044-8246
Susan Aasman, University of Groningen, The Netherlands, s.i.aasman@rug.nl, 0000-0003-1675-2998
Tom Slootweg, University of Groningen, The Netherlands, t.slootweg@rug.nl, 0000-0002-5181-8094
Rik Smit, University of Groningen, The Netherlands, p.h.smit@rug.nl, 0000-0002-9235-6869

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Nathalie Fridzema, Susan Aasman, Tom Slootweg, Rik Smit, *A Decade of transformation discourse: Sociotechnical imaginaries of the Dutch web between 1994–2004*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.15, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 141-161, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

study affirms that technology, as well as discourses about technology, are appropriated simultaneously in specific circumstances in which meaning is actively constructed. By looking into these complex dynamics, the chapter demonstrates that various interpretations of the meaning of the Dutch web were evident between 1994 and 2004. This specific historical period is critical to study because the decade before web 2.0 was a dense period including the disruption of traditional media, the emergence of various sociotechnical orders, and the normalization of a web culture that influenced our contemporary media landscapes.

This research contributes to studies of the early public web and can be described in three ways. I) Theoretically, the chapter offers the framework of sociotechnical imaginaries as a lens to study the development and understanding of technologies in a socio-historical context. II) In terms of methodology, the chapter demonstrates how an explorative analysis of discourse can be used to study a historical object systematically. And III), the research contributes empirically to the historiography of the public web in the Netherlands, offering opportunities to compare such lesser-studied narratives to dominant notions in internet history.

Thus, the research answers the question of what the prominent sociotechnical imaginaries of the web in the Netherlands between 1994 and 2004 are. Additionally, the chapter briefly elaborates on how the analysis of such a transformative period allows us to better position and interpret archived web materials and why it is important to scrutinize context-specific imaginaries.

1. Theoretical framework

Technologies are never merely material but are actively made sense of in discourse and through practice. Their uses and meanings are imagined by specific actors who are embedded in specific social, cultural, and historical contexts, and who have different degrees of discursive efficacy. The meaning of technology is, therefore, a site of discursive struggles and power dynamics. The web is no exception to this, as scholars (Flichy 2007; Stevenson 2013; van den Boomen 2014) have argued. Certain actors helped stabilize particular meanings of the web through their discourses. That is, certain ideas about the web and its uses have become more dominant than others, shaping how the web has been developed and used. Discourses, therefore, can help establish and normalize particular "sociotechnical imaginaries", which Jasanoff and Kim (2015, 4) conceptualize as "collectively held, institutionally stabilized, and publicly performed visions of desirable futures animated by shared understandings […] through, and supportive of, advances in science and technology."

Sociotechnical imaginaries are performative, in the sense that they are not merely symbolic, but actively shape a given technology's development as well as the policies and politics around it. By deconstructing sociotechnical imaginaries in texts and connecting them to their historical contexts, one can gain insight into the relationships between meaning (how the technology is interpreted or given meaning), materiality (how meanings are inscribed in the technology), and morality (how life ought to be lived, enabled or not by a technology) (Jasanoff and Kim 2015, 4). Moreover, the framework provided by Jasanoff and Kim suits this study's question and objectives particularly well because it recognizes that multiple sociotechnical imaginaries can exist simultaneously (23).

Following this perspective, sociotechnical imaginaries are mostly concerned with how possible futures of technologies are imagined. As Simone Natale and Gabriel Balbi (2014) point out, a fruitful distinction can be made between imaginations about future media technologies before they emerge and imaginations about a novel technology that has emerged. This distinction between future and new media is a dynamic one as future media can exist long before new media in various discourses such as science fiction without being grounded in reality. Interestingly, notions of future media like cyberspace that align with the actual development of the technology are best understood as prophetic afterward, and not as accurate predictions (Ernst and Schröter 2021, 37).

Following Natale and Balbi's conceptualization of a medium's life cycle (2014, 204), the Dutch public web between 1994 and 2004 is a communication network in its earlier period of introduction. Therefore, it is most suitable to look for 'new media imaginaries', specifically focusing on realms of the imagination pertaining to the novelness of the web. During this period, the technology's affordances are still flexible. To study this interplay between human imagination and the web's development, the authors propose to use the frame of interpretative flexibility from the SCOT tradition (social construction of technology). A "new technology often used in its early phase for different purposes by different social groups, and every group fights to impose a specific meaning on the novelty [...]. In a certain sense, the new technology is to be regarded not so much as a single technology as a continuum as possibilities" (Natale and Balbi 2014, 208). Following this framework allows us to study sociotechnical imaginaries, which are more grounded in reality—as opposed to future media—and specifically study common understandings of the web in the form of norms, values, and common sense, as well as critically look into the various social groups of influence.

The notion of temporality in internet imaginary work has also been conceptualized by Patrice Flichy in his trajectory of media imaginaries (2007, 10). Flichy understands technological reality as an ongoing process

between two poles of the *imaginaire*; utopia and ideology. Emergent utopian perspectives produce experimental initiatives that question an existing order. For these initiatives to become dominant technological assemblages, they must gain legitimacy and be mobilized within a new ideological framework that obscures certain facets of the prior (Lesage and Rinfret 2015, 3). Similar to Natale and Balbi (2014), a distinction is made between imaginaries about future media imaginaries, or utopias, and new media imaginaries, framed afore as ideologies.

Taking into account the notion of temporality helps to better select historical sources to identify sociotechnical imaginaries. As opposed to future media imaginaries—predominantly situated in texts like science fiction or written by futurologists—this study focuses on public discourses; texts (in various media forms) where reality is constructed through interpretation, performance, and social practice. Thus, the research adheres to the Foucauldian concept of discourse in which ideology and rhetoric are merged (Hodges 2015, 54). A specific focus is on metaphors concerning the web because these rhetorical devices were commonly used to help concretize the abstract and novel technology, and to help communicate values and purposes (e.g. *digital city*, *electronic highway*, or *global village*). The impact of metaphors is used in much scholarly work (Markham 2003; Van Dijk 2015) and is theorized by Lakoff and Johnson (2008). The latter authors argue that metaphors are central to defining our daily lives and have a structuring power that creates shared experiences.

2. Methodology

This research applies a Thematic Analysis (TA) to gain an understanding of which sociotechnical imaginaries about the web were present in the Netherlands, as well as how these various understandings developed over the period of interest. TA is suitable because it "allows the researcher to see and make sense of collective or shared meanings and experiences [...]. This method, then, is a way of identifying what is common to the way a topic is talked or written about and of making sense of those commonalities" (Braun and Clarke 2012, 57). Since sociotechnical imaginaries are, in the core, collective and idiosyncratic, TA is adequate to unravel hegemonic and counter-hegemonic notions, contributing to alternative readings of an idealized past and thus demystifying grand notions (Achugar 2017).

The research was structured as an exploratory sequential design. The first stage serves to acquire contextual knowledge through content analysis of various sources, as well as the identification of key points of interest and keywords. This is followed by a qualitative study aimed to explain the previously identified phenomena. In general, the design adhered to an abductive approach, allowing the implementation of existing theories as a

foundation while remaining open to modification based on emerging findings. This iterative interplay between data collection, analysis, and prior literature allows for a nuanced understanding of the object of study (Kennedy and Thornburg 2018). This approach is suitable for this type of historical research because it aims to situate dominant, historical narratives of the web in other, not poorly established contexts and thus build upon them. The table below explicates the procedures and analytical steps taken during the research. Besides offering a transparent step-by-step account of this research project, it also offers a methodological guide for similar research.

Table 1. Procedures and steps taken in the research process.

Procedures	Steps
Exploratory research (establish broader context)	o Exploration of the topic in public discourse and existing literature. o Content analysis of samples of newspapers and TV shows. o Identification of case studies and key phenomena.
Gathering and selecting data	o Gathering and organizing relevant discourse (TV shows, policy documents, popular books, texts from professional communities, radio shows, topical magazines, political party magazines). o Selecting a purposive sample based on previously identified important themes, context, metaphors, cases, and keywords that serve the research angles and objectives.
Describing and coding data	Textual analysis of sources. o Coding for subjects and themes; layout and structural organization; actors and institutions; ideological standpoints. o Coding for metaphors, keywords, and symbolic rhetoric. Contextual analysis of sources. o Reflecting on intertextuality in terms of conceptualizations and taking note of which texts are mentioned. o Noting down socio-political, and economic phenomena; keep a timeline of the research period. Identification and organizing of discursive statements (quotes).
Interpretation of data and mapping of sociotechnical imaginaries	Abductive categorization of sociotechnical imaginaries. o Categorizing sociotechnical imaginaries based on common understanding and mapping into broader discursive themes. o Supplementing sociotechnical imaginaries with quotes analysis. o Interpretation of dominant metaphors, thematic keywords, and particular rhetoric per sociotechnical imaginary.

The definitive sample used for TA comprises 46 Dutch media texts ranging across types of media and years of publication between 1994 and 2004 (Fridzema 2024). This period marks the beginning of the public availability of the web until the rise of social media and Web 2.0. The sample is purposively created; texts were selected throughout the research

process to be i) diverse in media types (from audiovisual news items to specialized magazines or policy papers), ii) spread across the period of interest, and iii) rich in potential to identify both dominant and alternative sociotechnical imaginaries. A critical note is that in these sources some of the more activist historical actors gained media preference, and increasingly assumed the role of spokespersons. This media preference may introduce certain biases in the types of imaginaries detected.

3. Findings

A total of five discursive, sociotechnical themes emerged from the analyzed discourses: civic; societal; economic; cultural; and ontological. Each theme includes at least three and up to six singular imaginaries that convey a specific vision of the Dutch web. Furthermore, each imaginary also includes predominant keywords, metaphors, and other symbolic rhetoric which convey the premise of the vision.

3.1 Civic

The civic sociotechnical theme includes five imaginaries that primarily convey a notion of the web being a democratic tool or place that could be beneficial for civil procedures in society or strengthen processes typically associated with the public sphere. Four imaginaries stem from discourses between 1994 and 1998, and emerged within the context of several local initiatives such as *Knoware*, *Internet Access Foundation* (IAF), *De Digitale Stad* (DDS) and *XS4ALL*[1]. The last two initiatives specifically sprouted from a highly educated, critical social class based in Amsterdam during the 1990s. The Dutch capital was a rich breeding ground for artists, squatters, and hackers to create an alternative culture (Olsthoorn 2015, 233; Apprich 2017). The Dutch hacker scene was part of an international network and quite active; there were multiple magazines in circulation and various hacker festivals were organized in the 1990s[2].

Despite these various contexts, it is important to emphasize that both initiatives operated locally during the mid-1990s due to practical boundaries such as hardware capabilities and pricing. This locality was the case for DDS in particular, which was, at its core, a virtual version of the city of Amsterdam and was thus mostly geared towards its inhabitants. In their early days, both DDS and XS4ALL aimed to provide services on a national level, but the following four imaginaries were part of a specific circle of

[1] Internet providers such as Knoware, Internet Access Foundation, and especially XS4ALL (Olsthoorn 2015) are examples of early initiatives that offered relatively cheap access to the web for everyone in the Netherlands outside the academic realm. De Digitale Stad (The Digital City) was the first online community in the Netherlands (Rommes, Van Oost, and Oudshourn 2003).

[2] Two hacker festivals were organized in the Netherlands during 1993: *Hacking at the end of the universe* (Seclist 1993) and *Next5minutes* (V2 n.d.).

enthusiasts that were predominantly bound to Amsterdam and part of its cultural scene. In other words, while both case studies are significant for Dutch web history, they represent a metropolitan view and do not take into consideration similar, peripheral activities and initiatives at the time[3].

Firstly, the web was imagined as connecting multiple, non-profit public spheres that would enable communication as well as provide public information. Various metaphorical constructions were used to underline this, such as *electronic highway, public domain,* and *global village* (all metaphors are translated from Dutch).

The second imaginary envisioned the web as a digital city which influences its meaning, usage, and design. Grounded in metaphors such as *digital city* and *digital infrastructure*, this imaginary reflected the essence of the DDS and seeped through on multiple levels such as interface design, jargon, social practices, usages, functions, etc. Interestingly, a significant amount of discourse reflected a critical awareness about the (over)usage of the 'city' metaphors:

> If that is your only metaphor, suddenly everything becomes like on-ramps and off-ramps and roadside restaurants, and in my eyes, that is not really necessary – Rop Gonggrijp, founder xs4all (Nieuwe economisch peil 1999).

Furthermore, the web was considered to be in its infancy, in an exploratory and experimental phase, which illustrates the novelty of the technology at the time. Metaphors like *testing ground* were often used in combination with other terms such as *innovative, testing, endless possibilities,* and *low threshold*. It is within this imaginary that utopian discourses develop in the form of future-oriented ideas about how the web will change reality significantly.

The fourth civic imaginary understands the web from a bottom-up perspective, highlighting its values as a decentralized, democratic space for ordinary users. This imaginary prevailed in hacktivist circles, particularly among those who founded XS4ALL. Within this framework, a counter-hegemonic narrative prevails; hackers should make the web available to a wider public (non-dominant groups in society), rather than serving a select group of privileged users (researchers and professionals). In other words, the web's accessibility was perceived as a universal right and the metaphors used reflected democratic ideals, conveying themes through discursive terms like *platform* and *network.*

The fifth and final civic imaginary emerged after 2000, coinciding with the gradual domestication of the web in society and its discussion in different contexts. As in the first imaginary, the web was approached as a

[3] The Digital City had other, popular versions in different cities, such as Groningen, which were developed locally (Digitale Stad Groningen 1998).

new public sphere, but here an emphasis was put on facilitating unbridled freedom of expression, even extreme and controversial opinions. This phenomenon was met with increasing annoyance by traditional media, as expressed in the quote below. Furthermore, it reflects the Netherlands' increasingly polarized political climate around the early 2000s[4]. Metaphors like *medium* prevail, emphasizing people's ability to express themselves in ways previously unattainable and connect with like-minded individuals online.

> Those group weblogs cause commotion, among other things, because you often see people hiding behind a pseudonym and the group itself. Just as in ordinary life, groups are often responsible for the most disturbance – Fransisco van Jole, journalist (NETWERK 2004).

3.2 Societal

After the web was more widely implemented in the Netherlands due to initiatives such as DSS and XS4ALL, but also other early providers such as the Internet Access Foundation (IAF) and Knoware (Olsthoorn 2015), other societal imaginaries emerged. These collective understandings situate the web as an impactful technology on a national level and thus touch upon themes such as social structures, daily life, and governance as a result of its domestication. In total, five societal imaginaries are identified from various public discourses across the decade of interest. Because these operate on a national scope—in contrast to the more concentrated civic imaginaries—the understandings are broad and, depending on the context, certain subtopics are more relevant than others. Therefore, it is worthwhile to provide a brief note on the Netherlands' socio-political development between 1994 and 2004. This period is characterized by trends related to neo-liberalism such as deregulation and privatization, an era of economic growth followed by a decline, the rise of populism, and the advancement of globalization (Mellink and Oudenampsen 2022).

The first societal imaginary positions the web in the context of the digital revolution and envisions the technology as essential for daily life. Recurring metaphors like *information society* and *digital infrastructure* reflect the sentiment of a pending large-scale societal transformation and its widespread use will enable such phenomena as *teleworking* and *teleshopping*. This sentiment began to emerge in the mid-1990s:

> For the first time, the contours of the information society become visible. [...] So far, computer networks are often associated with science or banking matters. But slowly, such technology is entering the living room – Fransisco van Jole, journalist (van Jole 1994, 185).

[4] At the end of the 1990s and beginning of the 2000s, the Netherlands saw an upsurge in populism and extreme right-ideologies. This notion became stronger after multiple (cyber)attacks and the 9/11 attacks.

The second imaginary reflects the notion that the web *will* drastically alter perceptions of reality either in a utopian or dystopian way. The first and second imaginary thus share a revolutionary rhetoric, but the latter one adopts solely future-oriented discourse. Utopian discourses mostly reflect watershed and phantasmagoric notions, a notable example of which is the idea of continuous synergy between the self and the technology.

> I believe in a future where biology and technology will merge, and the best of both worlds will prevail. I envision a future where we increasingly use technology to integrate with biology. [...] Ultimately, these two will just come together, giving rise to a new kind of humans who embody the best of both worlds – Paul Ostendorf, futurologist (Nieuwe economisch peil 2003).

Notable dystopian imaginaries are evident in public discourse as well. For example, the belief that a sense of community will get lost as no one would have to leave their house anymore whilst using the web, or the fear of addiction.

> The room in the future should never look like this. We need to continue doing normal errands, just keep loving each other. We should marry each other in city halls. Internet relationships scare me. For the first in days, tears well up in my eyes – Rogi Wieg, writer (Wieg 1996, 75).

Thirdly, the web is commonly understood as a technology that creates a social dichotomy between groups who already use computers and those who do not due to financial constraints and/or digital illiteracy. This narrative is apparent throughout the research period and includes a range of discourses pointing at the hegemonic position of particular social groups in the context of the web (*information-have-nots, wizzkids, the modemless, cybercitizens, digital newbies*). The fear of alienation due to the digital revolution or notions of a fragmented society were an ongoing concern since the early 1990s and thus were topics of discussion at the DDS and XS4ALL. Furthermore, within this imaginary, the founding of the *Digital Citizens Movement*[5] in 1994 and the subsequent government campaign *Internet for the Everyday*[6] are indicative of the societal apprehension at that time. Moving beyond the notion of estrangement, discourses post-2000s reflect a separation through fear and power; those who are comfortable with buying products online or knowledgeable about hacking computers, and those who are not.

---

[5] *Digitale Burgerbeweging* (digital citizen movement) was an organization founded in 1994 that wanted to put various rights of citiinternetke freedom of speech, in the Internet's development and legislation (Digitale Burger Beweging Nederland, 1996).

[6] *Internet voor alledag* (internet for every day) was a Dutch campaign with the then crown prince Willem Alexander as prominent actor to promote the implementation of the internet in people's daily routines. Iinternetom May 2001 until June 2003 (Internet voor alledag 2001).

The fourth societal imaginary involves the belief that its wider implementation will eventually reveal potential problems or (material) limits. Metaphorically, the web is commonly depicted as something concrete and tangible like *technical infrastructure*. Within this imaginary, many themes emerge throughout the decade that are dependent on context-specific influences. The most relevant are privacy and user rights; material limits of hardware/software; provocative online practices (unwanted intimacies, hacking, alternative subcultures, foul language, spam); (child) pornography; copyright; addiction; accessibility and literacy.

Finally, as a logical continuation of the previous imaginary, several discourses underline notions of control, governance, and legislation. A prevalent theme is the matter of responsibility: are users or platforms responsible for illegal music downloading; are parents responsible for children chatting online; are providers responsible for their client's behavior? Furthermore, the topics of censorship, encryption technology, and the privacy of citizens, were already prominent in the 1990s and re-emerged after 9/11 and the subsequent fear of terrorism. Pertaining to the latter, the question of how much the state should be able to monitor citizens' online behavior dominated discussion within this sociotechnical imaginary.

3.3 Economic

In the Netherlands, the period from 1994 to 2004 is mainly characterized by prosperity and increasing wealth (Mellink, Oudenampsen, and Woltring 2022). For the first time since the Second World War, economic growth numbers reached a record high and the unemployment rate was decreasing (Crone 2000). Simultaneously, the primacy of the welfare state was abandoned in favor of neoliberalism. In 1994, the Dutch government started *operation MDW* (market forces, deregulation, legislative quality) placing principles of a free market, privatization, and individualism at the core of government policy and practices (InfoMil 2023). The development of the web and digital society was a pivotal goal within this changing paradigm, influencing economic policies and vice versa. Central to this was the rise of the so-called 'new economy' (Gordon 2000). While ideological and public imaginaries of the web were evident in the 1990s, as discussed earlier, economic terms and notions were present from the beginning, yet gained prominence predominantly due to drastic governmental policy changes. As a result, the economic sociotechnical theme includes a rich set of discourses resulting in six imaginaries spanning the period of interest. The growing commercial hold on the web was widely discussed, most notably in talk shows, news broadcasts, current affairs programs, and documentaries.

The first imaginary highlighted the web's evolution as an important opportunity to gain a competitive edge in Europe and beyond; its national development is viewed concurrently as a measurement of success. Metaphors such as *information superhighway* reflect its impact on a large scale while symbolic rhetoric like *head start, falling behind,* and *competitive position* indicates a belief in partaking in a race. The Netherlands aspired to be among the first digital nations.

Secondly, within the economic context, the web symbolized another economic reality with new commercial opportunities for the private sector. A pivotal paradigm shift can be identified within this imaginary theme: the internet and the web are equated with 'the new economy', i.e. a new way of doing business digitally. Especially after 1995, a so-called 'new open market' emerged, prompting the search for new lucrative business models. This particular vision is prevalent in public discourses stemming from economic actors; this is how the market will work going forward. Symbolic rhetoric surrounding the web is marked by financial and hopeful language (i.e. *innovative, investing, infinite growth potential, success formula*) with references portraying the web as a *commercial paradise*.

Furthermore, many of these statements foreshadowed the rise of a new hegemonic force in the Netherlands: the new economic sector including internet businesses and related commercial actors. A stark difference is noticeable between these initiatives and those discussed in the civic imaginary theme. While the civic initiatives focused on the web's democratic capabilities, the new economy start-ups directed their attractions towards its commercial potential. Shared ideals like openness prevailed, but were reinterpreted within the context of efficiency and profitability.

The third economic imaginary involves the connection between the web and a distinctive entrepreneurial culture that evolved from 1995 onwards, giving rise to prominent figures like gurus and experts. Symbolic rhetoric like *internet gurus, experts, pioneers* is used, alongside more satirical stereotypes such as *boundless optimists, digital evangelists,* or *NASDAQ yuppies.* The discursive effects of forming this social group are noteworthy; the entrepreneurial spirit manifests itself in the forms of events like the famous First Tuesday drinks, and well-known gurus are consistently invited to share their perspectives on TV shows[7].

The fourth imaginary emerged from a starkly different context than previously discussed, namely after the internet market crashed. Cynicism and uncertainty had always surrounded the new economy, but after 2001 public discourses began viewing the new digital market as an illusion. The metaphorical portrayal of *the internet world* often created a divide between

[7] First Tuesday drinks were monthly network gatherings for entrepreneurs in the internet industry.

common-sense reality and 'that' group of people and businesses who blindly followed the *hype*. While the dotcom crash is often regarded as an American phenomenon, the Netherlands had its fair share of internet start-ups like Zon, Newconomy, and World Online—the latter of which went public, but ultimately crashed alongside the NASDAQ. As a result, many individuals lost their investments, a popular topic of discussion in Dutch media, further fueling the critical attitude of this imaginary towards the web's new economy.

Due in part to the failure of World Online, a fifth imaginary quickly emerged. The *web industry*—another metaphor to distinguish this societal group—was the subject of suspicion, specifically with regards to the workers' skill sets, the ethics of *doing internet business* on the stock market, and the industry's attitudes towards employees and consumers. This understanding extended from the third imaginary and further sustained this hegemonic group in society.

> I believe that [the internet world] consists nowadays of, unfortunately, 90% of people who are there solely to make quick money – Michael Frackers, founder of Planet internet (PER SALDO 2001).

Lastly, the sixth imaginary positions the web's economic structure as an *illusion* that was deliberately circulated for propagandistic purposes; the new economy was purposely created by a select group of people to generate profits. Following this understanding, start-ups like World Online created extensive marketing campaigns to spread disinformation, fostering false hope. This imaginary is arguably a logical continuation of discontent felt in society after the crash, compounded by the Netherlands being in a period of economic recession in 2002. Metaphors like *Walhalla* were used alongside such terms as *small talk, gambling,* and *throwing mud.*

3.4 Cultural

Through the gradual domestication of the web, various web cultures emerged between 1994 and 2004, encompassing many practices, jargon, and norms from an ever-growing group of users at the time. Similar to the societal imaginaries, the cultural theme is broad in terms of topics and scope. Depending on the specific discourses and social group in focus, cultural imaginaries may differ. Nevertheless, three distinctly recognizable categories are identified.

The first imaginary encompasses the very notion of 'web culture' and its perceived influence on daily life, habits, and practices, set against the backdrop of the material infrastructure at the time. Common metaphors included *cyberspace* and *virtual community*. Exemplary cultural jargon that emerged during the decade included *chatting, web surfing, scrolling,*

cyberspace traveler, in real time, and various electronic texts like *e-zines* or *e-mails*. As some aspects of early 1990s web culture normalized in the 2000s, other unique features also solidified. Examining a case like DDS reveals a rich history of Amsterdam's cultural scene going online for the first time. DDS was modeled as a digital city and thus virtual meeting spaces could resemble cafés, the metro, or post offices. Here, users could engage in practices like *irc'ing* or *mudding*[8] in so-called *metro meetings*. Furthermore, *newsgroups* emerged during this time, known to foster *flame wars*[9]. Topics such as *cyber romance, flirting*, and *sex* sparked significant interest in various public discourses, including popular books or TV/radio shows—a curiosity that endured well into the 2000s when chat services like MSN gained popularity.

Secondly, between 1994 and 1996, the web was envisioned as an additional layer of reality, called cyberspace, inhabited by various subcultures. Although the web was not yet fully enmeshed in everyday life, it was already perceived as a new, distinct space: an *artificial city, virtual city, electronic island,* or a *digital world*. Most statements from the mid-1990s initially assessed the web positively as a space for subcultures. Nevertheless, the notion of virtual communities and the spaces they inhabited gradually became associated with the emergence of extreme and contrarian opinions and attitudes (as discussed in the fifth civic imaginary). Generally, this imaginary includes a mix of legitimizing and mobilizing discursive statements, future-oriented utopian frameworks, and hegemony orientations related to community formation online.

Similar to the previous one, the third and final cultural imaginary emerged in discourses before 2000. Here, the notion of disembodiment is foregrounded and connected to creating different identities and/or *avatars*, separate from reality. The symbolic rhetoric deployed was predominantly spiritual or religious in tone, giving an esoteric dimension to the conception of the self in relation to technology. For example, some historical actors described the practice of becoming a *ghost* or *soul* online, falling in love through the mind, and leaving one's *shell* behind.

> The electronic possibilities have allowed me to create a really good avatar and make it do a variety of things. I do not see it as pretending to be someone else; I see it more as separating a part of yourself that you can then thoroughly explore – Yvonne la Grand, artist (De andere wereld van de zondagmiddag 2004).

3.5 Ontological

[8] IRC is the abbreviation for Internet Relay Chat. MUD is the abbreviation for Multi User Dungeon, referring to a type of online roleplay game (van de Boomen 1996).

[9] Textual disagreements with swear words and inappropriate language usage via e-mail or in newsgroups (van de Boomen 1996).

The final sociotechnical imaginary theme is ontological and encompasses all understandings that convey something about how the web's intrinsic existence, properties, and social reality are viewed. While most imaginaries have an ontological dimension, the understandings in this theme articulate how the technology of the web is perceived *a priori* and how this, in turn, shapes society's understanding of itself. Two ontological imaginaries are identified.

Firstly, the web's character is principally understood to be liberal and aligned with ideals of decentralization, freedom, openness, and individualism. Exemplary rhetoric includes *many-to-many attitude, unique features,* and *open nature.* The statements reflect a strong, ingrained understanding of the web characterized by notions of the technology being inherently associated with a set of conditions and rules, whether one agrees with them or not. Notably, some statements advocate the belief that nobody can own the web and that it does not possess a commercial character. As discussed previously, the rise of the new economy caused some bitterness.

The second ontological imaginary posits that the web is understood as a network influencing governing structures, technological infrastructures, and how individual actors (should) understand themselves and their relationships. This imaginary represents an interesting dichotomy, namely, the development of an increasingly networked society alongside the rise of individualism. The latter aspect can be contextualized within the neo-liberal surge in Dutch society that also informed the entrepreneurial spirit of the new economy in the late 1990s. A common metaphor used to convey this reality besides *network* or *infrastructure,* is *system*. Common rhetoric either emphasizes the interconnectedness of society through the web (e.g. *connections, circulate, organizational forms, links,* and *chain reactions*) or focuses on the individual (*tailormade, identity, your own culture*). Despite being present in certain discourse in the 1990s, the concept of the *network society* grew particularly popular in the 2000s and is often situated in discussions about the commercial aspects of the web. Furthermore, a prominent theme is the fear or danger of being too dependent on technological networks. Especially after 9/11, technological networks are imagined as vulnerable since one failure or hacker can trigger a chain reaction of problems.

> We have seen chain reactions where the power goes out for a whole day simply because a number of power plants fail. I think it is warranted to have a critical look at how things will work out if companies become increasingly dependent on each other through the internet – Roel Pieper, entrepreneur (Zembla 2004).

4. Discussion and conclusion

The study has identified a total of 21 sociotechnical imaginaries

categorized amongst 5 discursive themes. Each imaginary reflects a specific understanding of the Dutch web situated in a particular context, specified through social structure, scope, and time. The chapter demonstrates how the understandings were conveyed through the usage of various common metaphors and/or rhetoric. Despite the variety of interpretations and meanings, a few trends can be identified across the sample. Following Flichy's framework, most discourse is ideological as opposed to utopian, typically part of future media imaginaries. Once the web became publicly available, more widely used, and domesticated, its technological usage became normative and part of the status quo. In general, the research reveals an evolution in sociotechnical imaginaries transitioning from early grassroots democratic ideals to hopeful entrepreneurial visions, culminating in more critical perspectives post-2000. The latter turn can be positioned in various historical phenomena. Internationally, the influence of the 9/11 terror attacks or the frenzy due to the millennium bug problem were significant for creating notions of risks when it comes to the internet and web. Nationally, the Netherlands sees the rise of extreme populism and anti-globalization ideals, as well as the increasing implementation of the web as an infrastructure in all aspects of daily life creating a sense of dependence. Economically, the dot.com crash did a lot of detriment to the commercial optimistic mindset. Additionally, the act of buying goods online took some years to become normalized in the Netherlands, further influencing economic beliefs. A second trend is that the hegemonic notion is reflected in a significant amount of the sources and aptly illuminates the creation of dominant social groups. Most notable is the emergence of an entrepreneurial group that was at the forefront of many commercial initiatives after the mid-1990s and proved influential in creating the economic imaginaries of the web, replacing the more public (civic) imaginaries.

Likewise, hegemonic discourse brought forth a social dichotomy that developed between those who are able or willing to adapt to the new technology and those who are not. The afore-mentioned hegemonic power relations remained influential during the rise of web 2.0 and thus this research historicizes from where they originated. Finally, the research observed that imaginaries generally grew more critical, bitter, and/or hesitant after 2000 as discourse shifted towards the risks, vulnerabilities, and adverse effects of the web. This observation helps us to better understand the particular time and Dutch context of the turn of the century from a historical perspective. Futurism and optimism were significant themes towards the end of the 1990s but, as this study affirms, it is apparent that what the web *is* and how it was understood varies significantly before and after 2000.

Utilizing the framework of sociotechnical imaginaries to study the history of the web has proven vital for contextualizing meanings related to

Dutch web culture and practices in various ways. First, the research demonstrates the importance of looking beyond dominant notions and showcases that many imaginaries exist simultaneously. Also, these understandings are predicated on such dimensions as context, social group, or locality. Scrutinizing such scopes helps to better position ideologies while also enforcing the idea that preeminent notions do not have to be all-encompassing in terms of national beliefs. Furthermore, imaginaries do not exist independently but rather influence or inform each other. This is evident in the observation that metaphors or specific rhetoric are used across discourses but vary in meaning. For example, the ideal of 'openness' is used in the civic theme to frame the web along the lines of democratic ideals while the economic imaginary positions openness in terms of free markets and commercial benefit.

Ontologically, openness is perceived as an intrinsic characteristic of the web and refers to its technological structure. The difference in appropriation is also evident in the usage of metaphors. The term *information highway* has been used across the decade but positions the web in various ways; either as a communication opportunity between citizens, a network to connect countries on a global level, or a framework to enhance trade employing e-shopping or e-working. Thus, being aware of the various sociotechnical imaginaries, especially concerning the specific language used, is crucial for detecting alternative ideologies and contextualizing sources in historical research such as web-archived material. To elaborate on the latter, knowing which dominant or alternative imaginaries are of influence considering certain websites can help to better understand their content and meaning. Solely looking at the archived material without considering collective understandings of what actors thought the web's main function was, will disregard an important layer during the public web's early fluctuating years.

The case study of the public web in the Netherlands also demonstrates that popular techno-solutionist notions, mostly based on Big Tech institutional narratives of Silicon Valley (Morozov 2013, 13), should not be interpreted as global or solely significant. While American narratives have dominated internet history, the public is generally described as passive in such studies. Even in books that take a critical, non-technological deterministic perspective, like *From Counterculture to Cyberculture* by Turner (2006) or Morozov's *To Save Everything Click Here* (2013), one can still detect the inclination for top-down imaginaries from hegemonic groups. The chapter has shown that bottom-up imaginaries are also prevalent in society and can have significant discursive effects related to the web's development. The understanding of the web as causing a social dichotomy both in terms of skills, but also in culture, is a good example of this. Solely focusing on dominant understandings, or elite imaginaries, will result in

overlooking other important discourses and interpretations that are critical for understanding the web's everyday culture.

This does not imply disregarding top-down understandings or American narratives altogether; even in the specific context of the Netherlands, they prove to be influential and should be taken into account from a historical perspective. To elaborate, it is evident from the studied sources that America's techno-cultural hegemony was present, as concepts, actors, or ideas often appear in Dutch academic and popular discourse. For example, the usage of *information highway* is often in reference to Clinton and Gore's appropriation of the metaphor. Besides, influential popular-science texts like Rheingold's *Virtual Community* (1993) and Castells' *The Internet Galaxy* (2001) are referenced multiple times.

While all these examples clearly influence the web's development in the Netherlands, the Dutch initiatives are not mere copies of American counterparts. For example, the Dutch case of DDS is comparable to the famous pre-web forerunner The Well[10], however, it does not fit the description of a counterculture in the same fashion as Turner argued (2006). DDS, as well as XS4ALL, aimed to replicate the actual public communities in Dutch society virtually by working together with the government. In contrast, American initiatives such as The Well created anarchist communities to construct a sense of camaraderie that was missing in American post-war society and thus truly represented a counterculture (Turner 2006). This dichotomy is also identified by Apprich (2017), who positions the emergence of a specific European network critique in a larger commercial cyberculture. The complex intertwinement of both objectives is, according to Apprich, aptly represented in the dominant usage of the city metaphor. "This was less a matter of simulating the city, but rather a metaphorical transcription of the dynamics and diversity of urban processes" (2017, 80). DDS facilitated both civic or public functions in its center while also facilitating not-registered, niche cultures in its periphery. This demonstrates that while American narratives are influential and should not be ignored, dominant imaginaries should be contextualized and critically compared to specific contexts and scopes.

Future studies in Web History should aim to push beyond conventional histories such as the web as a utopian ideology based on Silicon Valley's idealized past. While the identified sociotechnical imaginaries provide a sufficient look into collective imaginations of the web, they are by no means generalizable for everyone in society as they are based on popular, public discourse. Future research will thus explore how the public web was understood and experienced by everyday users who are not part of a

[10] The Well is the abbreviation for The Whole Earth 'Lectronic Link which was one of the first online communities on the web, founded in 1985 (Turner 2005).

significant hegemonic group in society. To truly capture such lived experiences, methods like oral history interviews or the analysis of bottom-up sources like archived web materials need to be adapted. This is in line with calls from scholars like Janet Abbate who argue that internet histories should not solely be defined along material definitions. Other perspectives like usage experience or locality are important as well, whilst it is crucial to point out how such definitions bring their own socio-politics to the historiography of the internet (Abbate 2017). This means that we need to expand on our number and types of sources. Because our analysis of public media like popular TV shows and magazines were used besides sources from key actors or initiatives, the sample is predominantly based on established media with a limited number of historical actors. Further research aims to explore other bottom-up narratives using, for example, oral history interviews, but also discursive practices of early web users.

References

Abbate, Janet. 2017. “What and where is the internet? (Re) defining internet histories.” *Internet Histories* 1 (1-2): 8–14.

Achugar, Mariana. 2017. “Critical discourse analysis and history.” In *The routledge handbook of critical discourse* studies, edited by J. Flowerdew and J.E. Richardson, 298–311. Oxfordshire: Routledge.

“Afl. 3: Route 99: Werelden van internet – de opkomst van internet” Nieuwe economisch peil. (NOT, 1999). Television show.

“Afl. 22: internetgoeroe's slaan terug.” Nieuw economisch peil. (NOT, 2003). Television show.

Apprich, Clemens. 2017. *Technotopia: A media genealogy of net cultures*. Lanham: Rowman & Littlefield.

Braun, Virginia, and Victoria Clarke. 2012. “Thematic analysis.” In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological*, edited by V. Braun, et al., 57–71. Washington: American Psychological Association.

Brügger, Niels. 2018. *The archived web: doing history in the digital age*. Cambridge: MIT Press.

Castells, Manuel. 2001. *The internet galaxy: Reflections on the internet, business, and society*. Oxford: Oxford University Press.

“Computerterreur.” Zembla. (NPS, 2004). Television show.

Crone, Ferd. 2000. “Nieuwe economie, nieuw wassenaar? Naar een nieuw maatschappelijk vergelijk.” *Socialisme en Democratie* 57 (7-8): 333–344.

“De andere wereld van de zondagmiddag”. (IKON, 1998). Radio show.

Digitale Burgerbeweging Nederland. 1996. *Official Website.* December 27. Accessed 19 December 2023. https://web.archive.org/web/19961227042913/https://www.db.nl/.

Digitale Stad Groningen. 1998. *Official Website*. February 6. Accessed 19 December 2023. https://web.archive.org/web/19980206073946/http://dsg.nl/.

Ernst, Christoph and Jens Schröter. 2021. *Media futures: Theory and aesthetics*. London: Springer Nature.

Flichy, Patrice. 2007. *The internet imaginaire.* Cambridge: MIT Press.

Fridzema, Nathalie. 2024. *Purposive sample of Dutch media*. DataverseNL. https://doi.org/10.34894/PEGU7L

Gordon, Robert J. 2000. “Does the ‘new economy’ measure up to the great inventions of the past?” *Journal of Economic Perspectives* 14 (4): 49–74.

Hodges, Adam. 2015. “Intertextuality in discourse.” *The handbook of discourse analysis*, edited by H. E. Hamilton, D. Tannen and D. Schiffrin, 42–60. New Jersey: John Wiley & Sons.

Internet voor alledag. 2001. *Official Website.* June 5. Accessed 19 December 2023. https://web.archive.org/web/20010605022656/http://internetvooralledag.nl/.

InfoMil. 2023. “Marktwerking, deregulering en wetgevingskwaliteit (MDW) – project.” March 23. Accessed 19 December 2023. https://web.archive.org/web/20230323183830/https://www.infomil.nl/onderwerpen/integrale/activiteitenbesluit/regelgeving/overzicht/inwerkingtreding/virtuele-map/mdw-project/.

Jasanoff, Sheila and Sang-Hyun Kim. 2015. “Future imperfect: Science, technology, and the imaginations of modernity.” In *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power,* edited by S. Jasanoff and S. H. Kim, 1–33. Chicago: University of Chicago Press.

Kennedy, Brianna L., and Robert Thornberg. 2018. “Deduction, induction, and abduction.” In *The SAGE handbook of qualitative data collection*, edited by U. Flick, 49–64. Sage.

“Kijken, kijken… niet kopen”. PER SALDO. (RVU, 2001). Television show.

Lakoff, George and Mark Johnson. 2008. *Metaphors we live by*. Chicago: University of Chicago Press.

Lesage, Frédérik., and Louis Rinfret. 2015. “Shifting media imaginaries of the web.” *First Monday* 20 (10).

Markham, Annette N. “Metaphors reflecting and shaping the reality of the internet: Tool, place, way of being.” Unpublished manuscript, presented at the 4th annual conference of the International Association of internet Researchers (AoIR), Toronto, Canada, October 2003.

Mellink, Bram and Merijn Oudenampsen. 2022. *Neoliberalisme: een Nederlandse geschiedenis*. Amsterdam: Boom.

Morozov, Evgeny. 2013. *To save everything, click here: The folly of technological solutionism*. New York: PublicAffairs.

Natale, Simone, and Gabriele Balbi. 2014. “Media and the imaginary in history: The role of the fantastic in different stages of media change.” *Media History* 20 (2): 203–218.

“NETWERK”. (KRO, 2004). Television show.

Olsthoorn, Peter. 2015. *25 jaar internet in Nederland*. Amsterdam: FMT.

Rheingold, Howard. 1993. *The virtual community: Finding connection in a computerized world*. London: Secker and Warburg.

Rommes, Els, Van Oost, Ellen and Nelly Oudshourn. 2003. “Gender in het ontwerp van De DigitaleStad Amsterdam.” *Amsterdam sociologisch tijdschrift* 30 (1–2): 182–204.

Seclists. 1993. “Hacking at the end of the universe, summer congress.” May 21. Accessed 19 December 2023. https://seclists.org/interesting-people/1993/May/23.

Smit, Rik. “Platforms of memory: Social media and digital memory work” (PhD dissertation, University of Groningen, 2018).

Stevenson, Michael 2013. “The web as exception: The rise of new media publishing cultures” (PhD dissertation, University of Amsterdam, 2018).

Turner, Fred. 2005. “Where the counterculture met the new economy: The WELL and the origins of virtual community.” *Technology and culture* 46 (3): 485–512.

Turner, Fred. 2006. *From counterculture to cyberculture: Stewart Brand, the Whole Earth Network, and the rise of digital utopianism*. Chicago: University of Chicago Press.

V2 n.d. “Next 5 minutes.” Accessed 19 December 2023. https://web.archive.org/web/20220728124840/https://v2.nl/events/next-5-minutes/.

Van den Boomen, Marianne V. 1996. *Internet ABC voor vrouwen*. Amsterdam: Instituut voor Publiek en Politiek.

Van den Boomen, Marianne V. 2014. *Transcoding the digital: How metaphors matter in new media*. Amsterdam: Institute of Network Cultures.

Van Dijk, Teun A. 2015. "Critical discourse analysis." *The handbook of discourse analysis,* edited by H. E. Hamilton, D. Tannen, and D. Schiffrin, 466–485. New Jersey: John Wiley & Sons.

Van Jole, F. 1994. *De internet sensatie. Een reisverslag uit cyberspace.* Amsterdam: Rainbow Pocketboeken.

Wieg, R. 1996. "15 Mei." In *Alleen met het internet*, edited by W. Brands, 74–80. Amsterdam: Balans

Flirting and the web: The case study of Luxusbuerg

Carmen Noguera

Abstract: This research examines the role of the user in shaping and defining participatory platforms during the early years of the internet and the web. The focus of the study is Luxusbuerg, a Luxembourgish chat network created in 1996 as a channel on the Undernet International Relay Chat (IRC). The channel's success led its owners to develop a dedicated Luxembourg server in 1999, a web chat platform with channels tailored to users' interests. The chapter analyzes how intrinsic IRC elements shaped Luxusbuerg's evolution and user behavior, with a particular focus on the growing prevalence of flirting interactions. Finally, it includes an analysis of the #flirt and #queer channels, in order to ascertain their structural dynamics and user interactions over time.

Keywords: IRC, online dating, early participatory websites, online flirting, web chat

Luxusbuerg was a chat platform specific to Luxembourg created in 1996 as a channel on the Undernet International Relay Chat (IRC)[1], one of the largest real-time chat networks worldwide. The channel's popularity in Luxembourg drove its owners to set up their own server in 1999 with the support of Post Luxembourg.

Based on three oral interviews, an analysis of web archives, and a press corpus, this chapter first describes the platform's origins and how the fact that it was born under the wings of Undernet IRC marked its own existence. We especially highlight five main characteristics inherent to IRC, that were present during Luxusbuerg's lifetime, from its creation—following an international trend emerging at the time when channels based on geographical locations were being created and local linguistic communities were emerging—to its evolution. We also highlight the relevant role of users in such communities, in which they co-shaped and co-created the platform, following the trends marked by IRC's social structures. As the user-centered model reshaped the platform's user experience, flirting became the most popular use of the chat. We demonstrate how it drove the platform organization when it became its own channel-based server, constantly evolving and being shaped by its users. #Flirt became the most stable channel, with 400–500 people connecting simultaneously daily. The success of #flirt was also influenced by one of the distinctive features of the platform, the combination of online and offline interactions. This 'hybrid model' led to a community marked by the innervation of traditional and

[1] More about IRC may be found in Senft (2003). About IRC and Undernet see (Latzko-Toth 1998, 2010); (Stenberg 2002); and (Hinner 2000).

Carmen Noguera, University of Luxembourg, Luxembourg, carmen.noguera@uni.lu, 0000-0001-5275-3694

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Carmen Noguera, *Flirting and the web: The case study of Luxusbuerg*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.16, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 163-183, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

virtual communities feeding back to each other, what Latzko-Toth (1998) has defined as 'tribus IRC'.

Finally, by studying the first participatory websites such as Luxusbuerg, using web archives as a source, we aim to contextualize IRC and chat platforms as precursors of modern social media and one of the ancestors of instant messaging, following Ortner, Sinner and Jadin (2018), shedding light to the first experiences on online flirting and dating.

1. Origins of Luxusbuerg and influence of International Relay Chat

Luxusbuerg started in 1996 as an Undernet IRC channel[2]. It emerged as an alternative to a 'Luxembourg' channel that had already existed since at least 1994[3] on the same network, as explained by Raoul Mulheims, one of the founders of Luxusbuerg, in an oral interview in 2022. The Luxembourg Channel had a growing community of 20 to 40 people from 17 to 30 years old, all with a shared interest in technology. The founders of Luxusbuerg spent some time interacting with the 'Luxembourg' channel until they decided to create their own. The name 'Luxusbuerg' was a tribute to the national cultural heritage. It came from the Luxembourgish comic series Superjhemp: the hero was a parody of a stereotypical Luxembourger who lived in a city named Luxusbuerg. This saga "represents an almost inexhaustible repertoire of possible samples regarding Luxembourg identities" (IPSE 2011, 181). Soon, some rivalry between the two channels started, and since Luxusbuerg was very active, it surpassed the original one, with 200–300 people connecting simultaneously. Raoul Mulheims, together with two friends, Mehran Khalili and Christophe Leesch, decided to take it to the next level and to create a dedicated Luxembourg chat system that became operational on October 3, 1999 (Luxusbuerg 2002) and received support from Post Luxembourg ("Luxusbuerg" 1999; Tom Kettels, interview by author, September 28, 2022). The platform was launched and organized in five channels: #cafe, #teens, #20plus, #flirt, and #computers, according to the only registered version from 1999 in the Wayback Machine[4]. All of them disappeared after some time, while other channels

[2] IRC is a text-based chat system for instant messaging created in 1988. IRC evolved from a program with limited features at the beginning to an extensive complex technical infrastructure with four major independent networks. Undernet was one of them, created in 1992 with a user-centric approach, focusing on the users' needs and involving them in the network's governance.

[3] We do not have the precise creation date of the channel, and this is an estimation based on an oral interview with Raoul Mulheimms, one of the creators of Luxusbuerg. Based on this interview, we assumed that the Luxembourg channel existed in 1994, but we cannot confirm when it was exactly created and if it existed before that year.

[4] The first version of the Homepage of the new server Luxusbuerg was archived by the Wayback Machine on November 29, 1999. https://web.archive.org/web/19991129040331/http://www.luxusbuerg.lu/index.php3 Last accessed December 20, 2023.

emerged, except for the #flirt channel, which remained one of the main pillars of the chat network until it ceased its activities.

In the second part of the paper, we therefore analyze the impact of the inherent characteristics of Luxusbuerg in the development of flirting and do a close reading and analysis of the #flirt channel and its evolution through time, paying attention to its structure and user interactions. We also highlight the increasing consideration paid to security and legal issues to improve the users' experience, and the creation and implications of the #queer channel. Even though Luxusbuerg became its own platform in 1999, IRC characteristics played a significant role in its evolution, be it from a technical, organizational, social, or economic perspective. We can identify five main characteristics inherent to IRC that have defined Luxusbuerg since its inception: the rise of channels based on geographical locations, the emergence of local linguistic communities, the continuity of IRC's social structures, the users' role in co-shaping and co-creating the platform, and the synergies between traditional and virtual communities[5] consolidating the success of flirting as one of the main uses of the chat.

1.1 Rise of channels based on geographical locations

According to Latzko-Toth (1998), geographical location played a significant role in forming IRC channels. "Of the 196 most frequented channels in one of our surveys, more than a quarter were named after a locality: country (#chile, #france*...),* state or province (#arizona, #quebec...), or city (#auckland, #istanbul, #manila...)."[6]

To understand why people connected to a local community in an international chat, three main reasons can be proposed based on scholarly literature: the possibility of jumping quickly to the offline world, having quicker replies based on temporal closeness, and having a more straightforward conversation because the users share the same context (Latzko-Toth 2008, 1998; Billedo 2009; Dakhlia and Poels 2012; Velkovska 2002). These explanations align with the ones offered by the creators of the chat. For Raoul Mulheims, the possibility of going quickly to the offline world due to the small size of Luxembourg was one of the main motivations to connect to a local chat: "You meet online, and then you can say, let's meet this evening, and then we just get a drink." (Mulheims, interview by author, March 22, 2022). Similarly, Christophe Leesch, co-

[5] More about virtual communities on IRC in Liu (1999) and Jones (1997).

[6] Free translation of the author supported by DeepL. Original text: « Sur les 196 canaux les plus fréquentés lors d'un de nos sondages, plus du quart avaient pour nom une localité: pays (#chile, #france...), état ou province (#arizona, #quebec...), ou ville (#auckland, #istanbul, #manila...). » (Latzko-Toth 1998, 56)

founder of Luxusbuerg, explained in an interview in 1999 (“Luxusbuerg IRC” 1999, 290):

> A chat is a group of users who have something in common. Our channel is reserved for users who live in Luxembourg but do not necessarily have Luxembourgish nationality. The acquaintances that we make on the more international chats usually end in an email correspondence and keep their virtual character. The small size of our country also allows us to meet our chat friends in reality.

In addition, there is a close link between the creation of chats based on geographical locations and the flirt use—a general tendency to be found not only on IRC but also in France with the Minitel, as explained by Josiane Jouët, which may shed light on why the #flirt channel was the most popular in Luxusbuerg; users were able to meet offline and were also encouraged by the hybrid model based on the innervation of virtual and traditional communities.

> As the bodies are absent, the corporality is reintroduced by the words, and besides from the first contacts, it was the questions "H or F?" ("man or woman?") or "ASV" ("age, sex, city") Why? Because if I am in Dunkerque and you are in Marseille, it will be more difficult to contact you. That being said, there were people capable of crossing the whole of France to meet Minitel correspondents, but for future contacts, we were primarily looking for people from our region. (Dakhlia and Poels 2012, 227)[7].

1.2 Emergence of linguistic communities

Since its origins, Luxembourgish was the main language used to chat, even on an English-dominant platform such as IRC, while the content, the general news of the channel, and the navigation menu were both in English and Luxembourgish:

> There was a long discussion as far as I can remember. We said only Luxembourgish is not great because already, at the time, there was a strong non-Luxembourg-speaking community, so we needed another language. And we said, ok, the Internet language is English, let's create it in English. But the chat language in the public channels was typically Luxembourgish. (Mulheims, interview by author, March 22, 2022).

Following the international evolution of the IRC channels that gave prominence to local languages, Luxembourgish gained increasing importance over the years until it became an identity symbol celebrated by the local press. Mousel and Lulling (2002, 16–17) noted in D’Lëtzebuerger:

---

[7] Free translation of the author based on DeepL: « Comme les corps sont absents, la corporalité est réintroduite par les mots et d’ailleurs dès les premiers contacts, c’était les questions « H ou F ?» (« homme ou femme ? ») ou « ASV» (« âge, sexe, ville) Pourquoi? Parce que si je suis à Dunkerque et que tu es à Marseille, ce sera plus difficile de te rencontrer. Cela étant, il y avait des gens capables de traverser toute la France pour rencontrer des correspondants du minitel mais pour de futures rencontres, on cherchait prioritairement des gens de sa région ».

> It is quite clear that the internet as a new communication medium has already contributed greatly to the recent rise in the Luxembourgish written language, whether in the web or electronic mail. Never before has so much been written in Luxembourgish as in the last two to three years. A prime example is the website Luxusbuerg http://www. luxusbuerg lu, whose founders wanted to create a 'platform for the Luxembourg online society with their 'chat portal.'With an average of 3,400 users daily, the reception can already be described as a success. You can easily 'chat' with different people about the most diverse topics in Luxembourgish[8].

This shift from English to local languages is one of the characteristics inherent to the evolution of IRC chats. According to Latzko-Toth (1998, 46):

> [...] the English language, formerly used as a lingua franca on the main channels, has gradually given way to a mosaic of linguistic communities, especially on the Undernet where Quebecers tend to systematically create French versions of channels belonging to the English 'common core,' while on the same network, for some time now, the Malay language has been asserting itself as the one that has been growing most rapidly.

The rise of local languages is also perceived when doing a close reading of the #flirt channel language policy and its evolution, as traced by the Wayback Machine (Table 1). This phenomenon is particularly pertinent in Luxembourg, a multilingual country characterized by the prevalence of three main languages—French, German, and Luxembourgish—and English, taking on a predominant role in the early years of the web. Moreover, the use of English gained increased importance in the country due to the growing number of foreign residents[9]. We observed a discernible shift in language use by analyzing the evolution of the #flirt channel rules, as illustrated in Table 1: while the platform was initially more flexible with interactions in English, German, or French, a more restricted language policy was established in later years—this shift seemed to start in 2003 according to our analysis via the Wayback Machine, but we should treat this date cautiously taking into consideration the limitations and challenges of using web archives as a source[10]—implementing a "mandatory" use of Luxembourgish from 6h–24h to avoid "chaos and confusion." (Table 1). The inherent multilingualism introduced some tensions in the chat environment, particularly impacting the French-speaking community, which faced criticism for employing the French language in the chat, as explained

[8] Original text in German translated via DeepL.
https://persist.lu/ark:70795/j2d9wx/pages/16/articles/DTL429

[9] The percentage of foreign residents in the country increased from 29.7% in 1991, to 36.9% in 2001, reaching 43% in 2011. (Statec 2022)

[10] Web archives are not a mirror of the web as it was in the past. They are a re-construction, where multiple choices intervene Brügger 2008; Bachimont 2017). The very nature of web archives makes it challenging to use them as a source: an original is lacking; they are incomplete; they consist of a unique version, not a copy of the online web; and there is a temporal and spatial inconsistency between the archived fragments (Brügger 2018).

by Raoul Mulheims. In response to this issue, the chat management established a dedicated French-speaking channel, #francophone, as a remedial measure.

Table 1: Evolution #flirt channel rules

#Flirt channel rules 2000[11]	**#Flirt channel rules 2003**[12]
"We will try to keep Luxembourgish as the main language in the channel but French, English and German will be tolerated as long as the channel stays a "Luxembourgish one." We will NOT tolerate other languages under ANY circumstances, because that would result in an ultimate chaos with all this people. They may discuss as well in private". ("#Flirt guidelines," 2000)	"As the name implies, Luxusbuerg is a Luxembourgish chat. Therefore, we want in the #flirt that only Luxembourgish is written on the channel from 6h–24h. Then you can write English, French, and German, but even in measure. What we want to avoid in any case is that several languages are spoken on Channel 4, or that a language other than Luxembourgish prevails. That's just too much chaos and confusion. For those who want to speak French, there is the #francophone, or the private. And for all other languages the same applies: discuss in private". ("#Flirt Régelen", 2003)

For Caroline Dohmer, an Assistant Professor for Luxembourgish grammar and orthography at the University of Luxembourg, who was a moderator of the #flirt channel when she was 15 years old, Luxusbuerg played an essential role in the literacy and standardization of written Luxembourgish and in the generalization of the use of written Luxembourgish beyond text messages:

[11] Translation from Luxembourgish with Google Translate. Original text accessible via the Wayback Machine https://web.archive.org/web/20001206025800/http://www.luxusbuerg.lu:80/index.cgi?origin=luxusbuerg&site=flirt&lg=eng&display=gl

[12] Translation from Luxembourgish with Google Translate: Original text accessible via the Wayback Machine. https://web.archive.org/web/20030212121050/http://www.luxusbuerg.lu/

> No one used correct spelling. So, we just wrote in our own system. That is how Luxembourgish orthography is working until today. We have standardized orthography, but it is not taught in schools, so people do it as they think it is written (Dohmer, personal communication, June 13, 2023)

1.3 Continuity of IRC's social structures and organization

IRC's social structure functioned as a collective and simultaneous construction of users "unequally empowered" (Latzko-Toth 2014, 591) participating in the platform's design.

> Actors were shaped in tandem with the development of IRC. Its developers invented user categories such as operators, ordinary users, disruptive users, etcetera. New terms were created to designate different levels of channel operator status: auto-op, super-op, channel manager or founder, and so on, depending on the network, language, or specific channel culture. (Latzko-Toth 2014, 591).

Luxusbuerg's organization followed IRC's social structures, which persisted through the years. The website was organized and divided into channels, one of the main pillars of the structure of IRC: "Once connected to an IRC server, users join conversation spaces called channels, whose names are designated by the # sign." (Senft 2002, 258). We argue that social structures were so deeply rooted in the functioning of the platform that they persisted despite the changes, first when it became its own server in 1999, and secondly when it integrated a larger structure, the company Nvision, created by Luxusbuerg's owners in 2000.

IRC was based on a hierarchical system that often created tensions between the users and the operators. Elisabeth Reid defined the operators[13] as "people who have chosen to invest the time needed to set up and maintain the IRC program on their local machines for the benefit of other local users" (Reid 1991). They had the power to remove people from a channel if they misbehaved or to ban users if there was a significant offense, as defined in the channel rules. One of the most common conflicts within the IRC channels was the presumed abuse of power from operators towards users. The accusations of prejudice and injustice were frequent, and this conflict was no stranger to Luxusbuerg.

As explained by Raoul Mulheims, Luxusbuerg had its own regulation system, and actively monitoring the conversations was essential. Therefore, a significant role was given to the moderators in the platform. Luxusbuerg defined the operators as:

> Users whose job is to help people on the channel and make sure everyone follows the guidelines. They can change some channel settings (like the topic). They can remove users

[13] More about the role of the operators in Latzko-Toth (1998; 2014); Reid (1991); and "Undernetiquette and policies" (n.d.)

> from the channel if they misbehave (a 'kick'), as well as keep them permanently out of the channel (a 'ban') if they really act badly. ("What, who" 2000).

To become eligible, they needed to invest time and contribute substantially. If the operators started to kick out people without justification, they could lose their status. People offending others could never become eligible. Once they became operators, they needed to connect as much as possible, be wise, and mature in their behavior: "It became a club thing rather than anything else, and this also contributed to the quality of the experience, the user experience as we call it today," explained Raoul Mulheims.

Above the operators were the channel administrators[14], who were in charge of a specific channel and choosing its operators. The administrators were chosen by the top management, the co-founders, who eventually changed their roles after establishing their own company under which they managed Luxusbuerg ("What, Who" 2000).

1.4 Users co-shaping and co-creating the platform

According to Latzko Toth (2014), IRC was shaped as a collective construction of users actively participating in the platform's evolution. The users became the driving force behind the success of IRC by donating their time to the co-creation of the platform. They became actors, playing an essential role in the platform's governance, contributing to technical development, and reshaping functionalities based on their uses. Luxusbuerg users had been co-designing the platform since its origins. Most of the users, who were primarily students at the beginning, invested their time in helping and designing Luxusbuerg; it started as a community of users motivated by a common interest in technology. The more time the users spent, the more likely they were to become operators or administrators. Nevertheless, a significant change occurred in 2000, impacting the governance of the platform. Luxusbuerg, initially established as a non-profit entity, was transformed into a for-profit entity, selling advertisement space[15], following the creation of a proprietary company by its owners, namely Nvision. This strategic decision encountered resistance from the hardcore users who were against the platform becoming too commercial. This created some tensions in finding a balance between the free collaboration spirit from the origins and the commercial exploitation of the platform as a business. The resistance to this transformation may be

[14] More about the role of administrators in Latzko-Toth (2014; 2010)

[15] More about their advertisement strategy in "Quel type de publicité" 2002 https://web.archive.org/web/20020803184438/http://www.luxusbuerg.net/queltypepub.php

attributed in part to the pronounced influence of IRC's governance structure on Luxusbuerg's operational framework.

During the period spanning 2000 to 2002, the management of Luxusbuerg involved the participation of 120 individuals, while daily user connections ranged between 6,000 and 7,000. Notably, the platform continued to solicit assistance from users without offering financial compensation. In 2002, the Luxusbuerg Development Group issued a call for participation, seeking support for the technical advancement of Luxusbuerg:

> Luxusbuerg has created a new platform for all of you interested in the technical developments of Luxusbuerg. Your benefits: Ability to contribute to one of the biggest and technically advanced internet sites in Luxembourg/Work together in a team of more experienced programmers/Improve your skills. (Luxusbuerg 2002).

Also, in 2002, the platform launched a call for channel administrators, who had to submit their applications in teams of three. The solicitation outlined the following terms and conditions:

> You want to be responsible for a channel on Luxusbuerg? We give you the unique opportunity to apply as a team! Luxusbuerg allowed you to apply for a channel of your choice, where you can define the policies, rules, and activities. In order to join our team, we created this online submission page, where you can apply as a team for being responsible for a channel. ("New channels" 2002)

1.5 The hybrid model

One of the core elements of Luxusbuerg was the success of the hybrid model, where the traditional—normally based on a common geographic space preferably on a face-to-face mode of communication—and virtual communities merged, feeding back to each other. We have seen that offline meetings were not uncommon in IRC channels, and communities were both "technically mediated and constituted according to the possibility of face-to-face interactions" (Latzko-Toth 1998, 3). The users tended to build a consistent online persona through the pseudonym and the interactions, which acted as the basis for the beginning of the offline interactions. As explained by Bechar-Israeli, while users in IRC could change their pseudonyms—'nick' in IRC jargon—every few seconds if desired, the general trend was to keep them for a long time, becoming an identity-attached element. "The way to do so is to choose an original nick which conveys something about the person's 'self' and which will tempt other participants to strike up a conversation with that person" (Bechar-Israeli 1995). Jouët (1989) described the dynamic created by this virtual and offline construction of the self in her study of the AXE messaging system, where she explained how virtual participation played an essential role in

offline gatherings, with users calling each other by pseudonym, instead of by name.

The hybrid model was part of Luxusbuerg's success, as explained by Raoul Mulheims. “You meet online, and then you can say, let's meet this evening, and then we just get a drink,” which strengthened the club aspect, the camaraderie, friendship, or even sentimental relationship created among the members. “The elements of loyalty are, therefore, present and are constantly being developed.” (Mulheims, interview by author, March 22, 2022). Complementarily, Tom Weber, who was involved in the development of the connection to the chat through a web browser, as explained by Raoul Mulheims, stated in an interview published in 2000: “These are all successful because we are curious to discover the person who hides behind a pseudonym and to see if it corresponds to the image we have made.” (“Tom Weber” 2000)

Some of the encounters were integrated into the governance structure of the platform since there were different coordination meetings (management meetings, operator meetings, etc.). Others were opportunities for socialization beyond the virtual boundaries, from channel meetings with the members to massive Luxusbuerg parties with 600–800 people. These meetings strengthened the community spirit and improved the online experience, with the same users chatting on the platform and vice-versa. The first offline meeting took place in a pub in 1998; since then, similar ones have emerged in different locations. The first big party took place in December 1999, with 600 attendees (Luxusbuerg 1999), and was presented as an opportunity to go beyond the computer-based interactions:

> Luxembourg's online community is no longer confined to its computers: it's decided to get out there and celebrate by organizing a big party where we can finally meet in the flesh. And the party won't be held behind closed doors for the first time: everyone's invited.[16]

2. Navigating through the #flirt and #queer channel

2.1 The impact of the inherent characteristics of Luxusbuerg in the development of flirting

The user-centered approach reshaped the platform's user experience to the detriment of the hardcore users’ experience in some cases, as demonstrated by flirting becoming the most popular use of the chat, with around 400–500 people connecting at the same time daily. The hardcore users were disappointed because they felt the chat's essence was eroding

[16] Free translation from the author supported by DeepL. Original text in https://web.archive.org/web/20030730213049/http://www.luxusbuerg.lu/press/press_release_party_23-12-99.pdf

and they could not maintain the same level of conversation to which they were accustomed. In response, the management team created different channels when they launched the website to separate flirting from other conversations.

> It became so much about flirting that it was not possible anymore to run anything else in one single channel (…) so for the others, also for the guys that had been there for quite some time, I am not saying they were not interested in flirting at all, but also they wanted to have another conversation and be there because they knew each other well. (Mulheims, interview by author, March 22, 2022)

The hybrid model contributed to the success of the #flirt channel. This links with the observations made by authors such as Gyuillaume Latzko-Toth, Josiane Jouët, and Elizabeth Reid. Even though the users went through a pseudonym in the online environment, the objective was not to be anonymous but to create an online identity, which played an essential role in online flirting. Besides, the fact that the users were in geographic proximity and had many offline events contributed to the flirt's success. Billedo (2009) observes that when it came to flirting, many users preferred someone geographically close to them to chat because it was easier to meet in real life. Latzko-Toth (1998, 63) adds: "Any attempt at seduction in the 'virtual world' has, as an undertone, the hope of realization in 'real life'."[17]. That is why pseudonyms play such an essential role in many online communities, where they usually are indicative of the chatter's gender, age, or are unique to catch attention or sound familiar based on a shared culture to engage with other users. As Velkovska (2002, 206) explains, "the pseudonym is both a resource of the interactive strategy and a product of the exchange. It is a label that interlocutors temporarily associate with a person and thus constitutes the first element of self-typing and typification of others"[18].

2.2 Close reading of the #flirt channel

#Flirt was the most popular channel that lasted the platform's lifetime since 1999. It became a channel in its own right even before the creation of luxusbuerg.lu, as flirting was already one of the main uses as a channel at IRC Undernet. The channel was organized with a landing page and a navigation menu allowing access to features such as forum, message board, channel guidelines, initiatives such as Flirt Girl and Flirt Boy of the Month, and Miss and Mr. Flirt, in which the users could participate by sending their

[17] Free translation made by the author with the support of deepL. Original text: « toute démarche de séduction dans le "virtuel" comporte, en filigrane, l'espoir d'une concrétisation dans la "vraie vie" »

[18] Free translation made by the author with the support of DeepL. Original text accessible via the Wayback Machine. https://web.archive.org/web/20030730213049/http://www.luxusbuerg.lu/press/press_release_party_23-12-99.pdf

picture and description with their nickname, hobbies, favorite food, favorite music, among others. The news, announcements of parties, meetings, changes of channel operators, etc., were displayed on the channel's homepage. The community spirit and the participatory style were present throughout the channel's history. They announced the parties and events, shared pictures, and, as of 2006, included a short survey to take the users' opinions on channel improvements into account.

The initial archived record of #flirt, dating back to May 10, 2000, introduces the channel with the following description:

> "Are you single/lonely?You like to flirt?
>
> You want have fun?
>
> You like to talk to / meet new people?
>
> So don't hesitate, join the #Flirt-Channel by clicking here.
>
> Please respect our guidelines. Thanks :-)
>
> And remember: You do NOT have to be single to flirt!"

The last version from the #flirt channel was archived on September 22, 2009. One only needs to navigate through the different versions from May 10, 2000 to September 22, 2009 to see the similarities and differences. It showed a completely different look and feel that evolved together with the platform. Essentially, the spirit retained its original essence—a club atmosphere where updates about new administrators and operators, as well as farewells, were also featured. One of the main changes was with regard to security issues, as demonstrated by the channel rules, which became more specific as the platform grew. It likely evolved due to the lessons learned from various emerging negative practices, as evident when comparing the lists of prohibited actions and behaviors in 2000 and 2007. It is important to remember that Luxusbuerg served as an early example of a participatory website, with users and owners learning together through hands-on experience. Another example of the increased attention to security issues was the creation of the #adultflirt channel at some point in 2002 to separate the interactions of teenagers and adults. However, there were no measures to prevent someone from subscribing to a channel except for detecting misconduct based on the channel rules. For example, in the #adultflirt rules it was highlighted that minors were "banned immediately" and no nicks with numbers below 18 were allowed. ("Adultflirt Channel Regelen" 2007)

Table 2. List of forbidden behaviors

List of forbidden actions/behaviors 2000[19]	List of forbidden things/behaviors 2007[20]
No colors No repeating No flooding No insulting No Denial of Services attempts (like warscripts, nuking, ping-flooding, Trojan, or any other abuses) No advertising No scripts No clones (only line sharing with other identd will be allowed) No sexual harassment Don't threaten the Channel Operators No begging for Channel Operator No Sport events ("#Flirt guidelines", 2000)	Nicks no nicks that mean nothing (e.g.: kjashfdih) no racist nicks no nicks with sexual misconduct (many minors in #flirt) the OP has the right to self-assess when these points apply General rules no colors (not automatically recognized by the system) no bold / underlined / italic. This is mainly reserved for OPS and ADMINS no Caps Lock (capital letters) no repetition (no more than 5 times within 3 min.) keen CTCP / text flooding no insults vis-à-vis other chatters no insults to Operators & Admins. no sexual harassment do not talk too much about sports on the channel (privately with no problem) no CS (Cyber Sex) keng War / Fun / Mp3 / ... etc. scripter (nuking, icmp flooding, etc.) no URLs from other chats or sites that sell anything! (the only URLs that are allowed are from sites that have something to do with Luxusbuerg (partners) or by luxusbuerg itself (e.g. images that can be found on Luxusbuerg.lu etc.)) Do not write numbers, emails in the channel! no Idling (not more than 3 hours) no bots (eg like FlirtBOT, ChatBOT, CyberBOT.) no clones (except for connection sharing with different IDs) not for Channel Operator Status ("#Flirt Régelen", 2007)

[19] "Flirt Guidelines", 2000
https://web.archive.org/web/20001206025800/http://www.luxusbuerg.lu:80/index.cgi?origin=luxusbuerg&site=flirt&lg=eng&display=gl

[20] "Flirt Régelen", 2007
Translation from Luxembourgish with Google Translate. See original text:
https://web.archive.org/web/20070807094739/http://www.luxusbuerg.lu/index.php?tab=content&channel=flirt&ContentID=67

Based on the limited messages retrieved from forums and message boards accessible through various archived versions obtained from the Wayback Machine, it is evident that the majority of these communications were concise and conducted in Luxembourgish, with occasional entries in English. Both the message board and forum discussions appeared as somewhat disordered dialogues, frequently lacking coherence or continuity. However, it is imperative to acknowledge the inherent constraints when evaluating the extent of widespread public engagement based on a scant selection of excerpts retrieved via the Wayback Machine. This is particularly true for forums and message boards, where immediacy plays a key role, akin to the dynamic nature of social media. Web archives may fall short of capturing the fluidity and real-time interactions intrinsic to these platforms[21]. Indeed, Ortner, Sinner, and Jadin (2018) considered IRC to be one of the precursors of modern social media and one of the ancestors of instant messaging. Moreover, we cannot forget that one-to-one conversations played a fundamental role on the platform, being the preferred way of interaction for flirting, established privately, hence not harvested due to privacy issues. Some examples of the interactions found in the message boards are calling attention to a particular person, like "I love you, Nancy" or to the general channel users, such as "moien" in the message board. In the forum, messages were devoid of titles and arranged chronologically, with the most recent ones appearing first. Many of the messages served as calls for attention to create conversations, such as messages generating expectations about a user's identity after a nickname change (Table 4) or the self-promotion of websites (Table 4). Channel rules regulated such practices, particularly the promotion of other chats and commercial websites, which was strictly prohibited.

Table 3. Message board

Date	Nick	Subject	Message
2000-04-16	Dexio15	I love you Nancy	Nancy, I just wanted to tell you how much I love you![22]

[21] More about the challenges of archiving social media in Marshall and Shipman (2012), Vlassenroot et al. (2019), Byrne (2017; April 2018), among others.

[22] Translation from DeepL. Original text archived via the Wayback Machine on 27 August 2002. https://web.archive.org/web/20020827150801/http://www.luxusbuerg.lu/index.cgi?origin=luxusbuerg&site=flirt&lg=eng&display=bbs&bbs=flirt&action=message&id=13 Last accessed 19 December 2023

2000-03-12	exit	chat meeting	How about a chat meeting soon? any suggestions, proposals, etc.? Purely out of curiosity .. what sort of environment do *you* think is good / appropriate for a chat meeting? [23]
2000-09-23	dragola	moien	hello dear why don't you chat with me[24]

Table 4: Forum[25]

Nance Huh!! Wanted to immortalize me here for a while =) hmmm... that's what you meant.. :) until then then CU!! Ciao (2001-04-17 – 33)
ZERO-NICE From the looks of it, no one has come here to look for a long time, but still! please type View all my homepages: http://zero-nice.da.ru (2001-03-25 - 32)
Flatterma hello, I greet all of you from the chat who know me, but seeing that I recently changed my nickname, no one knows who I am, so just take all my greetings to the chat "FLIRT"!!!

2.3 Flirting and the queer

In 2005, Luxusbuerg opened the #queer channel. It was managed in collaboration with the association Rosa Lëtzebuerg, a national LGBTIQ+ organization in Luxembourg. Raoul Mulheims explained that the channel succeeded especially in creating a sense of community, a complementary place of integration, as stated by the welcome message: “The crew of #queer welcomes you all on their journey to a slightly different world, a world where man with man, woman with woman, and man with woman are

[23] Archived via Wayback Machine on 4 July 2002 https://web.archive.org/web/20020704222550/http://www.luxusbuerg.lu/index.cgi?origin=luxusbuerg&site=flirt&lg=eng&display=bbs&bbs=flirt&action=message&id=8 Last accessed 19 December 2023

[24] Translation from Google Translate. Original text in Luxembourgish: https://web.archive.org/web/20020704032451/http://www.luxusbuerg.lu/index.cgi?origin=luxusbuerg&site=flirt&lg=eng&display=bbs&bbs=flirt&action=message&id=37

[25] Translation from Google Translate. Original text in Luxembourgish: https://web.archive.org/web/20020615142519/http://www.luxusbuerg.lu/index.cgi?origin=luxusbuerg&site=flirt&lg=eng&display=forum 3

welcome and where there are no borders."[26] As with other IRC channels, the chat became a complementary place of integration. For Raoul Mulheims, this channel contributed to the identity-building of the LGBTIQ+ community in Luxembourg:

> It was something for the coming-out stage. It's exactly what you need, first thing you get in touch, you do not go to other real-life meetings and so on, for people not knowing about your sexual identity, and trying out and getting in touch with peers. The guys from the association told us that it was a really great thing that we had this forum because, under the protection of anonymity, there were a lot of people. (Mulheims, personal communication, 2022).

This vision aligns with studies such as Chaplin's (2014) on the Minitel and how "virtual spaces" contributed to creating "new forms of lesbian identity untethered to specific locations, organizations, embodiment, or proximity" (Chaplin 2014, 452). As stated by Dame-Griff (2023), the internet shaped transgender identity and activism from the 1980s to the present, and trans people online exploited different digital infrastructures in the early days of the internet to build a community. For the author, the development of the internet and the transgender life histories and identities are inextricably linked.

Through close analysis of the #queer channel, we could see that the rules required users to be at least 16 years old to connect to the channel ("Queer's rules," August 20, 2007). Based on our analysis, the only channel that stipulated an age restriction was the #adultflirt for individuals 18 years and older. Apart from that, the #queer channel followed the same dynamic as the others, with news about channel parties, plans, and updates on administrators and operators:

> "Barbecue on the Weiswampecher Lake on the 7th of August (2005)
>
> Finally, it is time for the #Queer to have its first meeting, and indeed on the 7th of August in Weiswampech on the lake, with grilling, swimming, and having fun. Everyone is welcome; you are also welcome to bring your colleagues, children, or pets. The day would only fail if the weather wasn't good. Everyone must bring their own food and drink. A grill is definitely provided. For further information, email us at Queer-admins@luxusbuerg.lu We also know how to drive with carpools, so we'll see everything until then.[27]
>
> As the title says, we have a new admin in Queer today. A warm welcome to PetitPrince. Continue to have a lot of fun and great work together in Queer. [13/05/05 @ 23:47]
>
> Hello, two ladies have arrived. Wildgemse as operator and Boogie as supporter. Have a lot of

[26] According to the information saved on the Wayback Machine, the queer channel was created in January 2005, although we cannot confirm the creation date. Translation from Google Translate . Original text archived via the Wayback Machine on 7 April 2005. https://web.archive.org/web/20050407225811/http://www.luxusbuerg.lu/index.php?tab=news&channel=queer&PHPSESSID=14379147775969f756bdd0166904e318 Last accessed 19 December 2023

[27] Translation from DeepL. Original text on https://web.archive.org/web/20051029194413/http://www.luxusbuerg.lu/?p=844

fun. [06/07/05 @ 20:19] [28]

Consistent with the pattern observed in other channels, #queer also included a section on statistics with data pertaining to recent activities. This encompassed information on the most and least talkative users and what they called "the big numbers," in which they highlighted the users' more relevant behaviors with a certain sense of humor. For instance, "lucy_maus could not decide whether to go or stay and visited the channel 5,592 times. time-lu did not agree either, 5,442 out and in :)" or "wildgemse wanted to tell others what he/she was doing—6,651 descriptions. Furthermore Hey_Mr_Dj—5,805 descriptions"[29].

Conclusion

Through this study, we have seen that the origins of the chat, as part of IRC Undernet, with its social structure so profoundly rooted in the functioning of the platform, exerted a profound influence on its evolutionary trajectory, as well as on user roles and expectations. Throughout its history, first as a channel on Undernet IRC in 1996, then becoming a web server in 1999 and transforming itself into a go-for-profit under the company Nvision in 2000, the users remained active participants who reshaped and redefined Luxusbuerg by designing, programming, and co-creating the governance rules as the platform evolved.

The evolution of the platform faced some tensions with hardcore users. The first issue was discontent among hardcore users at the direction the platform was taking, due to the growing prevalence of flirt use. While users played an essential role in reshaping the platform as it expanded, some hardcore users felt that they were shifting away from the platform's original purpose of fostering connections and engaging in conversations among friends. To keep both user groups engaged, the management team decided to create the website following the same channel-oriented structure as IRC with distinct channels for various interactions, separating the flirting use from other conversations. The second issue was with the platform's shift towards commercial goals. Recognizing the potential business benefits, the owners transitioned Luxusbuerg into a for-profit venture, creating tensions among the hardcore users and as they strove to balance not-for-profit ideals

[28] Translation from Google Translate. Original text on https://web.archive.org/web/20051029195503/http://www.luxusbuerg.lu/index.php?tab=news&channel=queer&PHPSESSID=0ff97dddac74e143a0373519e8973437 archived via the Wayback Machine on 29 October 2005. Last accessed 3 January 2024.

[29] Translation from Google Translate. Original text on https://web.archive.org/web/20051215204333/http://www.luxusbuerg.lu:80/index.php?tab=toptalkers&channel=queer&PHPSESSID=24ad9ce66776ab985a03b90edd2e995d archived via the Wayback Machine on 15 December 2005. Last accessed 4 January 2024.

with commercial success. Maintaining a delicate equilibrium between commercial exploitation and soliciting non-remunerated user collaboration reminiscent of the platform's early years posed a significant challenge for the management team.

We have seen that the hybrid model—combining online and offline interactions— has been essential to the platform's success. It laid the foundation for a robust 'community' spirit and contributed to the success of flirting on the platform, strengthened by the country's spatial specificities. Notably, for the LGBTIQ+ community, the chat network became a complementary space of inclusion and played a crucial role in fostering sociability and shared experiences.

Finally, the history of this website and its analysis with the support of web archives as a source has helped us contextualize and shed light on contemporary practices related to users' interactions, flirting, and online dating. As Raoul Mulheims explained during his interview, with chat platforms such as Luxusbuerg emerged the essence not only of today's online dating, but also the foundations on which the whole social communications part has been built up.

References

Bechar-Israeli, Haya. 2006. “FROM 〈Bonehead〉 TO 〈cLoNehEAd: NICKNAMES, PLAY, AND IDENTITY ON INTERNET RELAY CHAT1.” *Journal of Computer-Mediated Communication* 1, no. 2. https://doi.org/10.1111/j.1083-6101.1995.tb00325.x.

Bachimont, Bruno. 2017. “L’Archive Du Web : Une Nouvelle HerméNeutique Des Traces ?” *Web Corpora*. June 21. https://webcorpora.hypotheses.org/288.

Brügger, Niels. 2008. “The Archived Website and Website Philology: A New Type of Historical Document?” *Nordicom Review* 29, no. 2: 155–175. https://www.nordicom.gu.se/sites/default/files/kapitel-pdf/270_brugger.pdf

Brügger, Niels. 2018. *The Archived Web: Doing History in the Digital Age*. Cambridge, MA: MIT Press.

Byrne, Elisabeth. 1994. “Cyberfusion: The Formation of Relationships on Internet Relay Chat. Introduction.” Irchelp. . https://www.irchelp.org/communication-research/academic/byrne-e-cyberfusion-1993/thesis1-intro.html

Byrne, Elisabeth. 2017. “The Challenges of Web Archiving Social Media.” Irchelp. December 4. https://blogs.bl.uk/webarchive/2017/04/the-challenges-of-web-archiving-social-media.html

Chaplin, Tamara. 2014. “Lesbians Online: Queer Identity and Community Formation on the French Minitel.” *Journal of the History of Sexuality* 23, no. 3: 451–72.

Dakhlia, Jamil and Géraldine Poels. 2012. “Le Minitel Rose : Du Flirt éLectronique... Et plus, Si AffinitéS. Entretien Avec Josiane JouëT.” *Le Temps Des MéDias*, 2 no.12: 221–28. https://doi.org/10.3917/tdm.019.0221.

Hinner, Kajetan. 2000. “Statistics of Major IRC Networks: Methods and Summary of User Count.” *M/C Journal* 3 no. 4. https://doi.org/10.5204/mcj.1867

Dame-Griff, Avery. 2023. *The Two Revolutions: A History of the Transgender Internet*. New York, USA: New York University Press.

IPSE – Identités Politiques Sociétés Espaces, 2014. *Doing Identity in Luxembourg: Subjective Appropriations – Institutional Attributions – Socio-Cultural Milieus*, edited by IPSE – Identités Politiques Sociétés Espaces. 1. Aufl. Bielefeld: transcript Verlag.

Jones, Quentin. 1997. “Virtual-Communities, Virtual Settlements and Cyber-Archaeology: A Theoretical Outline.” *Journal of Computer-Mediated Communication* 3, no. 3: JCMC331. https://doi.org/10.1111/j.1083-6101.1997.tb00075.x

Klein, Charlie and François Peltier. 2022. “La démographie luxembourgeoise en chiffres 2022.” Statec, November 24. http://statistiques.public.lu/fr/publications/series/en-chiffres/2022/20220511.html

Latzko-Toth, Guillaume. 1998. “À La Rencontre Des Tribus IRC: Le Cas d’une Communauté d’usagers Québécois de l’Internet Relay Chat.” Mémoire Présenté Comme Exigence Partielle de la Maîtrise en Communication Université du Québec à Montréal, Université du Québéc. https://archivesic.ccsd.cnrs.fr/sic_00461232

Latzko-Toth, Guillaume. 2010. “La Co-Construction d’un Dispositif Sociotechnique de Communication : Le Cas de l’Internet Relay Chat.” Doctoral Thesis, Université du Québec à Montréal. https://theses.hal.science/tel-00543964.

Latzko-Toth, Guillaume. 2008 “L’Internet Relay Chat : Un Cas Exemplaire de Dispositif Sociotechnique.” *Commposite* 4, no. 1: 52–73.

Latzko-Toth, Guillaume. 2014. “Users as Co-Designers of Software-Based Media: The Co-Construction of Internet Relay Chat.” *Canadian Journal of Communication* 39, no. 4: 577–96. https://doi.org/10.22230/cjc.2014v39n4a2783.

Liu, Geoffrey Z. 1999. “Virtual Community Presence in Internet Relay Chatting.” *Journal of Computer-Mediated Communication* 5, no. 1: JCMC514. https://doi.org/10.1111/j.1083-6101.1999.tb00334.x

Luxusbuerg. “#Adultflirt Channel Regelen.” August 20, 2007, https://web.archive.org/web/20070820152019/http://www.luxusbuerg.lu/index.php?tab=content&channel=adultflirt&ContentID=79&PHPSESSID=fa7f8ecf5b33a04975fb85ee427fb4dc Retrieved from Internet Archive. Last access 3 January 2024.

Luxusbuerg. “Development Group.” February 8, 2002. https://web.archive.org/web/20020208164404/http://www.luxusbuerg.lu/index.cgi?origin=luxusbuerg&site=root&lg=eng&display=developmentgroup. Retrieved from Internet Archive. Last access 30 May 2022.

Luxusbuerg. “#Flirt guidelines, December 6, 2000.” https://web.archive.org/web/20001206025800/http://www.luxusbuerg.lu:80/index.cgi?origin=luxusbuerg&site=flirt&lg=eng&display=gl Retrieved from Internet Archive. Last access 20 December 2023.

Luxusbuerg. “#Flirt Régelen, February 12, 2003.” https://web.archive.org/web/20030212121050/http://www.luxusbuerg.lu/ Retrieved from Internet Archive. Last access 7 January 2014

Luxusbuerg. “Luxusbuerg, c'est quoi ?” August 3, 2002 https://web.archive.org/web/20020803183510/http://www.luxusbuerg.net/infoluxusbuerg.php Retrieved from Internet Archive. Last accessed 28 May 2022

Luxusbuerg. “#queer TopTalkers.” December 15, 2005. https://web.archive.org/web/20051215204333/http://www.luxusbuerg.lu:80/index.php?tab=toptalkers&channel=queer&PHPSESSID=24ad9ce66776ab985a03b90edd2e995d Retrieved from Internet Archive. Last access 4 January 2024.

Luxusbuerg. “Quel type de publicité sur Luxusbuerg ?” August 3, 2002, https://web.archive.org/web/20020803184438/http://www.luxusbuerg.net/queltypepub.php

Luxusbuerg. “New Channels on Luxusbuerg.” February 14, 2002. https://web.archive.org/web/20020214021831/http://www.luxusbuerg.lu/index.cgi?origin=luxusbuerg&site=root&display=newchannelsjan02&lg=eng. Last access 30 May 2022.

Luxusbuerg. “Wannna go virtual?” December 23, 1999. https://web.archive.org/web/20030730213049/http://www.luxusbuerg.lu/press/press_release_party_23-12-99.pdf Last access 6 March 2024.

Luxusbuerg. “What, Who and How.” July 7, 2000. https://web.archive.org/web/20000707235127/http://www.luxusbuerg.lu/index.cgi?origin=luxusbuerg&site=root&lg=eng&display=whatwhoLast accessed 17 December 2023

jls. “Luxusbuerg : «chatter» En Luxembourgeois.” *D’Lëtzebuerger Land*, October 29,

1999. https://persist.lu/ark:70795/cmx0np/pages/42/articles/DTL531
Explorator. “Luxusbuerg IRC ChatHomepage. Entretien Avec Christophe Leesch.” 1999.
Marshall, Catherine C., and Frank M. Shipman. 2012. “On the Institutional Archiving of Social Media.” In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, 1–10. JCDL ’12. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2232817.2232819
Mirashi, Mandar. 1993. “The History of the Undernet.” August 1993. https://www.ibiblio.org/pub/academic/communications/irc/undernet/undernet-history Accessed 17 December 2023
Mousel, Pierre, and Jérôme Lulling. 2002. “Internet: Gefahr Oder Hilfe Für Das Luxemburgische?” *D’Lëtzebuerger Land* 49, no. 51/52 (December 20, 2002): 16–17. https://persist.lu/ark:70795/j2d9wx/pages/16/articles/DTL429
Reid, Elizabeth. 1991. “Electropolis. Communication and Community on Internet Relay Chat. Adapted from an Honours Thesis Written at the University of Melbourne (Australia) in 1991.” IRCHelp.org. https://www.irchelp.org/misc/electropolis.html Accessed May 16, 2023.
Senft, Theresa. 2003. “Internet Relay Chat.” In *Encyclopedia of New Media: An Essential Reference to Communication and Technology*. ed. Steve Jones, 256–258. Thousand Oaks, London, New Delhi: Sage.
“Undernetiquette and Policies,” n.d. https://www.ibiblio.org/pub/academic/communications/irc/undernet/Undernetiquette-and-policies. Last modified 1993. Last accessed 17 December 2023.
Stenberg, Daniel. 2002. “History of IRC (Internet Relay Chat).” *Daniel Stenberg* (blog). https://daniel.haxx.se/irchistory.html
Explorator. “Tom Weber. Say It Again Tom!” 2000.
Velkovska, Julia. 2002. “L’intimité anonyme dans les conversations électroniques sur les webchats.” *Sociologie du travail* 44, no. 2: 193–213. https://doi.org/10.4000/sdt.32951
Vlassenroot, Eveline, Sally Chambers, Emmanuel Di Pretoro, Friedel Geeraert, Gerald Haesendonck, Alejandra Michel, and Peter Mechant. 2019. “Web Archives as a Data Resource for Digital Scholars.” *International Journal of Digital Humanities* 1, no. 1: 85–111. https://doi.org/10.1007/s42803-019-00007-7

The Online presence of the Danish public sector from 2010 to 2022: Generating an archived web corpus

Tanja Svarre, Mette Skov

Abstract: This chapter presents the generation of a web archive corpus with the purpose of studying the development of Danish public sector websites from 2010 to 2022. Websites constitute an important element in shaping electronic governments. Few studies have carried out longitudinal studies of government websites based on archived web materials. In this study, which spans three levels of administration, archived web data is gathered to analyze governments at the local, regional, and national levels in Denmark across four selected years.

Keywords: public sector, online information, e-government, web archives.

1. Introduction

E-government designates the use of technology to improve access to information and services for government stakeholders, such as citizens, companies, and fellow government administrations at the local, national, and international levels (Layne and Lee 2001). The term "e-government" started to emerge in the late 1990s with the growth of the internet (Grönlund and Horan 2005; Skiftenes Flak et al. 2009).

Various models have attempted to map the stages of e-government development, indicating that governments progress at different rates (Rooks, Matzat, and Sadowski 2017). According to Layne and Lee (2001), e-government begins as the initial stage, which is characterized by a presence akin to a catalog; it then progresses to the transaction stage, where tasks such as online form submissions become possible, followed by the vertical integration of information and systems and, finally, the horizontal integration of information, which is the highest stage. Recently, there has been a shift in the e-government literature toward smart government, suggesting that with smart technologies, digital governments can offer even more advanced solutions to their stakeholders (Lemke et al. 2020; Hujran et al. 2023).

Prior studies have highlighted factors that both facilitate and hinder the progress of e-government. At a broad level, these factors include technological, organizational, and environmental aspects (Zhang, Xu, and Xiao 2014) and, more specifically, finance, socioeconomics, politics, the government level and type, management practices, digital skills, cultural

Tanja Svarre, Aalborg University, Denmark, tanjasj@ikp.aau.dk, 0000-0002-5468-0406
Mette Skov, Aalborg University, Denmark, skov@ikp.aau.dk, 0000-0002-8821-0314

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Tanja Svarre, Mette Skov, *The Online presence of the Danish public sector from 2010 to 2022: Generating an archived web corpus*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.17, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 185-197, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

models, and the social environment (Ronchi 2019; Ingrams et al. 2020; Cahlikova 2021). The diversity of these factors emphasizes the complexity of e-government implementation, which is not necessarily a straightforward process.

A central element of e-governments are government websites, where governments share information with, offer services to, and interact with citizens (Sandoval-Almazan and Gil-Garcia 2012). Like the general e-government maturity models presented above, the authors have also brought forward models to characterize government websites in particular. One categorization, suggested by Fan (2018), operates at the following five levels: 1) websites as one-way communication channels for information; 2) support for two-way communication, such as through SMS, mobile apps, and social media; 3) transaction-enabled websites that enable citizens to carry out various transactions; 4) citizen participation, whereby citizens can provide input and participate in polls; and 5) one-stop portals, through which citizens can access services across governments and functionalities. An alternative categorization was offered by Fietkiewicz, Mainka, and Stock (2017) that operates with pillars, where pillars 1–3 correspond to Fan's (2018) levels 1–3 but pillar 4 concerns the operability and integration of services, while pillar 5 relates to e-participation for citizens. From these frameworks, it is clear that the assessment of e-government also takes place at the specific website level.

Like many countries, Denmark has embraced the challenge of developing its government's digital dimension. Renowned for its high level of digitalization (Pedersen 2018; Flensburg and Lai 2021), the country has consistently ranked as the world's most digitalized nation over the past three years (United Nations 2023). While government websites may not be the most frequently used platforms by citizens (Flensburg and Lai 2021), they are the core drivers of e-government initiatives. This chapter aims to discuss the methodological challenges related to the construction of a corpus for studying the online public sector in Denmark and to contribute to our understanding of the progression of digitalization in Denmark's public sector from 2010 to 2022. Archived web data is our primary source, facilitating comparative studies over time. The following research questions guide our inquiry: 1) How can a corpus be established to enable longitudinal studies of e-government development? and 2) What were the overall characteristics of Danish e-government websites in the period from 2010 to 2022?

2. Related work

Several longitudinal studies have examined the development of public sector websites over time. Within the field of internet history and web

archiving, only two studies have specifically focused on public sector websites (Schafer 2017; Raffal 2018). The following section presents an overview of these two studies, but first, we introduce other longitudinal work on public sector websites. The research aims and methodological approaches used in these studies are described to provide a contextual background for the present study.

The general methodical approach to studying e-government websites over time involves variations in inspection methods. These inspection methods can manifest in different forms, but they generally involve one or more evaluators conducting an assessment of one or more interfaces, typically based on a predefined set of criteria (Nielsen 1994; Hollingsed and Novick 2007). Some studies have adopted established government indexes that enabled comparisons between government units, typically across different countries, while others have chosen to conduct their own assessments of websites. Moreover, the focus and theoretical framework guiding the evaluations can vary. We elaborate on these studies below.

In an early study by Shi (2006), the accessibility of 30 Chinese provincial government websites and eight Australian state-level websites were compared in late 2004 and again in the autumn of 2005. Guided by a predefined set of assessment criteria, this research aimed to evaluate accessibility for people with disabilities. Most Chinese websites exhibited severe accessibility issues, while all but one of the Australian governments had only minor problems. This was surprising, given China's favorable ranking in an international e-government assessment at the time. Between the two data collection periods, no changes were observed on Australian websites, while one Chinese government reduced its accessibility problems in the repeat assessment in 2005. Lazar and colleagues (2013) also focused on accessibility in their examination of government websites in Maryland in 2009 and 2012. Fifteen webpages were manually evaluated in 2009, and 25 in 2012. A comparative analysis between the two years showed a slight improvement in the number of violations related to accessibility issues. The numbers were divided between pages with no violations (one out of 25 in 2012) and those with a significant number of violations. Most violations were found on websites that did not use a state template, suggesting that adopting a standardized template could be a means of improving accessibility.

Another study used an 86-item checklist to analyze the evolution of Spanish public hospital websites between 2005 and 2008 (García-Lacalle, Pina, and Royo 2011), finding that there were more hospitals with websites at the end of the study period than at the start. However, due to the substantial number of hospitals still lacking an online presence in 2008, the authors concluded that online information was not a priority for many hospitals during that period. Among the hospitals with a web presence, the

predominant focus was on providing information rather than facilitating interaction. Furthermore, the study identified influencing factors that determined whether hospitals had an online presence. These included the size of the hospital, the level of managerial freedom, and, notably in 2008, the extent of pressure from the outside world.

Garcia-Murillo (2013) also studied web presence, government effectiveness, and accountability over time, aiming to discern their impact on the perception of corruption. The analysis was based on various international data and indicators of corruption, governance, regulatory quality, web presence, and others, covering 208 countries over the years 2002–2005 and 2008. The findings showed that web presence, along with the effectiveness of government and accountability, had a positive impact on the perception of corruption in the years studied.

In their study examining the online presence of smaller cities, Feeney and Brown (2017) manually evaluated 500 municipal government websites in both 2010 and 2014 based on a predefined protocol that emphasized information, e-services, utility, transparency, and civic engagement. Over the two years considered in the study, it was found that all five parameters showed improvement from the first to the second evaluation in both basic and advanced features on the municipal websites. Similar to the findings of Lazar et al. (2013), Feeney and Brown's (2017) statistics revealed a significant variation among municipalities.

Epstein (2022) also studied local US governments but focused on larger municipalities with populations exceeding 50,000. The research was taken from four years between 2000 and 2019. The data collection was built upon surveys sent out to selected municipalities during these years, identifying e-government characteristics such as media use and the available services. It was found that municipalities had initially been slow to start several services, but the pace accelerated in the latter half of the period. The author found a correlation between city size and the adoption of e-government services, while median income or poverty rates in the municipalities did not have a significant effect.

Ingrams et al. (2020) used an established e-governance performance index and manual inspection to investigate the world's 80 largest cities in 2003, 2009, and 2016. A cluster analysis of the measurement variables showed differences in the studied cities at different stages of e-government development. The results indicated that GDP, population size, and regional competition influenced all identified development stages, while democratic levels mainly impacted higher development stages.

In the realm of archived web data, few longitudinal studies have focused on public sector websites. Schafer (2017) analyzed the French state, drawing from various sources, such as The Internet Archive, French web archives, and other news-related archives, newsgroups, interviews, and state

reports. Through several core events in the 1990s, including the launch of websites for entities such as the Ministry of Industry, the Louvre Museum, and the French Railways, the study explored political initiatives and reactions to regulate the emergence of the World Wide Web in France. It delved into how France had adopted and appropriated the web, considering the influence of French culture and values. In a subsequent study, Raffal (2018) used data from the UK Web Archive to study the evolution of the British Ministry of Defence and Armed Forces over a five-year period from 1996 to 2013. The corpus was supplemented with additional relevant sources to provide contextual perspectives on the development observed on the collected websites. Focusing on online communication with citizens, particularly in the context of recruiting members for the British Army, the analysis centered on website content and link structures. The author found a distinct focus on recruitment, noting changes in the terminology used to recruit new soldiers over time. In addition, the number of channels used for recruitment increased throughout the period of study. Furthermore, the Ministry of Defence was found to have used the web to shape its agenda and manage public perceptions.

As shown in this review, and highlighted by Epstein (2022), only a few studies have empirically investigated the development of e-government over time. These studies often explored similar levels of administration, such as municipalities, cities, or hospitals. While many were based on various inspection and survey methods, their use of web archive materials was limited. Rather than using surveys and inspection methods, this chapter outlines the creation of a web archive corpus to understand the development of e-government across the public sector in Denmark.

3. Establishing the corpus

The aim of this paper is to present the generation of a web corpus for studying online public information following the administrative structure of Denmark across the national (government), regional (regions), and local (municipalities) levels (Gjerding 2005; Chatzopoulou and Poulsen 2017). Following a structural reform in 2007 (Vrangbæk 2010), the country is now divided into five regions and 98 municipalities. At the national level, the number of ministries and ministry administrations varies according to specific governments and their focus, making it subject to change.

In Denmark, the Danish national web archive, *Netarkivet*, handles the collection and curation of Danish web resources through various types of crawls, including broad, selective, and event crawls (Schostag and Fønss-Jørgensen 2012; Brügger, Nielsen, and Laursen 2020). To cover the entirety of the public Denmark online, the corpus consists of two components. National-level websites undergo a selective crawl covering ministries,

government agencies, and related websites conducted four times a year. Regional- and local-level administrations are not included in this selective crawl; instead, they are captured in a broad crawl, which is also carried out four times a year. Upon data delivery from *Netarkivet*, an ETL description is carried out. ETL stands for extract, transform, and load, explicating the data extracted from, in this case, a web archive for a specific purpose (Fage-Butler, Ledderer, and Brügger 2022). In this study, the ETL description specified that the corpus should include the entire selective crawl for ministries and government agencies, along with all regions and municipal websites from the broad crawl identified by web addresses. The crawls were defined for spring 2010, 2014, 2018, and 2022. *Netarkivet* provided the data in WARC files (Maemura 2023) stored on a Linux server with a Solr interface to facilitate the search. The WARC format enabled sub-extracts from the corpus based on specific metadata for analysis purposes and offered insights into the nature of the corpus.

However, defining and indexing the corpus does not guarantee the inclusion of all relevant content, as various authors have highlighted when working with archived web data. Incompleteness is expected due to curatorial and technical challenges, coupled with the considerable time spent harvesting large datasets (Brügger, Nielsen, and Laursen 2020). Consequently, sites may be captured with varying degrees of completeness, or may not be captured at all, and thus appear as blind spots (Maemura 2018; Raffal 2018; Donig et al. 2023). We consider this condition as we transition into the analysis phase of this chapter.

4. Preliminary results

Data analysis can be approached in different ways for large web archive corpora, as outlined by Nielsen (2021). These methods include 1) variations in measuring numbers and sizes; 2) conducting text analysis focusing on word frequencies, languages, and topic modeling; 3) link analysis to identify networks between corpora websites using outgoing links; and 4) searching the source code for insights such as domains. This chapter focuses on the first type by analyzing file and domain types across the four selected years. Future papers will present both text and link analysis, along with topic modeling (Murakami et al. 2017).

Prior to the removal of duplicates, the raw corpus consisted of a total of 16,695,320 files (Fage-Butler, Ledderer, and Brügger 2022). Table 1 illustrates the distribution across the four years, with 2018 having the smallest count, at 2,917,777 files, while 2014 recorded the highest, at 4,947,301 files. The numbers indicate that fewer files were shared online and harvested in 2018 compared to the other years.

Figure 1. Distribution of files in 2010, 2014, 2018, and 2022.

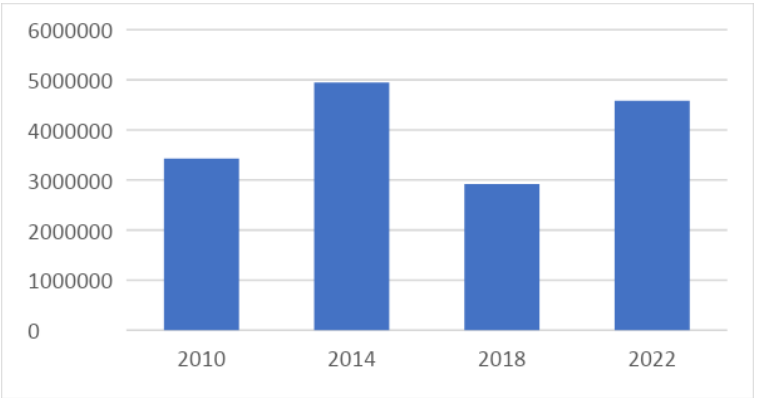


Table 1 breaks down the numbers in Figure 1 to present the dominant file types. As observed in Figure 1, commonly used files on websites included HTML, images, PDF files, and the category "other". The years 2014 and 2018 showed a higher prevalence of images and a lower count of HTML files compared to 2010 and 2022. The category "Other" included types of files such as .js (JavaScript) and .css (Cascading Style Sheets).

Table 1. File types and numbers found in the corpus for the four years.

File types	2010 (% of 2010)	2014 (% of 2014)	2018 (% of 2018)	2022 (% of 2022)
HTML	2,871,108 (83.7)	3,524,990 (71.3)	2,100,565 (72.0)	3,656,309 (79.8)
Image	338,698 (9.9)	502,156 (10.2)	360,039 (12.3)	250,578 (5.5)
Other	94,834 (2.8)	729,503 (14.7)	156,954 (5.4)	396,816 (8.7)
PDF	85,479 (2.5)	85,347 (1.7)	267,671 (9.2)	183,615 (4.0)
Text	38,497 (1.1)	104,747 (2.1)	30,447 (1.0)	93,355 (2.0)
Word	407 (0.0)	24 (0.0)	35 (0.0)	2 (0.0)
Video	346 (0.0)	323 (0.0)	305 (0.0)	344 (0.0)
Audio	270 (0.0)	211 (0.0)	1761	65 (0.0)
Excel	5 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Total	3,429,644	4,947,301	2,917,777	4,581,084

We also conducted an analysis of the distribution of domain types in the corpus. Municipalities and regions were represented by their domain names in the ETL descriptions, with the expected number known in advance—that is, 98 municipalities and five regions. However, a comparison with Table 2 reveals some blind spots in the corpus. Thus, only 44 municipalities were captured in 2022, although the other years showed better representation. Regional representation was comparable to this, with only one or two regions captured, as opposed to the expected five.

As mentioned, the selective crawl on ministries and government agencies included not only the anticipated entities but also related institutions and websites. To specifically isolate ministries and government agencies, we classified the domains into three categories—namely, 1) ministries and government agencies, 2) state institutions, and 3) others. The first category included entities such as the Ministry of Employment and the Danish Authority of Social Services and Housing. The second category was included in the corpus to identify institutions operating at the state level but not in the form of an agency or ministry. Among these was borger.dk, the national portal providing citizens' access to information across local, regional, and state levels, offering information from hospitals, local municipalities, housing and tax authorities, the public message system (e-boks.dk), and more. Other examples of state institutions in this category included the State Archives, the Accident Investigation Board of Denmark, and the Geological Survey of Denmark and Greenland. The category "Others" encompassed a variety of web domains. Some were directly associated with state institutions, such as sites providing information on study support in Scandinavian countries and the unit governing traffic on the bridge between Funen and Zealand. Moreover, educational institutions, such as universities and university colleges, were placed in the "Other" category. Lastly, websites loosely linked to government administrations, such as the campaign site for Tour de France in Denmark, a national lottery website, and the Union of Municipalities, were also included.

With these definitions in place, Table 2 shows that 2010 and 2022 had the highest occurrence of ministries, government institutions, and state institutions, with 2010 harvesting a significantly higher number of websites in the "Other" category. While Table 1 reveals that 2014 had the largest number of files, in Table 2, the year is shown to be ranked second lowest in terms of domains. This discrepancy suggests that there was a more extensive harvest of domains in 2014 compared to the other three years. Further exploration of this will be conducted in subsequent analyses.

Table 2. Domains retrieved from the corpus for the four years.

	2010 (% of 2010)	2014 (% of 2014)	2018 (% of 2018)	2022 (% of 2022)
Municipalities	89 (16.3)	72 (30.8)	89 (38.7)	44 (13.5)
Regions	2 (0.4)	1 (0.4)	1 (0.4)	1 (0.3)
Ministries and government agencies	59 (10.8)	50 (21.4)	46 (20.0)	69 (21.2)
State institutions	53 (9.7)	35 (15.0)	32 (13.9)	54 (16.6)
Others	343 (62.8)	76 (32.5)	62 (27.0)	157 (48.3)
Total	546	234	230	325

5. Discussion and next steps

This chapter details the construction of a corpus for studying the online public sector in Denmark, of which a preliminary analysis was conducted. As demonstrated both in the literature and through empirical work, archived web data does not necessarily capture the entire web when it is crawled for national archives. This inherent limitation should be considered when analyzing the data in future work. Nevertheless, the corpus presented in this chapter contributes valuable insights into e-government based on archived web materials.

Future work will extend the analyses of the data and include a more in-depth examination of the text and links within the harvested websites. In addition, to detect potential variances between the different types of public administrations, we will also analyze how the local and national levels of e-government have evolved over time. The categorizations of e-government websites (Fietkiewicz, Mainka, and Stock 2017; Fan 2018) can serve as the theoretical basis for these comparisons. Several studies have investigated the role of background measures, such as outside pressure, size, and regulatory quality (García-Lacalle, Pina, and Royo 2011; Garcia-Murillo 2013; Epstein 2022; Ingrams et al. 2020), in the development of e-government. Future work could incorporate similar background measures when analyzing municipalities.

6. Acknowledgments

This work was supported by Netlab at Aarhus University and the Aalborg University SSH Dean's Strategic Fund. The authors wish to express their appreciation for the valuable technical and analytical support

provided by Ulrich Karstoft Have (AU), Matias Kokholm Appel (AAU), and Signe Birgit Sørensen (AAU).

References

Brügger, Niels, Janne Nielsen, and Ditte Laursen. 2020. "Big Data Experiments with the Archived Web: Methodological Reflections on Studying the Development of a Nation's Web." *First Monday*, February. https://doi.org/10.5210/fm.v25i3.10384.

Cahlikova, Tereza. 2021. "Drivers of and Barriers to E-Government." In *The Introduction of E-Government in Switzerland: Many Sparks, No Fire*, edited by Tereza Cahlikova, 45–68. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-78624-3_3.

Chatzopoulou, Sevasti, and Birgitte Poulsen. 2017. "Combining Centralization and Decentralization in Danish Public Administration." In *The Palgrave Handbook of Decentralisation in Europe*, edited by J. Ruano and M. Profiroiu. London: Palgrave Macmillan. https://doi-org.zorac.aub.aau.dk/10.1007/978-3-319-32437-1_11.

Donig, Simon, Markus Eckl, Sebastian Gassner, and Malte Rehbein. 2023. "Web Archive Analytics: Blind Spots and Silences in Distant Readings of the Archived Web." *Digital Scholarship in the Humanities* 38, no. 3: 1033–48. https://doi.org/10.1093/llc/fqad014.

Epstein, Ben. 2022. "Two Decades of e-Government Diffusion among Local Governments in the United States." *Government Information Quarterly* 39, no. 2: 101665. https://doi.org/10.1016/j.giq.2021.101665.

Fage-Butler, Antoinette, Loni Ledderer, and Niels Brügger. 2022. "Proposing Methods to Explore the Evolution of the Term 'mHealth' on the Danish Web Archive." *First Monday*, January. https://doi.org/10.5210/fm.v27i1.11675.

Fan, Qiuyan. 2018. "A Longitudinal Evaluation of e-Government at the Local Level in Greater Western Sydney (GWS) Australia." *International Journal of Public Administration* 41, no. 1: 13–21. https://doi.org/10.1080/01900692.2016.1242621.

Feeney, Mary K., and Adrian Brown. 2017. "Are Small Cities Online? Content, Ranking, and Variation of U.S. Municipal Websites." *Government Information Quarterly*, Open Innovation in the Public Sector 34, no. 1: 62–74. https://doi.org/10.1016/j.giq.2016.10.005.

Fietkiewicz, Kaja J., Agnes Mainka, and Wolfgang G. Stock. 2017. "eGovernment in Cities of the Knowledge Society. An Empirical Investigation of Smart Cities' Governmental Websites." *Government Information Quarterly, Open Innovation in the Public Sector* 34, no. 1: 75–83. https://doi.org/10.1016/j.giq.2016.08.003.

Flensburg, Sofie, and Signe Sophus Lai. 2021. "Networks of Power. Analysing the Evolution of the Danish Internet Infrastructure." *Internet Histories* 5, no. 2: 79–100. https://doi.org/10.1080/24701475.2020.1759010.

García-Lacalle, Javier, Vicente Pina, and Sonia Royo. 2011. "The Unpromising Quality and Evolution of Spanish Public Hospital Web Sites." *Online Information Review* 35, no. 1: 86–112. https://doi.org/10.1108/14684521111113605.

Garcia-Murillo, Martha. 2013. "Does a Government Web Presence Reduce Perceptions of Corruption?" *Information Technology for Development* 19, no. 2: 151–75. https://doi.org/10.1080/02681102.2012.751574.

Gjerding, Allan N. 2005. "The Danish Structural Reform of Government." Aalborg University. https://vbn.aau.dk/ws/portalfiles/portal/166299/abstractfil.pdf.

Grönlund, Åke, and Thomas A. Horan. 2005. "Introducing e-Gov: History, Definitions, and Issues." *Communications of the Association for Information Systems* 15, no. 1. https://doi.org/10.17705/1CAIS.01539.

Hollingsed, Tasha, and David G. Novick. 2007. "Usability Inspection Methods after 15 Years of Research and Practice". In *Proceedings of the 25th Annual ACM International Conference on Design of Communication*, 249–55. SIGDOC '07. New York, NY: ACM. https://doi.org/10.1145/1297144.1297200.

Hujran, Omar, Ayman Alarabiat, Ahmad Samed Al-Adwan, and Mutaz Al-Debei. 2023. "Digitally Transforming Electronic Governments into Smart Governments: SMARTGOV, an Extended Maturity Model." *Information Development* 39, no. 4: 811–34. https://doi.org/10.1177/02666669211054188.

Ingrams, Alex, Aroon Manoharan, Lisa Schmidthuber, and Marc Holzer. 2020. "Stages and Determinants of e-Government Development: A Twelve-Year Longitudinal Study of Global Cities." *International Public Management Journal* 23, no. 6: 731–69. https://doi.org/10.1080/10967494.2018.1467987.

Layne, Karen, and Jungwoo Lee. 2001. "Developing Fully Functional e-Government: A Four Stage Model." *Government Information Quarterly* 18, no. 2: 122–36. https://doi.org/10.1016/S0740-624X(01)00066-1.

Lazar, Jonathan, Brian Wentz, Abdulelah Almalhem, Alexander Catinella, Catalin Antonescu, Yeveniy Aynbinder, Michael Bands, et al. 2013. "A Longitudinal Study of State Government Homepage Accessibility in Maryland and the Role of Web Page Templates for Improving Accessibility." *Government Information Quarterly* 30, no. 3: 289–99. https://doi.org/10.1016/j.giq.2013.03.003.

Lemke, Florian, Kuldar Taveter, Regina Erlenheim, Ingrid Pappel, Dirk Draheim, and Marijn Janssen. 2020. "Stage Models for Moving from e-Government to Smart Government". In *Electronic Governance and Open Society: Challenges in Eurasia*, edited by Andrei Chugunov, Igor Khodachek, Yuri Misnikov, and Dmitrii Trutnev, 152–64. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-39296-3_12.

Maemura, Emily. 2018. "What's Cached Is Prologue: Reviewing Recent Web Archives Research towards Supporting Scholarly Use." *Proceedings of the Association for Information Science and Technology* 55, no. 1: 327–36. https://doi.org/10.1002/pra2.2018.14505501036.

Maemura, Emily. 2023. "All WARC and No Playback: The Materialities of Data-Centered Web Archives Research." *Big Data & Society* 10, no. 1: 20539517231163172. https://doi.org/10.1177/20539517231163172.

Murakami, Akira, Paul Thompson, Susan Hunston, and Dominik Vajn. 2017. "'What Is This Corpus About?': Using Topic Modelling to Explore a Specialised Corpus." *Corpora* 12, no. 2: 243–77. https://doi.org/10.3366/cor.2017.0118.

Nielsen, Jakob. 1994. "Usability Inspection Methods." In *Conference Companion on Human Factors in Computing Systems – CHI '94*, 413–14. Boston: ACM Press. https://doi.org/10.1145/259963.260531.

Nielsen, Janne. 2021. "Quantitative Approaches to the Danish Web Archive." In *The Past*

Web, edited by D. Gomes, 165–79. Berlin: Springer.
Pedersen, Keld. 2018. "e-Government Transformations: Challenges and Strategies." *Transforming Government: People, Process and Policy* 12, no. 1: 84–109. https://doi.org/10.1108/TG-06-2017-0028.
Raffal, Harry. 2018. "Tracing the Online Development of the Ministry of Defence and Armed Forces through the UK Web Archive." *Internet Histories* 2, no. 1–2: 156–78. https://doi.org/10.1080/24701475.2018.1456739.
Ronchi, Alfredo M. 2019. "e-Government: Background, Today's Implementation and Future Trends." In *E-Democracy: Toward a New Model of (Inter)Active Society*, edited by Alfredo M. Ronchi, 93–196. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-01596-1_5.
Rooks, Gerrit, Uwe Matzat, and Bert Sadowski. 2017. "An Empirical Test of Stage Models of e-Government Development: Evidence from Dutch Municipalities." *The Information Society* 33, no. 4: 215–25. https://doi.org/10.1080/01972243.2017.1318194.
Sandoval-Almazan, Rodrigo, and J. Ramon Gil-Garcia. 2012. "Are Government Internet Portals Evolving towards More Interaction, Participation, and Collaboration? Revisiting the Rhetoric of e-Government among Municipalities." *Government Information Quarterly* 29, no. 1: S72–81. https://doi.org/10.1016/j.giq.2011.09.004.
Schafer, Valérie. 2017. "From Far Away to a Click Away: The French State and Public Services in the 1990s." In *The Web as History: Using Web Archives to Understand the past and the Present*, edited by Niels Brügger and Ralph Schroeder, 117–36. London: UCL Press. https://discovery.ucl.ac.uk/id/eprint/1542998/1/The-Web-as-History.pdf.
Schostag, Sabine, and Eva Fønss-Jørgensen. 2012. "Webarchiving: Legal Deposit of Internet in Denmark. A Curatorial Perspective." *Microform & Digitization Review* 41, no. 3–4. https://doi.org/10.1515/mir-2012-0018.
Shi, Yuquan. 2006. "e-Government Web Site Accessibility in Australia and China." *Social Science Computer Review* 24, no. 3: 378–85. https://doi.org/10.1177/0894439305283707.
Skiftenes Flak, Leif, Willy Dertz, Arild Jansen, John Krogstie, Ingrid Spjelkavik, and Svein Ølnes. 2009. "What Is the Value of eGovernment – and How Can We Actually Realize It?" *Transforming Government: People, Process and Policy* 3, no. 3: 220–26. https://doi.org/10.1108/17506160910979333.
United Nations. 2023. "Compare Countries." https://publicadministration.un.org/egovkb/en-us/Data/Compare-Countries.
Vrangbæk, Karsten. 2010. "Structural Reform in Denmark, 2007–09: Central Reform Processes in a Decentralised Environment." *Local Government Studies*, April. https://doi.org/10.1080/03003930903560562.
Zhang, Hui, Xiaolin Xu, and Jianying Xiao. 2014. "Diffusion of e-Government: A Literature Review and Directions for Future Directions." *Government Information Quarterly* 31, no. 4: 631–36. https://doi.org/10.1016/j.giq.2013.10.013.

SECTION 5
Multi-level methods for studying web archives

Web archives and hyperlink analyses: The case of videnskab.dk 2009–2022

Niels Brügger, Katharina Sølling Dahlman

Abstract: This chapter demonstrates how the use of a national web archive in hyperlinked network analyses may prove an indispensable source when conducting not only historical but also contemporary analyses of a given website. Our analyses are based on the case of videnskab.dk, a Danish journalistic website disseminating research-related knowledge to the public. Focus is on the examination of hyperlinks related to videnskab.dk in the years of 2009, 2014, 2018, and 2022, followed by a network analysis of videnskab.dk in relation to similar transnational websites. Our results showcase what insights may be gained when conducting analyses with and without access to a national web archive, respectively, highlighting the impact and importance of data collections when studying the online web..

Keywords: web archive, hyperlink network analysis, actor types, historical analysis, contemporary analysis.

1. Introduction

This chapter investigates how the holdings of a national web archive can be used to shed light on the hyperlinks related to one individual website. The study explores the case of the Danish science website videnskab.dk, and it is primarily based on content from the national Danish web archive Netarkivet. Videnskab.dk is a journalistic website that disseminates research-related knowledge to the wider public, like scientificamerican.com in the US, futura-sciences.com in France, and scinexx.de in Germany, and it has been chosen as a case because historical hyperlink network analyses of the website were conducted as part of a larger evaluation project of the many activities of the website (explained in more detail below).

The aim of the following is twofold. Firstly, to provide empirical results about the historical development of the hyperlinks related to the website videnskab.dk, with a focus on the changing main actor types to which it is connected. Secondly, to showcase that the archived web is not only useful for historical studies, but it is also an indispensable source type for contemporary analyses, in particular hyperlink analysis because web archives are (probably) the only place where in-links to any given website can be found, in contrast to out-links that are known to the website owner, and that can be collected from the website itself on the online web. As part of the latter aim a transnational hyperlink network analysis of the online web is included to highlight what could be done if national web archives

Niels Brügger, Aarhus University, Denmark, nb@cc.au.dk, 0000-0003-1787-1980
Katharina Sølling Dahlman, Aarhus University, Denmark, katharina@j-p.dk

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Niels Brügger, Katharina Sølling Dahlman, *Web archives and hyperlink analyses: The case of videnskab.dk 2009–2022*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.19, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 201-222, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

could be combined.

Thus, the overall research question is: What characterizes the changes of actor types in the hyperlink network of videnskab.dk?

1. Context of the study

To better understand the following hyperlink analyses some context is needed, including information about why this study was made, where the data came from, what characterizes network analysis of hyperlinks and the archived web, and finally how the available data were prepared for analysis.

2.1 The starting point: Evaluating videnskab.dk

Videnskab.dk was founded in 2008 to promote and communicate research-related knowledge to the wider public, and in 2023 videnskab.dk had 18 employees, 12 full-time and 6 part-time (Degn et al. 2023, 22). In 2022, after 15 years of the website's existence, the Danish Agency for Higher Education and Science, which provided funding for the website, sought its evaluation. The Centre for Cultural Evaluation at Aarhus University was commissioned to perform this evaluation, and the authors of this chapter were invited to contribute analyses of the hyperlink structure around videnskab.dk.

To cover as many facets of the evaluation of the website's activities as possible, a very broad analytical design was chosen, including (1) an analysis of the website and its content, with a focus on genre, design, functionality, and journalistic communication, (2) an analysis of social media presence and communication strategies (Twitter, Facebook, Instagram, and LinkedIn), (3) a field study and interviews with management and staff, (4) interviews with researchers who have contributed to articles on videnskab.dk, (5) interviews with science journalists/editors from other media who published articles based on content from videnskab.dk, (6) questionnaires and interviews with teachers and pupils (elementary and high school), and the two elements that constitute the basis of this chapter, (7) analyses of hyperlinks extracted from Netarkivet, from the period 2009–2022, and (8) a network analysis of outgoing hyperlinks from international websites of similar type, that is journalistic websites that disseminate research.

The study was conducted in 2022 by researchers with different backgrounds to cover the various approaches, and the final evaluation report was published in March 2023 (Degn et al. 2023). In the following, focus is only on the hyperlink analyses (points (7) and (8) above), and the results that did not find room in the final report (Degn et al. 2023, 32–36) are unfolded in more detail. The report was written in Danish, but a brief

summary in English can be found on page 4 in the report.

2.2 Getting the data: The national Danish web archive Netarkivet

Since 2005, the Danish web has been collected by the national Danish web archive Netarkivet at the Royal Danish Library (see http://netarkivet.dk). Netarkivet collects the entire Danish web domain .dk four times each year, along with a limited amount of Danish web material on other web domains. In recent years, Netarkivet has enabled researchers to extract and obtain content for research purposes. Based on this service, data containing all links to and from videnskab.dk for the years 2009, 2014, 2018, and 2022 was extracted. The raw data files contained between 100,000 and 200,000 links each year: 141,903 in 2009; 233,886 in 2014; 127,234 in 2018; and 113,706 in 2022.

A few limitations that may influence the results have to be addressed. Firstly, in contrast to Netarkivet's collection of the Danish web that is almost complete, social media platforms such as Facebook, Twitter, and YouTube have not been collected in a systematic and exhaustive manner. This implies that links from the web to social media are present, whereas the opposite may not be true. Secondly, the following hyperlink analyses do not place videnskab.dk in the complete link graph of all links on the Danish web, which amounts to 10–12 billion links. Rather, videnskab.dk is positioned within its immediate context, defined as links one iteration away from the website. This includes links from videnskab.dk to other websites, links from other websites to videnskab.dk, and links in and out of all these websites. While this approach makes the analysis more focused, it comes at the expense of completeness (a few examples of studies of the Danish web exist, e.g. Brügger et al. 2017; Brügger et al. 2020).

2.3 Hyperlink (network) analyses and the archived web

The methodological history of network analysis dates back to the 1930s (Moreno 1934) and has been used to study diverse topics (refer to Wasserman and Faust 1994, 5–6, for an extensive list). In the mid-1990s, the advent of the web as a media platform offered new opportunities to study networks, because the web is characterized by concrete connections manifested as hyperlinks. This led to the inception of hyperlink network analysis around 1997, one of the first articles in this new sub-field being Jackson (1997). In the following years, hyperlink network analysis gained prominence in internet studies (see early overviews in Foot et al. 2003, 4–8; Park and Thelwall 2003) and within the software industry, with Google's PageRank playing a pivotal role (Brin and Page 1998). By the early 2010s, the widespread availability of web archives gave rise to a new branch of

hyperlink network analysis: hyperlink network analysis of the archived web. Weltevrede and Helmond's historical study of the Dutch blogosphere stands as one of the first examples, based on the holdings of the Internet Archive (Weltevrede and Helmond 2012). Shortly thereafter, discussions on how the specificities of the archived web as a source affect network analysis are added to this literature (Brügger 2013). However, as of today, the number of network analyses studying the archived web remains limited, predominantly adopting a historical perspective (e.g. Meyer et al. 2017; Weber 2017; Cowls and Bright 2017; Ackland and Evans 2017; Webster 2017; Brügger 2021, 2022; Fage-Butler et al. 2022), whereas the archived web is not studied as a source that can shed new light on contemporary hyperlink networks (for a brief introduction to network analysis, hyperlink network analysis, and the archived web, see Stevenson and Ben-David 2018). This chapter aims to bridge this gap by studying both the contemporary web and the past web with the archived web as a source.

The network analysis of videnskab.dk's hyperlinks is based on standard network analytical concepts (Wasserman and Faust 1994), where the value of an entity is a function of its relations to other entities in the network. The nodes of the network are entire websites (and not individual web pages), while a hyperlink constitutes the edge, and the number of concrete hyperlinks between two nodes determines the weight of the edge. In addition, as outlined below, websites are categorized into actor types, which then serve as an attribute of the node. Since hyperlinks point from one website to another website, the network is directed. The analysis focuses on three ways of measuring centrality: in-degree centrality (the number of edges pointing to a given website), out-degree centrality (the number of edges pointing from a given website to other websites), and betweenness centrality (how often a node is present on the shortest path between two nodes, in other words how often it functions as a bridge). It is important to note that a website can control its out-degree, but not its in-degree or its betweenness centrality.

However, when using the archived web as the source for hyperlink (network) analyses, two limitations related to the nature of the archived web must be considered, in contrast to conducting hyperlink network analyses of the online web (Brügger 2013—for a general introduction to the archived web as a source, see Brügger 2019). Firstly, due to the method of web archiving and the organization of the collection, the same web page is likely to exist in the archive more than once, even within a limited period of time. In some cases, it may be an identical copy, while in others, it may be a version, that is two web pages with the same URL but different content from different points in time. Therefore, versions of the same web page from two different points in time are excluded if they link to precisely the

same websites, even though their content may differ, thus retaining only one version of each web page in the dataset. This curation approach aims to reflect what the web actually looked like in the past (and not what it looked like in the web archive) while reducing the number of links considerably.

Secondly, since Netarkivet collects the entire .dk web domain four times a year, material meeting the criteria for this analysis (in- and out-going links from videnskab.dk + one iteration) can be archived at different times during a calendar year. Consequently, the analysis of each of the four years has a temporal extension of one year, wherein links that were not simultaneously present online are analyzed as if they were. In other words, each annual link graph becomes temporally inconsistent. This inconsistency is accepted because it provides a more comprehensive link graph compared to a link graph based on only one week or one month per year (for a discussion on the balancing of temporal inconsistency and completeness, see Brügger 2019, 22–25).

2.4 Preparing the data for analysis

The csv files extracted from Netarkivet were processed in Excel to prepare them for analysis using the network analysis software Gephi (gephi.org) and to perform certain descriptive statistics. To simplify the network, a cut-off level of 100 was applied to the edge weight, that is: edges connecting two websites with fewer than 100 links were excluded from the dataset. Upon initial test analyses, it became clear that further data cleaning was necessary, and additional information needed to be incorporated.

Initially, the dataset contained nodes with very high weights, raising questions about whether the numerous links were actual links to videnskab.dk (and other websites) or if they were the result of recurring website construction elements, (menu, navigation, footer, and the like). These components could lead to a high weight even if all links were, strictly speaking, identical. To remove these types of links, the edge table was manually checked for suspicious edges, including the following:

Table 1: Suspicious rows in the edge table.

source	target	weight
ronniandersson.dk	facebook.com	8305
ronniandersson.dk	google.com	8305
ronniandersson.dk	twitter.com	8305
ronniandersson.dk	upworth.dk	8305

Table 1 illustrates instances where five different edges from the same node display an identical number of links. This clearly indicates that the links were found in a footer or a similar element present on all pages of the

websites. Manual checks were conducted on such cases by examining the website in Netarkivet's browser view. When evaluating the link types, inspiration was drawn from the categorization proposed by Ryfe et al. (2016), which distinguishes four types of links:

> [...] navigation, commercial, social, and citation. Navigation helps users find content. Commercial involves linking practices for earning money from others. Social includes sharing content via social media feeds and/or offering users opportunities to share content. Citation directs users to information in an effort to establish the credibility of news reports. (Ryfe et al. 2016, 42)

Since our analysis primarily focuses on 'citation' links related to videnskab.dk, web features such as share buttons to social media and similar elements are not considered links and are thus excluded from the dataset.

Secondly, the dataset had to be enriched with information on the actor types within videnskab.dk's link graph, since this could not be immediately deduced from the web addresses. Therefore, the top 50 websites (measured by weight) in the edge tables for in- and out-links were manually checked either in Netarkivet or on the online web, to determine their respective categories. The list of categories was developed through an iterative process, establishing a new category if a website on the list did not fit one of the already identified categories. To reduce complexity, websites were assigned to a single category only. This grounded approach led to the following list of categories:

- Science website (e.g. sciencenorway.no)
- Mainstream media (such as national daily newspapers and weekly magazines)
- Niche media (niche media *with* a journalistic/editorial approach)
- Alternative media (niche media *without* a journalistic/editorial approach)
- Research institution
- Library
- Education
- Association
- Official (e.g. Ministries, Health Care system)
- Publisher (academic publisher, either publishing house or journal)
- Academic portal (e.g. researchgate.net)
- Blog
- Encyclopedia (e.g. wikipedia.org)
- Discussion forum
- Other

3. Results: videnskab.dk in the hyperlink network 2022, and actor types in the past

The primary focus of the analysis of videnskab.dk centers on the website's link graph as it appeared when the evaluation report was drafted, i.e. in the year 2022. The analysis comprises (1) descriptive statistics based on the number of links to and from videnskab.dk, with a particular focus on actor types, and (2) a network analysis that includes videnskab.dk's position in the network. The difference between these two approaches lies in the fact that the statistical analysis gives an overview of actor types and individual websites but does not provide insights into the centrality of each website in the network—whether the websites connected to videnskab.dk are themselves central or not. This dimension is elucidated through the network analysis. Looking back from 2022, the study examines the website's hyperlinks in 2009, 2014, and 2018, followed by a brief outline of some of the major developments. This historical dimension specifically focuses on the actor types of the hyperlink and not on the network as such, aligning with the analysis presented in the published evaluation report. Finally, a brief analysis of videnskab.dk in the transnational web landscape is included. An Appendix with the figures that are not included in the following can be found in the Zenodo community 'Book chapter Web archives and hyperlink analyses' at https://zenodo.org/communities/resaw2023_chapter.

3.1 The hyperlink network 2022

The first analytical step is to examine the top50 actor types linked either through out-links or in-links from videnskab.dk. Most links from videnskab.dk point to either academic publishers or research institutions, comprising nearly half of the top-50. The remainder is primarily linked to mainstream media and other reputable scientific websites (Figure 1 in appendix). Libraries and research institutions dominate the in-linkers, constituting just below half of top-50. Blogs, discussion fora, and alternative media also contribute to in-links as much as mainstream media (Figure 2 in appendix).

Comparing linked-to and in-linking actor types, research institutions are prevalent in both cases, while mainstream media also hold significance, albeit to a lesser extent. Unsurprisingly, videnskab.dk does not link to more (scientifically) dubious websites from alternative media, but these websites link back to videnskab.dk.

The second analytical step examines the distribution of individual actors within the top50 concerning the number of concrete links (Figure 3 and 4 in appendix). Regardless of whether the focus is on out- or in-links to videnskab.dk, the structure remains the same: very few actors possess a very

high number of links, followed by a mid-group of approximately 10 actors with fewer links, and then a long tail of actors with very few links.

A closer look at in-linking actors reveals a number of characteristics: (1) the highest in-linking website is a Nordic science-related website (sciencenordic.com), similar to videnskab.dk and with which videnskab.dk collaborates; (2) many research institutions among the top in-linkers have a substantial number of concrete links (high edge weight); (3) among the top10 in-linkers, niche media like ing.dk (a journalistic website on technology and science) and two alternative media—a news aggregator (godt-nyt.dk) and a website about drug use (psychedelia.dk)—are noteworthy; (4) the rest of top50 includes a mix of mainstream media, two alternative media (uriasposten.net—an anti-elite website—and nomedica.dk—an anti-medical science website),personal blogs (e.g. lykkeliginoerdland.dk, a blog aimed at informing women about 'hard science'), and discussion fora such as ingeniordebat.dk (an engineering forum with 647 members) and musclezone.dk (a bodybuilding forum); (5) finally, it is worth noting that public libraries link to videnskab.dk, but generally with very few concrete links.

Moving on to the actual network analysis, our focus is on how videnskab.dk is positioned within its immediate hyperlink network and identifying the characteristics of other nodes in this network. A few network statistics: edges are only included if they have a weight above 100 (as previously mentioned); the network comprises 1,147 nodes and 1,679 edges; the network diameter is 4, indicating that 4 steps are needed to travel between the two farthest nodes; the average degree of nodes is 1.164, indicating that each node is connected to a little more than one other node; the average weighted degree is 371.983, representing the average number of edges weighted with the weight of each edge; and the graph density is 0.001, measured on a scale between 1 and 0, where 1 implies that all potential edges are realized, and 0 signifies none.

Figure 1 illustrates the network with a focus on out-degree that is the number of outgoing hyperlinks. Unsurprisingly, many of the actor types and specific actors already identified in the statistics of top-50 most linked to from videnskab.dk are visible, but new actors also emerge as central out-linking nodes, notably psyx.blogspot.com, a blog for a psychotherapist and sexologist. Also, it is worth noting that two major out-linkers are alternative media, 23.dk (likely due to its Wikipedia structure) and uriasposten.net, known for being a link central. Blogs, given their inherent nature, are also central out-linkers (ing.dk, version2.dk, minkusinemaria.dk, lykkeliginoerdland.dk). Finally, it is worth noting that mainstream media and libraries do not play a substantial role in the out-degree network contrasting with their prominence when focusing solely on links in

and out from videnskab.dk.

Figure 1: The near out-degree network of videnskab.dk. Nodes are sized according to their out-degree, edges according to weight (graph spatializsed with Fruchterman Reingold (area 5.000, gravity 1.,0), zoomed for better readability. For the full network visualizsation, see Figure 5 in appendix.

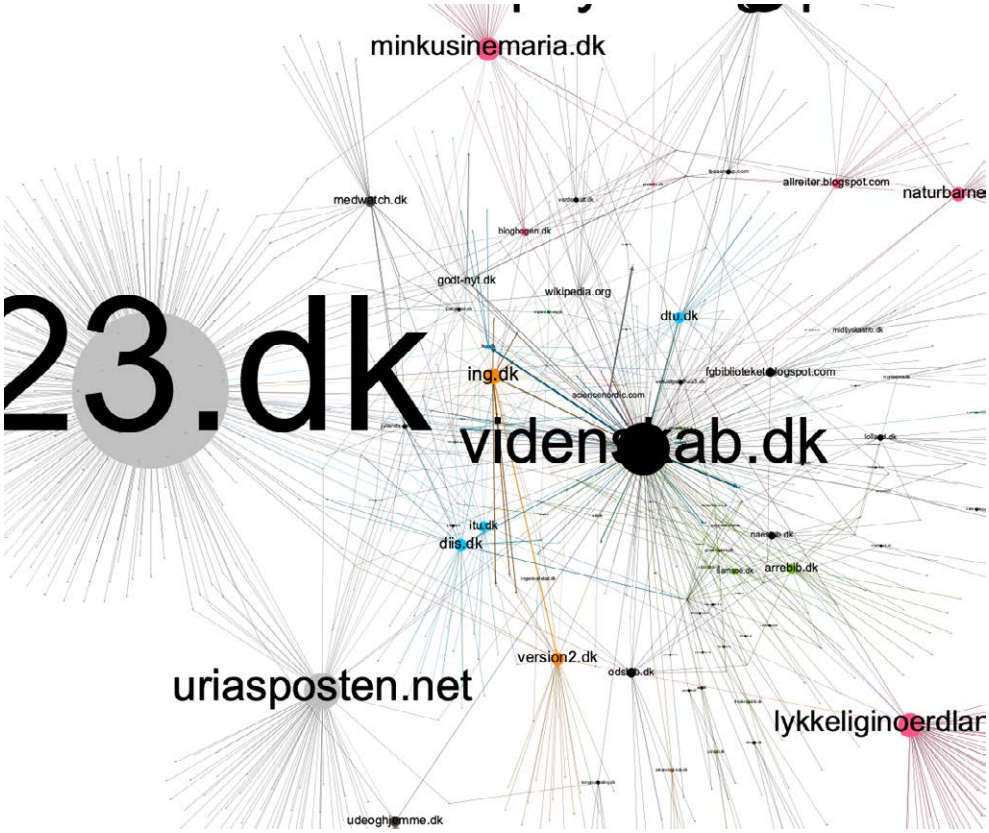


Figure 2: The near in-degree network of videnskab.dk. Nodes are sized according to their in-degree, edges according to weight (graph spatializsed with Fruchterman Reingold (area 5.000, gravity 1.,0), zoomed for better readability. For the full network visualizsation, see Figure 6 in appendix.

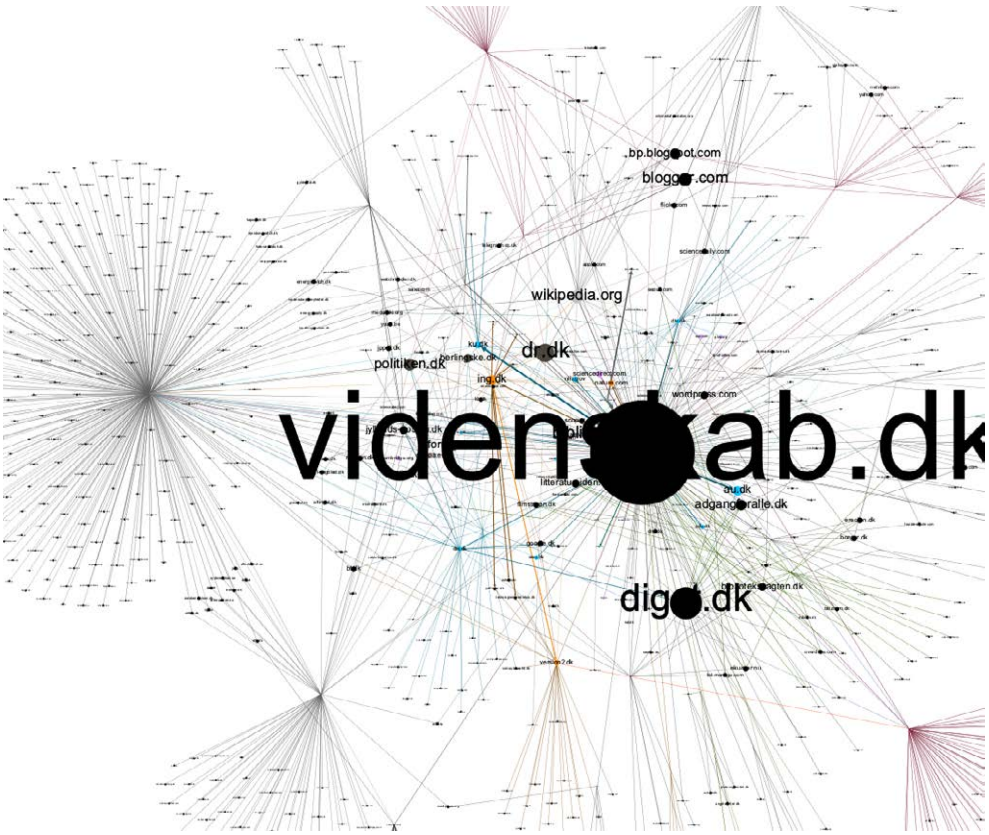
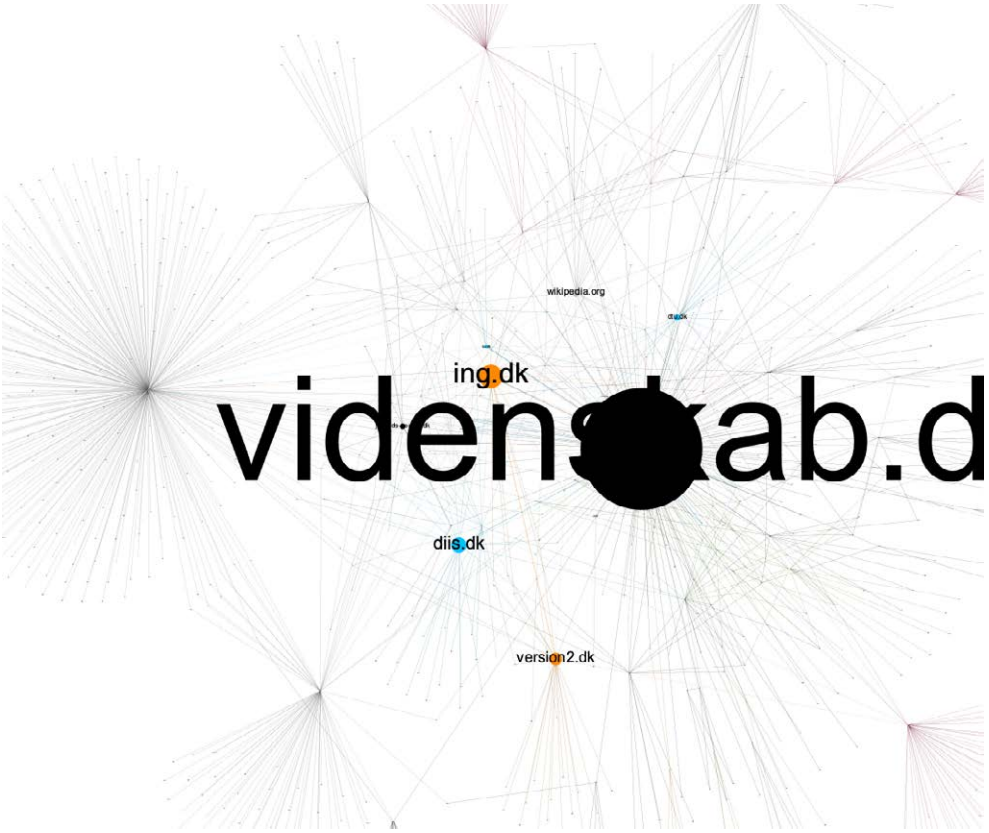


Figure 2 visualizes the network with a focus on in-degree that is the number of ingoing hyperlinks. When analyzing the in-degree network, it is important to bear in mind that any given website cannot, by itself, produce a high in-degree, since it does not decide who links to it. The in-degree thus provides insights into how popular a website is if a link is considered a popularity measure. Actors from the top-50 list of links related to videnskab.dk are also present in the in-degree network, with mainstream media and research institutions being central. New websites appear as central in-linked nodes, such as digst.dk (the Agency for Digital Government), labradorcms.com (a website for digital publishing), and blogger.com. Notably absent in the in-degree network are libraries. Finally, a stark difference observed when comparing the in-degree network to the out-degree network in Figure 1 is that nodes central in the out-degree network are not central in the in-degree network, except for ing.dk. This

suggests that the heavily out-linking websites may not be popular for incoming links, making them relevant only in the out-degree network if directly accessed (by typing their web address in the browser) because they are not likely to be visited by users who arrive at them through an in-link.

Figure 3: The near betweenness-centrality network of videnskab.dk. Nodes are sized according to their betweenness centrality (the bigger the node, the more it functions as a bridge in the network), edges according to weight (graph spatializsed with Fruchterman Reingold (area 5.000, gravity 1.,0), zoomed for better readability. For the full network visualizsation see Figure 7 in appendix.



The third segment of the network analysis focuses on betweenness centrality, identifying websites that play a central role as bridges between other nodes (Figure 3). Unsurprisingly, videnskab.dk emerges as the most central bridge, and apart from a handful of nodes (ing.dk, version2.dk, the research institutions diis.dk, dtu.dk, and ku.dk, and one mainstream media), there are minimal important bridges. This means that actor types such as publishers, libraries, mainstream media, alternative media, and blogs, do not serve as bridges enabling connections between the different actor types.

When comparing the out-degree, in-degree, and betweenness networks, it becomes evident that different actors take central roles depending on the network focus. Only a few actors are central in more than one network, notably ing.dk and version2.dk, along with a few research institutions, most notably ku.dk (the University of Copenhagen). Surprisingly, libraries are not central in any of the networks.

Determining the role of videnskab.dk in the link network in relation to other actors highlights the significance of those in top 50 in-linkers to videnskab.dk, that is the actors who deliberately point to videnskab.dk and potentially send their own users in that direction. However, their value for videnskab.dk grows with their centrality in the network. In other words: it is interesting when an actor links to videnskab.dk, but it is even more interesting if the linking node holds a central position in the network. To investigate this, one needs to consider actors in the top 50 of in-linkers with these nodes' centrality within each of the three network measures: out-degree, in-degree, and betweenness centrality. If a website not only links significantly to videnskab.dk (among the top 50 in-linkers) but also ranks high in one or more of these measures, it signifies that the website is particularly important for videnskab.dk. An analysis along these lines reveals four websites as the most crucial: ing.dk (the journalistic website on technology and science), diis.dk, dtu.dk (websites from two research institutions), and wikipedia.org. Others are also important, but to a lesser extent: e23.dk, lykkeliginoerdland.dk, minkusinemaria.dk, naturbarnet.dk (a blog about healthy living), version2.dk, udeoghjemme.dk (a weekly magazine), information.dk (a mainstream media), ku.dk, au.dk (research institutions). Notably, some strong in-linkers to videnskab.dk (and of whom there were many) do not play a significant role in the broader network: science websites, libraries, and discussion fora. Moreover, the absence of expected actor types that would either link to videnskab.dk or be part of the network is noteworthy, including local newspapers, NGOs, companies, and primary and high schools, which are key target groups for videnskab.dk.

In conclusion, the linking patterns in 2022 suggest that videnskab.dk supports its own ethos as a serious scientific publisher by linking to academic publishers and research institutions. However, the actors linking to videnskab.dk form a much more heterogeneous group: research institutions are still important players, but they are supplemented to various degrees by niche and mainstream media with an interest in videnskab.dk's topics, along with alternative media, blogs, and discussion fora. Thus, videnskab.dk is embedded in networks where the dissemination of scientific knowledge is key, but it also interacts with actors promoting views aligned with its ethos.

3.2 Hyperlinked actor types 2009, 2014, 2018

This section delves into videnskab.dk's out-links and in-links in 2009, 2014, and 2018, providing an overview of actors appearing across these years, followed by a historical analysis comparing these results with those from 2022 to identify and discuss historical developments.

3.2.1 2009: Unreciprocated attention

Looking at videnskab.dk's out-links in the year 2009 (Figure 8 in appendix), research institutions emerge as the predominant actor type, followed by mainstream media. Positioned in the middle are actors such as scientific websites and alternative media, while education, officials, blogs, and discussion forums occupy the lower positions. This suggests a consistent presence of both 'scientific' and 'non-scientific' sources throughout (with reference to differences regarding scientific engagement or association). Although the distribution of concrete links from these actors may exhibit some unevenness (Figure 10 in appendix), this use of both scientific and non-scientific sources appears prevalent here as well, since both feature prominently at either end of the distribution spectrum, as exemplified by the presence of both research institutions and alternative media among those with the highest link count.

Turning to in-links, blogs take the lead as the most prevalent type of actor, appearing more than twice as much as any of the other actor types (Figure 9 in appendix). Research institutions, which dominate out-links, shift to the bottom of the in-links, establishing a contrast between videnskab.dk's out-links and in-links. Furthermore, compared to the above, the consistent presence of scientifically and not scientifically engaged actors is small in both the distribution of actors and the distribution of concrete links (Figure 11 in appendix). Both are largely constituted by actors who may be considered less scientifically engaged, such as discussion forums, mainstream media, and, notably, blogs.

3.2.2 2014: Closing in

In the year 2014, research institutions continue to dominate as the most prevalent actor in videnskab.dk's out-links (Figure 12 in appendix). Mainstream media, however, has been surpassed by publishers, who were not present in the out-links from 2009. The appearance of yet another new actor, academic portals, accompanies the disappearance of discussion forums, blogs, and alternative media. Additionally, the size of education has doubled. The presence of actors more scientifically engaged now outweighs those who may be considered less so. This shift is also apparent in the

distribution of concrete links (Figure 14 in appendix), where research institutions and other scientific actors largely constitute the head of the distribution.

Examining the in-links (Figure 13 in appendix), blogs remain the most prevalent actor, but their size is no longer more than twice that of every other present actor. Mainstream media, niche media, and research institutions almost reach the same level of prevalence as blogs, somewhat evening out the top. Furthermore, several new types of actors appear, namely scientific websites, encyclopedias, libraries, and academic portals. Although these are positioned throughout the middle and the bottom, the contrast between videnskab.dk's out-links and in-links appears less pronounced compared to the figures from 2009. The presence of scientific actors has increased and expanded, as emphasized by the distribution of concrete links (Figure 15 in appendix), with research institutions holding some of the highest link counts.

3.2.3 2018: One step forward, two steps back

Largely the same actors present in 2014 are also present in the out-links from 2018, with research institutions now being preceded by publishers. Scientific websites have almost doubled in size, ranking as the fourth most prevalent actor (Figure 16 in appendix). While still partially uneven, the distribution of concrete links appears somewhat more flattened with an elongated tail (Figure 18 in appendix), possibly indicating a sharpened preference for certain types of sources, such as education and research institutions, which have the highest link counts.

Videnskab.dk's in-links in 2018 somewhat mirror the out-links from 2009, taking two steps back, as blogs have once again grown to twice the size of any other present actor, maintaining their position as the most prevalent actor (Figure 17 in appendix). While the appearance of research institutions remains largely the same as in 2014, this actor is now preceded by alternative media. Furthermore, there is an increase in discussion forums, while actors such as encyclopedias and libraries have disappeared. In other words, the contrast between videnskab.dk's out-links and in-links appears more significant when compared to 2014, with both the prevalence and presence of more scientifically engaged actors diminished. While research institutions still hold some of the highest link counts in the distribution of concrete links (Figure 19 in appendix), they are now accompanied by actors such as blogs and alternative media, emphasizing the aforementioned changes.

3.3 The development of actor types related to videnskab.dk

When comparing videnskab.dk's out-links and in-links over the years, a contrast emerges between who videnskab.dk links to, and who links to videnskab.dk. Videnskab.dk's out-links increasingly target more scientific actors, such as research institutions, publishers, education, and similar entities. On the other hand, actors who are not scientifically engaged, namely blogs, alternative media, and discussion forums, continue to appear as in-linkers to videnskab.dk—actors to whom videnskab.dk, apart from the year 2009, does not link (Figure 8 in appendix; Figure 12 in appendix; Figure 16 in appendix; Figure 1 in appendix).

However, the contrast appears to diminish, as blogs, which were the most prevalent actor in videnskab.dk's in-links throughout 2009, 2014, and 2018 (Figure 9 in appendix; Figure 13 in appendix; Figure 17 in appendix), are finally preceded by libraries and research institutions in 2022 (Figure 2 in appendix). Furthermore, there is a continuous increase in link counts from scientific actors. Building on Terveen and Hills' understanding of a website's hyperlink connectivity as a reflection of the website's credibility and perceived quality, described as a positive correlation (Park and Thelwall 2003, 13), these results may suggest a development in which videnskab.dk is increasingly recognized as a credible source of science—at least in the eyes of other scientific actors. At the same time, this may also point to a trend in which videnskab.dk is less cited by the general population, whether through discussion forums or blogs, which videnskab.dk itself has described as an important target group (Degn et al. 2023, 5–6).

The prevalence of certain actor types may also reflect societal and technological changes. Thus, the decrease in citations of videnskab.dk cited by the general population could be attributed to social media's partial takeover of blog-related activities, which, as stated earlier, has been excluded from the datasets. Likewise, the transformation of libraries from being minimally present to becoming the most prevalent actor in videnskab.dk's in-links in 2022 (Figure 2 in appendix) may mirror the institutional development libraries have undergone since digitalization, characterized by a growing demand for users' electronic access to scientific journals and papers (Povlsen 2016).

The continuing decrease of non-scientific actors in videnskab.dk's out-links over the years could also be time-related, reflecting videnskab.dk's gradual foothold since its establishment in 2008. This development may have reduced the need to produce content based on what is already popular among the general population. At the same time, this might also explain why videnskab.dk's in-links appear to be converging with its out-links

(actor-wise) over the years, shaping the identity of the videnskab.dk of today.

As demonstrated above, the use of archived web data has offered a deeper understanding of the 'whats' and 'whys' surrounding present-day videnskab.dk by providing insights into older versions hereof. As such, the analysis of 'what has been' can serve as the foundation for the analysis of 'what is', proving the archived web to be an invaluable source not only for historical but also contemporary analyses.

3.4 A transnational perspective

Having examined the actors present in videnskab.dk's out-links and in-links across the years 2009, 2014, 2018, and 2022, the decision was made to examine videnskab.dk from an international perspective. This involved conducting a network analysis of videnskab.dk alongside similar international scientific websites to identify potential differences or similarities in connections and use of sources.

The analyzed network is based on hyperlinked citations, referring to out-links found in the content of each website, which have been harvested—thus not collected from a web archive—in 2022 from videnskab.dk, and a corresponding English (newscientist.com), American (scientificamerican.com), French (futura-sciences.com), German (scinexx.dk), Norwegian (forskning.no), and Swedish (fof.se) website. These sites were chosen for their journalistic profile and orientation towards conveying scientific content, which resemble that of videnskab.dk.

Merging the edges allowed us to examine the in-degree of each node as a reflection of the number of unique connections, facilitating the identification of shared sources among the individual websites. The analysis showed that approximately one-tenth of the nodes in the network could be identified as shared sources, with each node having no more than 1.195 connections (degree value), resulting in a sparse network graph with a low density score of 0.002. Additionally, a positive modularity score suggested a tighter connection within the clusters (each representing the individual websites chosen and their dedicated out-links). Thus, one might describe the network as somewhat polarized (Smith et al. 2016), also visible in the overview of the graph (Figure 4 in appendix).

Most of the shared sources can be identified as scientific actors such as academic portals and publishers (Figure 21 in appendix). Notably, the ones with the highest in-degree tend to be engaged in various branches of the natural sciences, such as nature.com, ncbi.nlm.nih.gov, pubmed.ncbi.nih.gov, and nasa.gov, to name a few. The use of these specific sources may suggest a shared appreciation, potentially owing to

their accessibility for general research or their emphasis on content related to the natural sciences. This might indicate a perceived academic or international value, which could be of interest considering their shared Top-Level Domains (TLDs). Zooming in, it becomes apparent that videnskab.dk has a higher number of shared sources with the other Nordic websites, specifically forskning.no and fof.se, along with the American website, scientificamerican.com, than with the remaining websites. This may imply more similarities in content among these platforms.

Taking a broader view of sources or connections in general, the Nordic websites share structural similarities by having a large number of edges compared to the remaining websites, thus acting as larger hubs. Furthermore, upon merging the edges, videnskab.dk appears to have more distinct or individual connections than any other website in the network, indicating a wider use of different sources (Figure 5 in appendix).

Having examined what can be characterized as formal connections in the network, we decided to delve deeper into the Country Code Top-Level Domains (ccTLDs) of each source, as a means of exploring what Park and Thelwall refer to as the "trans-national knowledge flow" (2003, 12). This also allowed for an exploration of how we might understand the said notion of 'different' sources. Given our interest in the connectivity between the chosen websites, we concentrated on the ccTLDs associated with their respective nationalities, resulting in the following list of identified ccTLDs:

- .se (Swedish)
- .no (Norwegian)
- .us / .gov (American)
- .uk (English)
- .fr (French)
- .dk (Danish)
- .de (German)
- "other" (unidentified ccTLDs)

Results showed that videnskab.dk exhibited the largest variety of ccTLDs (Figure 6 in appendix). This was somewhat mirrored by the Norwegian website, where sources also demonstrated a wide variety of ccTLDs compared to the other websites. However, most of videnskab.dk's sources shared the website's own national origin (Danish), a pattern also observed in the other Nordic websites. For instance, forskning.no and fof.se mainly connect to Norwegian and Swedish sources, respectively, suggesting a shared 'favoring' of national sources among the Nordic websites.

While the use of national sources was also evident among the remaining

websites, sources representing neither the website's own nationality nor that of the others in the network, but instead those categorized as "other", were more prevalent. This might indicate a more widespread use of sources in terms of nationality, as national ccTLDs seemed to be less favored in these instances. On the other hand, since the specific ccDLTs in the "other" category have not been identified, it remains unclear whether the "other" category constitutes a broad range of ccTLDs (apart from those identified), or the same unidentified ccTLD—potentially representing a narrower, rather than widespread, use of (international) sources.

All of the aforementioned points indicate that videnskab.dk shares more similarities with the other Nordic websites, both in terms of structure and sources, effectively distinguishing itself from the American and the European websites. With only a few sources serving as common denominators, the graph presents a rather polarized network, portraying a sense of disconnectivity in an otherwise globalized world. Nevertheless, given the temporal limitations of the method used for data collection, along with the continuous evolution of online web content (Brügger 2019), the results might merely be a reflection of the temporary, emphasizing the need for further investigation.

4 Discussion

In this section, we will briefly discuss some potential implications of the results and the methods.

Analyzing hyperlinks, including actor types and their positions in the hyperlink network, provides an opportunity to uncover the structure that underpins one of the main communicative infrastructures in contemporary society—the web. Hyperlinks can be likened to the 'roads' that allow a web user to 'travel' from one communicative entity to another. However, this map of roads is not readily visible when navigating hyperlinks from one website to another. Users are embedded in the web landscape and cannot see its structure from above. And not only are all the potential roads not visible, so is the role of the interlinked entities, the websites, because their status, that is their centrality, cannot be fully understood as such while moving around on the web. Only a network analysis can provide an overview map of the interconnected websites and the distinct role each one plays on the entire map. However, mapping the roads does not reveal information about the content of websites, their creators, or the frequency of visits. To complete the picture, one has to include analyses of website content and user statistics. Nevertheless, providing a map of hyperlinks is a valuable first step in comprehending the nature of our communicative infrastructure.

The international hyperlink network analysis of videnskab.dk and the contemporary online web indicates numerous narratives when extending the scope to include websites and hyperlinks beyond the nation of origin for each website. However, given the often closed nature of national web archives, expanding this analysis to a comprehensive transnational analysis would require access to all relevant national web archives, including easy access to the hyperlinks. Unfortunately, as for now, such an analysis is not feasible due to the lack of transnational research infrastructure between national web archives.

Web archives such as Netarkivet can be valuable tools for mapping and identifying website content through their recordings of various hyperlink data, such as link paths or link positions. However, depending on the design and software used, certain tools may struggle to perform an accurate reading of the code embedded within the structure of specific hyperlinks. This can lead to inaccurate data and introduce uncertainty regarding the validity of any findings. Ensuring a tool's alignment with a chosen collection of hyperlinks can be challenging with large datasets constituted by hyperlinks with different structures. While this calls for practical solutions, questioning how a tool's interpretation of data differs from our own may be a valuable step in determining the usability and validity hereof. Thus, apart from demonstrating web archives to be an indispensable source in analyses of the web, historical as contemporary, we also encourage a critical reflection on the data and findings generated through the use of these.

5 Conclusion and next steps

As demonstrated in this chapter, the archived web is not only a valuable source for analyzing the past, but also enhancing our understanding of the present communicative infrastructure of the web and its hyperlinks, particularly in providing information about in-going hyperlinks to websites, which cannot be collected from the live web. It thus makes a plea that web archives are not only significant for historical studies, but also for contemporary investigations.

While the present study is limited to focusing on one website and one iteration of hyperlinks from this website, it serves as a model to inspire broader studies, such as those exploring entire national web domains as outlined by Brügger et al. (2020). It can—and should—encourage more focused analyses of the most central nodes.

Finally, to conduct an exhaustive analysis of the hyperlink network in which any given website is embedded, it is pivotal to create research infrastructures that extend beyond the borders of national web archives.

Acknowledgements

We would like to express our gratitude to the contributors to the evaluation report of videnskab.dk: Hans-Peter Degn, Christiane Særkjær, Line Hassall Thomsen, and Maja Sonne Damkjær. Special thanks to the IT developer at the Department of Media and Journalism Studies, Ulrich Karstoft Have, and Netarkivet for facilitating the extraction of the data.

References

Ackland, Robert, and Ann Evans. 2017. "Using the web to examine the evolution of the abortion debate in Australia, 2005–2015." In *The web as history: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 159–189. London: UCL Press.

Beaudouin, Valérie, Zeynep Pehlivan, Peter Stirling. 2018. "Exploring the memory of the First World War using web archives: Web graphs seen from different angles." In *The SAGE handbook of web history*, edited by Niels Brügger and Ian Milligan, 441–463. London: SAGE.

Brin, Sergey, and Lawrence Page. 1998. "The anatomy of a large-scale hypertextual Web search engine." *Computer Networks and ISDN Systems* 30, no. 1: 107–117, https://snap.stanford.edu/class/cs224w-readings/Brin98Anatomy.pdf.

Brügger, Niels. 2013. "Historical Network Analysis of the Web." *Social Science Computer Review* 31, no. 3: 306–321. https://doi.org/10.1177/089443931245426

Brügger, Niels. 2019. "Understanding the archived web as a historical source." In *The SAGE handbook of web history*, edited by Niels Brügger and Ian Milligan, 16–29. London: SAGE.

Brügger, Niels. 2021. "Digital humanities and web archives: Possible new paths for combining datasets." *International Journal of Digital Humanities* 2, no. 1–3: 145–168.

Brügger, Niels. 2022. "Tracing a historical development of conspiracy theory networks on the web: The hyperlink network of vaccine hesitancy on the Danish web 2006–2015." *Convergence* 28, no. 4: 962–982. https://doi.org/10.1177/13548565221104989

Brügger, Niels, Ditte Laursen, and Janne Nielsen. 2017. "Exploring the domain names of the Danish web." In *The web as history: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 62–80. London: UCL Press.

Brügger, Niels, Janne Nielsen, Ditte Laursen. 2020. "Big data experiments with the archived Web: Methodological reflections on studying the development of a nation's Web." *First Monday* 25, no. 3. https://firstmonday.org/ojs/index.php/fm/article/view/10384.

Cowls, Josh, and Jonathan Bright. 2017. "International hyperlinks in online news media." In *The web as history: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 101-116. London: UCL Press.

Degn, Hans-Peter, Christiane Særkjær, Line Hassall Thomsen, Maja Sonne Damkjær, and

Niels Brügger. 2023. “Evaluering af Videnskab.dk”. Aarhus: Center for Kulturevaluering, Aarhus Universitet. https://ufm.dk/publikationer/2023/evaluering-af-videnskab.dk.

Fage-Butler, Antoinette, Loni Ledderer, and Niels Brügger. 2022. “Proposing methods to explore the evolution of the term ‘mHealth’on the Danish Web archive.” *First Monday* 27, no. 1. https://firstmonday.org/ojs/index.php/fm/article/view/11675.

Foot, Kirsten, Steven M. Schneider, Meghan Dougherty, Michael Xenos, and Elena Larsen. 2003. “Analyzing Linking Practices: Candidate Sites in the 2002 US Electoral Web Sphere.” *Journal of Computer-Mediated Communication* 8, no. 4. https://doi.org/10.1111/j.1083-6101.2003.tb00220.x.

Jackson, Michele H. 1997. “Assessing the Structure of Communication on the World Wide Web.” *Journal of Computer-Mediated Communication* 3, no. 1. https://doi.org/10.1111/j.1083-6101.1997.tb00063.x.

Meyer, Eric T., Taha Yasseri, Scott A. Hale, Josh Cowls, Ralph Schroeder, and Helen Margetts. 2017. “Analysing the UK web domain and exploring 15 years of UK universities on the web.” In *The web as history: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 83–100. London: UCL Press.

Moreno, Jacob L. (1934). *Who shall survive? A New Approach to the Problem of Human Interrelations*. Washington, DC: Nervous and Mental Disease Publishing.

Park, Han Woo, and Mike Thelwall. 2003. “Hyperlink Analyses of the World Wide Web: A Review.” *Journal of Computer-Mediated Communication* 8, no. 4. https://doi.org/10.1111/j.1083-6101.2003.tb00223.x.

Povlsen, Karen Klitgaard. 2016. “BØGER! BØGER! BØGER!” In *Dansk Mediehistorie, vol 4.*, edited by Klaus Bruhn Jensen. Frederiksberg C: Samfundslitteratur.

Ryfe, David, Donica Mensing, and Richard Kelley. 2016. “What is the meaning of a news link?” *Digital Journalism* 4, no. 1: 41–54.

Smith, Marc A., Lee Raine, Ben Schneiderman, and Itai Himelboim. 2014. “Mapping Twitter Topic Networks: From Polarized Crowds to Community Clusters.” *Pew Research Center*, February 20, 2014. https://www.pewresearch.org/internet/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters/.

Stevenson, Michael, and Anat Ben-David. 2018. “Network analysis for web history.” In *The SAGE handbook of web history*, edited by Niels Brügger and Ian Milligan, 125–137. London: SAGE.

Wasserman, Stanley, and Katherine Faust. 2009 [1994]. *Social network analysis: Methods and applications*. Cambridge: Cambridge UP.

Weber, Matthew S. 2017. “The tumultuous history of news on the web.” In *The web as history: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 83–100. London: UCL Press.

Webster, Peter. 2017. “Religious discourse in the archived web: Rowan Williams, Archbishop of Canterbury, and the sharia law controversy of 2008.” In *The web as history: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder, 190–203. London: UCL Press.

Weltevrede, Esther, and Anne Helmond. “Where Do Bloggers Blog? Platform Transitions within the Historical Dutch Blogosphere.” *First Monday*, February 2, 2012. https://doi.org/10.5210/fm.v17i2.3775.

Do user comments belong to journalistic articles? A brief visual history of user interaction on selected German and American news websites 1996–2024

Johannes Paßmann, Martina Schories, Paul Heinicker

Abstract: The chapter reconstructs a brief history of online commenting, based on the position comments have to journalistic articles on news websites. Its key assumption is derived from paratext theory: Changes in spatial and temporal proximity of texts in the periphery of a main text—such as comments on the same web page as a journalistic article as compared to posts in a separate forum—indicate controversies over relevance of participants in a public discourse. Studying transformations of online comments is thus considered an access point to studying histories of public spheres. With help of a software the authors and colleagues developed, changes in commenting sections are traced and visualized. These changes are detected in a data sample of archived web pages provided by the Internet Archive.

Keywords: commenting, web archive, paratext, news websites, user comments.

In recent decades, online comments have experienced a fluctuating reputation, embodying both the hope and disillusionment of the democratic potential of the internet. Popular cultural memory of online comments' history might paint a clear picture: a participatory culture lasting until the late 2000s was followed by a decay of commenting practices within the platformized web and its influx of new users and devices. However, closer inspection indicates a more complex history.

Disruptive communication practices were rampant on the internet even before the World Wide Web was established. Practices like 'flaming', for example, were already widespread on the Usenet (Kiechle 2022). In our data analysis outlined below, we found, for example, that the *Los Angeles Times* had already shut down its bulletin boards in 1996 due to racial slurs (latimes.com 1996, later replaced by discussion boards the following year, latimes.com 1997). In interviews with some of the earliest bloggers, we learned that when user comments were introduced to blogs in the late 1990s, blog owners could hark back on practices of moderating problematic Usenet discussions in the 1980s.[1]

[1] Interview conducted by Lisa Gerzen and Johannes Paßmann with Dori Smith (Blog "The Backup Brain") on October 25, 2023 in San Francisco, Interview conducted by Lisa Gerzen and Johannes Paßmann with Matt Haughey (Blog "A Whole Lotta Nothing", Founder of *Metafilter*) on October 27, 2023 in Portland.

Johannes Paßmann, Ruhr-University Bochum, Germany, johannes.passmann@rub.de, 0000-0002-2822-6082
Martina Schories, Ruhr-University Bochum, Germany, martina.schories@rub.de, 0009-0003-1322-5356
Paul Heinicker, Ruhr-University Bochum, Germany, paul.heinicker@rub.de, 0009-0001-7695-974X

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Johannes Paßmann, Martina Schories, Paul Heinicker, *Do user comments belong to journalistic articles? A brief visual history of user interaction on selected German and American news websites 1996–2024*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.20, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 223-246, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

Still, something changed: a recurring pattern in interviews about the history of the internet as a public medium is that, at certain points in time, such as the 'eternal September' of 1993, the 9/11 terrorist attacks in 2001, the proliferation of smartphones and the 'platformization' of the web around 2010, or the Trump election and Brexit vote in 2016, commenting cultures transformed. This is also the case for practices of the Usenet: 'flaming' differed significantly from later practices of 'hate speech' (Kiechle 2022). As a result, examinations of online commenting histories will inevitably uncover both continuities and transformations in the technologies and practices of user-generated content.

We argue that one entry point into understanding these continuities and transformations is the positioning of users' content on journalistic websites. The placement of this content marks a zone of contact between those who published a text and those commenting on it. The changing positions of users' and producers' content to one another can be read as traces to the transformations of commenting practices specifically and public discourse online more generally. These changes can be studied through web archives. As a result, web archives and their preserved snapshots of websites can provide a useful lens for studying histories of public spheres online, as public spheres are not only determined by the question of *who may speak* but also of *who may speak when and where*. We want to demonstrate this approach by studying how journalistic media have positioned themselves in relation to these (at each point in time: new) modes of participation.

The question of *who may speak when and where* positions online commenting in a historical and media theoretical context that predates the internet considerably. Gérard Genette's theory of the *paratext* delves into the "undefined zone", or the "intermediary zone between the off-text and the text" (C. Douchet and A. Compagnon, as quoted in Genette 2010, 2. I.o. "zone indécise" and "zone intermédiaire", Genette 1987, 8). Every text is surrounded by numerous peripheral texts, with some, such as titles and book covers, considered *paratexts*, while others simply constitute *discourse*. For this distinction between (more relevant) paratext and (less relevant) discourse, the material position of those texts is crucial. What is in spatial and temporal proximity to a main text has much higher chances to be considered (noteworthy) paratext.

This chapter consequently traces the positioning of user-generated content on news websites over time. We understand the changes of the way journalistic texts and those of their readers are positioned to one another as indicators for a controversy around the larger question of whether users' texts should be considered paratexts to articles or just discourse. The spatial and temporal positioning is not the only actor in the negotiation of that

controversy, but a crucial one that is worth following in a web-historical context.

1. Sample

For this purpose, we leverage a specific, yet limited access to a history of online commenting: data of archived news websites obtained through the Internet Archive (IA) in the context of the *Archives Unleashed Cohort Program*. We analyze the data using the 'Technograph', a tool we developed that allows us to trace and visualize commenting systems and their changes. At its core is an automated pattern recognition process realized through the programming language R, which looks for structural evidence of commenting systems in the HTML code of these websites, such as occurrences of certain form tags or used frameworks.[2]

This process of extracting and visualizing patterns is a starting point for a qualitative analysis of the archived snapshots that the tool pointed to. Hence, in the analysis outlined below, the function of our software is to help us navigate through the multiplicity of data provided by the IA. This facilitates a more efficient use of the visual interface of the IA's Wayback Machine (IAWM) in a next step, allowing us to focus primarily on a set of archived web pages and situate them in a broader web-historical context.

Generally, our method has three blind spots. Firstly, we rely on data from a web archive, and these archives are constitutively incomplete (Brügger 2018; 2008). In our case, commenting sections are often not archived at all. Sometimes, for example, when they are implemented with JavaScript, we only find very rudimentary traces in the archive. The second blind spot of our methodology is that we primarily focus on web artifacts rather than on the practices and actor-networks they have been part of (Paßmann and Gerzen 2024).

We counter both blind spots by comparing various cases—spanning time and different countries. Thus, the central question of this chapter is not about the single positioning of a commenting section in relation to a news article, but rather about what changes in these positions over time might reveal. To make these formal and structural changes in the web pages more legible, we have developed what we call a Historiogram (Figure 1). It illustrates the chronological development of commenting practices with a specific focus on their position within news websites. The Historiogram follows the structure of a dual-axis chart. The x-axis encodes the temporal

[2] The Technograph has been developed by Martina Schories in cooperation with Lisa Gerzen, Robert Jansma, Anne Helmond, and Johannes Paßmann. A beta version is available via https://shiny.sfb1472.uni-siegen.de/b03-technograph/ and the code repository can be found on GitHub: https://github.com/SFB1472/tdp-b03-technograph.

course of observations from 1996 to 2024 and the y-axis represents the actual position of the observed commenting sections. The Technograph creates simple Historiograms automatically. However, the one displayed below has been built manually in order to show more details in a single figure and to add information we found browsing the IAWM (on the basis of the snapshots the Technograph pointed us to).

For each media outlet, the figure distinguishes between two categories: user posts that are situated directly on the article's web page (on site) and those that are not (off site). In order to make the aforementioned limits of our observation visible, we visualize periods of uncertain data with a transparent color gradient. In addition, we used small circles to mark points in time where we found hints of major redesigns of the respective websites. The results of these redesigns are partly illustrated on the right-hand side of the diagram with corresponding screenshots. All redesigns are explicitly referenced in the text.

We systematically analyzed the websites of four nationwide German print media outlets considered 'quality media' with a vested interest in providing a public discourse. This selection includes two daily newspapers (Frankfurter Allgemeine Zeitung, faz.net and Süddeutsche Zeitung, sz.de), one weekly paper (Die Zeit, zeit.de), and one weekly magazine (Der Spiegel, spiegel.de). However, this sample introduces a third blind spot, which we can only partially counter: a bias in the selection of analyzed websites. To mitigate this, we also sampled the website of a regional daily newspaper (Augsburger Allgemeine) situated on the periphery of 'quality' newspapers. For an international comparison, we added the New York Times (NYT), and for a non-German, non-global perspective, we analyzed the Los Angeles Times (LAT). Despite these efforts, a bias towards 'quality' newspapers persists.

2. Comments as paratexts?

Reading the Comments, a book Joseph M. Reagle published in 2016 about "Likers, Haters and Manipulators at the Bottom of the Web", argues that the position of online comments in "the bottom half" of websites characterizes their status. As a marginal medium, they had often been neglected—too often, he argues. Their positioning reflected their marginality as a zone that was "much like California during its gold rush […] lively and lawless." While reluctance to read the comments was understandable, Reagle advocates it was "wise to understand them", countering the observed tendency in the mid-2010s to disable or ignore commenting sections as the bottom half (Reagle 2015, 3).

We also contend that the question of where comments are situated holds semantic significance. In literary and media theory, this argument has been elaborated most prominently by Genette (2010; 1997) for paratexts of books. Texts positioned in the periphery of a main text contribute to its meaning, and this meaning also depends on the material positioning of the potential paratext in the periphery to the main text. For instance, a critic's remark printed on the back cover of a novel, due to its proximity to the main text, becomes a "peritext", which means in most cases that it can be considered a paratext. When the same remark remains "anywhere outside the book", the situation is less clear: it could, under certain circumstances, be considered an "epitext" (Genette 2010, 344).

In essence much paratext research follows the logic of Goffman's "frame analysis" (Goffman 1986 [1974]). Paratexts function as a kind of frame for understanding the main text (Stanitzek 2005; Dembeck 2007). In recent decades, the concept of paratext has been used primarily to expand the boundary between text and discourse. The prevailing argument has been the necessity to broaden the concept of text, recognizing more elements in the periphery of a main text as integral to it. Along these lines, trailers have been considered paratexts of movies (Hediger 2004; Zons 2007), packages, controllers, and similar items have been categorized as paratexts of computer games (Jones and Thiruvathukal 2012), and gaming streams have been considered the main text with the games themselves functioning as paratexts (Consalvo 2017), among other examples.

Paratext research, advocating for extending the boundaries of the units of analysis, echoes a typical argument of the second half of the 20th century: Similar to Goffman's argument of the "frame", most media theories have advocated to extend the boundaries of human action (e.g., considering media the "extensions of man", McLuhan 1994), and the "project" of actor network theory (ANT) was "to extend the list and modify the shapes and figures of those assembled as participants […]" (Latour 2005, 72). This move towards extension of the object of analysis has also been the main argument of paratext theory and research.

As argued earlier, today, it would be of limited benefit to pursue this path further by simply extending the textual boundaries of online texts to online comments and advocate—similarly to Reagle—for their inclusion into the textual unit of main text and paratext (Paßmann 2023). Rather than participating in the boundary work of determining what belongs to the frame of the text, i.e. what is paratext and what is not, we contend that this question itself marks a controversy that should be studied with the concept of paratexts (ibid.). All kinds of actors, including websites, participate in this boundary work, and this boundary work can be rephrased as the question: what should (not) be considered a paratext?

When zeit.de relaunched its modes of user interaction in September 2009, they published an article explaining their updates (zeit.de 2009). They noted that the most obvious change was the "positioning of user comments directly below the articles," stating, "This conforms with our idea that debates are an important part of a text" (ibid.). In Genette's sense, this would be the most explicit form of authorizing a text in the periphery, rendering it a paratext. This authorization functions through two acts: the quoted explanation, and the placement of comments in direct neighborhood to the articles.

This extension (and delimitation) of textual boundaries is a daily practice for people dealing with online comments. When the editors of Zeit Online write that comments belong to the text and simultaneously redesign the commenting section to make comments readable alongside the main text (i.e. the article), they extend the textual boundaries of their article. Comments are not considered part of the article, but they gain recognition as their (legitimate) periphery. On the one hand, there are practices of "sorting texts out" of this legitimate periphery: moderation of commenting sections decides what may be visible in the texts' periphery. Journalists also contribute to delimitation with their speech acts, such as when they claim that they never read the comments because they are useless or even harmful (Paßmann 2023). When Reagle (see above) advocates for "reading the comments", he also works on the boundaries of the main text. In that case, his argument supports an extension of the text boundaries; *reading the comments* means recognizing comments as paratexts.

This implies a plethora of actors negotiating the text boundaries between online articles and their comments, among them journalists, commenters, moderators, websites and their positioning of texts to another, and, last but not least, academic research, by advocating for extending the boundaries—in the (typical 20th-century) tradition of paratext theory. The relevant boundary being negotiated here is not so much between *text* and *paratext*, but rather between *paratext* and *discourse*. This is a fundamental zone of conflict because the distinction between paratext and discourse challenges what and who belongs where.

We argue that the negotiation of these boundaries is a central practice in the history of online commenting. Online commenting is inherently concerned with the negotiation of the paratext/discourse distinction. The way users' content is positioned on websites in relation to the main text is a crucial factor in this negotiation practice. At stake in this negotiation is the question of recognition: in a quite Hegelian sense, websites participate in determining which voices are recognized as (peripheral) members of the main text.

3. Entering the zone of conflict: The example of the NYT

Whereas zeit.de appears to align itself more or less directly with its commenters—or, more precisely, their comments—, the NYT appears to have been keen to maintain a clear distinction between comments and the journalists' articles. The NYT has a longstanding tradition of intensive user interaction. In the earliest archived snapshots from 1996, there is a call to "join the discussion in the new forums" (nyt.com 1996). This interaction remains in dedicated spaces of specific web pages—and not below the articles.

The first texts with users' comments visible at the bottom half of the web page (within the same window) were found in archived snapshots from 2009 (nyt.com 2009). However, all these texts are framed as blog posts—in this case from the NYT blog "Room for Debate": roomfordebate.blogs.nytimes.com. This blog is subtitled "a running commentary on the news" (ibid.). This implies that comments are not directed at the NYT authors' articles, but rather at *the news*. The editors only initiate discussions on current topics, such as the rescue of the American International Group (AIG) with "$170 billion in United States taxpayer bailout money." External specialists are invited to comment on these news topics, and in the bottom half of the web page, below the posts from expert bloggers, users' comments on the topic are displayed. In the right column of the page, selected user comments on all currently discussed topics are prominently visible as "Comments of the Moment" (ibid.). These comments appear on the same page as posts from expert bloggers, who, at least in the cases we found in the archive, are not members of the NYT. Even here, comments cannot be considered paratexts to journalistic main texts (or even blogging main texts) since there is not a main text these comments refer to; there is just a 'topic'. They are, in Genette's sense, merely discourse from the perspective of journalistic articles.

During the 2010s, the NYT seems to have extended its textual boundaries considerably. The first articles featuring a 'comments' icon at the end of the text date from February 2010 (nyt.com 2010a). Not all articles had these commenting options at that time. Clicking the comments icon (in the shape of a speech balloon) opens a dedicated subsite for 'readers' comments' (nyt.com 2010b, see Fig. 1 no. 8). This subsite begins with a very brief teaser of the article (one sentence) and its heading. Below that, comments are displayed in full length. On the one hand, comments are explicitly referred to in the main text. However, it requires the user's action to read the comments, and even after that, the main text and comments are not visible on the same web page. Without delving too deeply into Genette's nomenclature here, this commenting section might be considered an

"epitext", a peripheral text that is not as easily recognizable as a paratext as the "peritexts", which are in direct proximity to a main text. Whether or not comments 'belong' to the articles and would thus be a paratext (in the position of an epitext) is consequently not clearly determined by the NYT. This changed in the mid-2010s.

In snapshots from August 2016, we, for the first time, found comments displayed on the same web page as articles. In the right column, not even on the bottom half, but directly beside the text and between other boxes in that column, such as 'related texts' and 'trending', three 'recent comments' are displayed (nyt.com 2016b). Below these three comments are two links, one named 'see all comments', and another one 'write a comment'. A click on 'see all comments' opens a margin that fills the whole right column of the page (see Fig. 1 no. 9).

The margin has two tabs, one with 'all comments', and one with 'readers' picks'. On top of the margin, a notice reads: "The comments section is closed. To submit a letter to the editor for publication, write to letters@nytimes.com" (ibid.). The comments are threaded and marked with counted 'Recommends'. To the left of the 'Recommend' counter is a Facebook-like thumb, and to the right of it are Facebook and Twitter icons. The 'readers' picks' are ranked according to the number of 'Recommends'. On top of the article, in a row right below the heading, a speech balloon is depicted with the number of comments this article 'received' (see Fig. 1 no. 10). This balloon is repeated at the bottom of the article in larger size—not below it, but indented into its last paragraph. A click on the balloon again opens the comments in the margin, filling the right column of the page.

This website design approach undertakes several efforts to present the comments as something to read (and write) while or after reading the article. In other words, the design strongly attempts to render the comments paratexts. After (at least) twenty years of intensive user interaction that was very keen to keep comments away from the main text, the design found in snapshots from 2016 even exceeds that of classic weblogs displaying comments on the bottom half. Commenting sections are indented into the main text itself, displayed next to the text so that, as a result, at least the first three 'most recent' comments are perceived during reading the main text, rather than afterward.

With one more click, opting to open the commenting section fills the screen with a binocular (or *stereoscopic*, if you will) view of two almost equal columns: the article on the left and the comments on the right. Moreover, as the number of comments is displayed on three points on the website, the design offensively implies that commenting is a frequent, usual, and popular practice that *one does* when reading NYT articles. This raises the question of why, after decades of consistently separating articles and

comments, in 2016, the NYT brought them even closer together than classic blogs. The fact that this is the year of the Brexit referendum and the Trump election begs interpretation.

The popularity markers remain on the NYT article's websites until the present day (nyt.com 2024). The number of comments still appears three times on the website, though in a more modest manner. The speech balloons are much smaller and not indented into the main article text. Clicking on the balloons still opens the comments in a dedicated margin in the right column of the website, but the stereoscopic view has disappeared. While in 2016, text and comments could be fully read next to one another, in the 2024 interface, first, the comments overlap the article. Second, the article is turned dark. And third, one cannot scroll the article anymore when comments are open. This pattern is even more noticeable on the NYT for iPad app, where even less of the commented article is visible (approximately 20% of the—dark—screen), and for the NYT mobile app for Android smartphones, where the comments appear in an entirely new window. Moreover, in the current NYT design, all comments are at least one click away from the article text; no comment is displayed by default on the same web page as the article.

In that sense, to a large extent, the NYT reverts to the pre-2016 textual order. In this historical view, the offensive speech act, or rather: web-design act, of ostensively rendering comments paratexts appears as an exception. The NYT's web and app design no longer suggests by default to read the articles with their comments. In contrast to the mid-2010s, today, one can easily read the NYT articles without users' comments.

4. German quality papers

For all four cases of German quality print media websites we analyzed, we found a pattern generally similar to the sequence outlined above for the NYT. They all started with separate forums (frequently positioned in the tradition of letters to the editors) and, over the 2000s and 2010s, gradually brought user content closer to the articles. From there, however, they experimented with a diversity of strategies.

The Süddeutsche Zeitung introduced online comments displayed below the articles by default in late spring 2007 (sz.de 2007, see Fig. 1 no. 5), after hosting a forum (i.e. an architecture containing user posts on a separate web page) for article comments for at least four years already (sz.de 2003). The earliest example of a commenting section below an article we found for zeit.de dates to July 2006 (zeit.de 2006). In a snapshot from November 2005, we found hints to 'Leser-Kommentare' (reader's comments) next to an article text (right column), seemingly linking to another web page

(zeit.de 2005). A ‘forum’, however, can already be found in April 1999—and this forum is hosted not only on a separate web page, but also under its own third-level domain (zeit.de 1999).

For spiegel.de, we found the first forum in snapshots from 1997 (spiegel.de 1997). In June 2001, a subsite for letters to the editors concerning online articles was introduced (spiegel.de 2001b), and a ‘mailto’ link for these ‘online letters’ was displayed below the articles in those days (spiegel.de 2001a). Selected (e-mailed) letters to the editors were published on a separate website (spiegel.de 2001b). The first time we found user comments below an article was in a snapshot from March 2010 (with posts from January, spiegel.de 2010). The first snapshots we found of these posts labeled as comments (‘Kommentar’) date back to 2014 (spiegel.de 2014a).

The website of the *FAZ* is comparably slim before 2001; we could not find any trace of user or reader interaction (faz.net 1997). The newspaper launched its own ‘portal’ FAZ.NET in January 2001 (de.wikipedia.org 2024). In snapshots of that year, we, for the first time, found hints to a forum (a non-loadable graphic named “Foren & Chat”) (faz.net 2001). In January 2002, a forum subpage was archived for the first time (faz.net 2002). Until mid-June 2005, the forum was linked on the starting page (faz.net 2005). In a snapshot from November 2005, we, for the first time, found ‘readers’ opinions’ displayed below the articles. Visible on the same web page are only the comments’ headings (which refer to the article, faz.net 2005).

In October 2011, the FAZ conducted a major relaunch (faz.net 2011). As a result, the number of comments an article ‘received’ was displayed in a speech balloon next to the article’s teaser (on the starting page, for example). At the end of the article itself, there is an icon inviting comments. To read the comments, one must click a specific tab at the beginning of the article (see Fig. 1 no. 1). Once the tab is open, the article disappears (ibid).

That means, on the one hand, the links to the comments are prominent and commenting is framed as a popular practice. On the other hand, comments and articles do not appear in the same window. In that sense, the FAZ’s commenting system, as relaunched in 2011, shares similarities in terms of paratextuality with the current presentation of comments in the NYT.

Three years later, the FAZ discards the tabbed commenting system again and relocates comments below the articles. However, again, only the comments’ headlines are visible by default (faz.net 2014). The comments’ peritexts (their headlines) are peritexts for the articles, as they are positioned on the same web page, but not the comments themselves (see Fig. 1, no. 2).

Today, the website’s textual order is again comparable to that of the NYT, as comments only appear in the right column of the website upon

demand. When the comments are opened, they can, given the browser's window is wide enough, be displayed in the stereoscopic view, similar to what the NYT had from 2016 on (and later changed). However, as of today, for FAZ, users need to create an account and log in to write and read comments, placing the comments one degree further away from the main text.

On a general level, both the FAZ and NYT not only follow the two steps from forum to comments as peritext, but also also share the third step of providing a separate commenting space that one has to actively opt into in order to read the comments—and in order to read the comments and the articles (more or less easily) in the same window.

The current Süddeutsche Zeitung (SZ) appears to follow a similar layout at first glance. Positioned above the commentable articles, just below the headline, is a speech balloon with a counter indicating the number of comments. When clicked, a designated margin on the right side of the web page opens, allowing for a stereoscopic reading. The differences from the FAZ and NYT mainly concern the conditions for commenting and reading. For the NYT, a paid subscription is required not only to access articles, but also to read and post comments). On the other hand, the FAZ requires only a (free) account to read and post comments, a privilege limited to articles without a paywall (something most articles have). With the SZ, an account is only required for posting comments, not to read them. However, similar to the FAZ, the SZ has a mix of paywalled and free articles.

The history of SZ's online comments is a bit more complicated, with only a few highlights mentioned here.[3] They disabled their commenting system in 2015 and introduced a specific forum that aimed not to comment on individual articles (as paratexts to texts), but rather on *topics*—like how journalists comment on topics, too (Wüllner 2015). This resembles the NYT's *Room for Debate* subsite, observed in snapshots from 2009. Briefly, before transitioning from comments back to a forum), starting in 2014 each article subpage linked to the *rivva Debattenmonitor,* an external website that displayed all public comments, tweets, and posts about this article (sz.de 2015).

In January 2021, the SZ changed the audience dialogue software from *Disqus*—a commenting system prevalent worldwide for the past two decades that also changed the way they considered comments, problematizing them more and more (Paßmann, Helmond, and Jansma 2023)—to *Talk*, a software developed by Mozilla, the Washington Post, and

[3] Lisa Gerzen and Johannes Paßmann also interviewed the person responsible for SZ's comments section since the mid-2000s. Some aspects from that interview were published in Paßmann and Gerzen 2024. More detailed publications, including the SZ's specific audience interaction history, will follow.

the NYT. However, for the purpose of this article, our focus will be on the most fundamental changes in the textual order of articles and user-generated content. It was not until June 2022 that the SZ announced a 'Relaunch' of readers' comments on their own web pages:

The commenting function will be part of the articles again. Right next to the opinions of the authors are the opinions of the readers. The reason for this change is, in addition to the associated appreciation for reader comments, also the clearer reference to the article and passage in question. This initially applies to all opinion articles on SZ.de. But it will gradually be expanded to include other articles (sz.de 2022).

First of all, we can see here that the semantics of recognition of readers are, as in most other explanations of editors (re-)introducing comments, setting the tone: engaging users as *more than recipients*. For the chronology of the SZ, however our research, reveals a cyclical journey, starting with (off-site) forums, transitioning to (on-site) comments, reverting to (off-site) forum posts commenting on topics addressed by the editors, too, and recently, reintroducing (on-site) comments (see Fig. 1). However, the way they return to (on-site) comments has transformed.

The reintroduction of commenting sections is, at least initially, limited to articles that are commentary in a journalistic sense. This return to comments aims to redefine them not so much as comments to journalistic texts, but as something that could be termed, alluding to British sports culture, *co-comments*. In this context, the posts are intended to comment on what journalists have commented on and, ideally, complement the main comments. These revived comments have also transformed: paywalls now restrict access to commenting, introducing a selection, or, to use the literal Latin sense of the word ('eligere'), a *re-elitarization*, as choices are made. The SZ prioritizes commenters who pay for their service. While this is a general development that we cannot elaborate on here, the introduction of paywalls signifies a shift in the concept of audience. The media in question are moving away from a focus solely on "getting noticed by many" (Werber et al. 2023) online. Rather, there is an increasing focus on identifying and catering to *their* audience.

Spiegel.de has a similarly, perhaps even more, complicated history following their introduction of users' comments below articles in 2010 (officially labeled as comments only in 2014, spiegel.de 2014a, see Fig. 1 no. 3). What makes this history complicated is that firstly, the Spiegel has its own nomenclature of this genre: the discussions unfold in a forum, however, these forums are often linked under the articles, starting in 2014 with a call to comment. In our understanding, this would classify it as a commenting function, because it is displayed in the same window as the

article (rendering it a peritext), and the content posted is, to varying degrees, readable when only the article is opened.

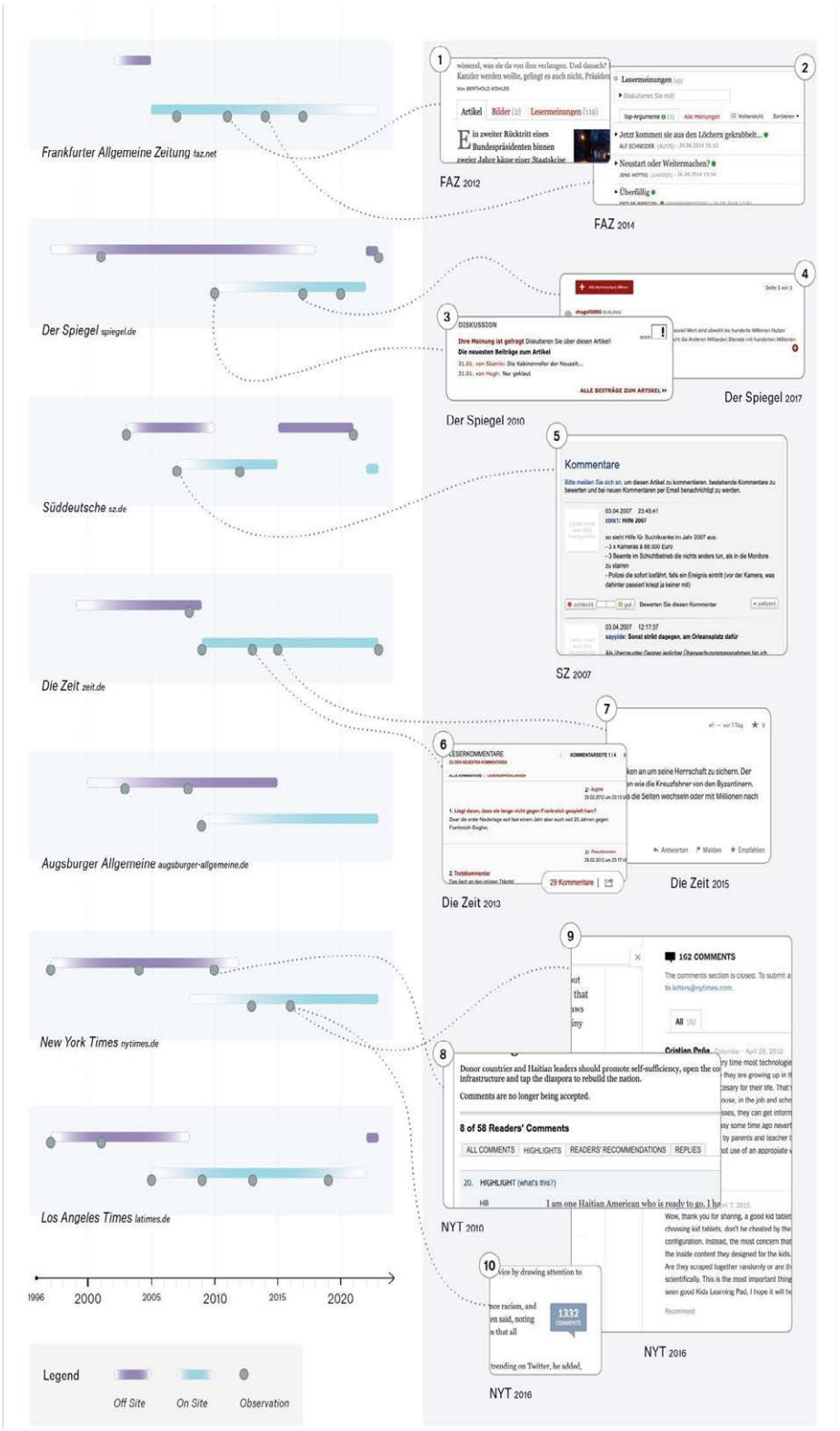
Secondly, as with the other examples discussed earlier, not all articles are open for comments. Some have no option for user interaction at all, while others have topical forums linked on the left margin of the article's web page (spiegel.de 2014b). Thirdly, Spiegel experiments with its user posts. For instance, a redesign first observed in snapshots from 2017, discards the left margin but still allows comments to be displayed below the articles (spiegel.de 2018, see Fig. 1 no. 4). The overview we are outlining here remains thus on a somewhat superficial level.

However, two updates introduced by Spiegel seem noteworthy here. After positioning their 'forum' posts as comments under (selected) articles in 2010 (see Fig. 1 no. 3), they underwent a 'reformed' website design in December 2019 that also included an updated commenting function (spiegel.de 2020). The editors expressed that it was "about time to rethink the commenting section" (ibid.). The update brought several new features reminiscent of social media platforms, such as 'ignore user', 'report comment', a liking function, up- and downvotes, and threaded replies. Most interestingly, the update included a tab with comments recommended by the editors. In this sense, some comments are, using Genette's terminology, authorized by those who wrote the main text. While the commenting section remains an 'undefined zone', certain comments are explicitly recognized as paratexts.

In a second major update, exactly four years later, introducing 'Spiegel Debatte', the editors wrote: "After months of development and a lot of feedback from our users, SPIEGEL Debatte is replacing the previous comment section under articles on SPIEGEL.de" (spiegel.de 2023). This debate is intended to focus on topics defined by the editors on a daily basis and is exclusive to subscribers. They call this feature a 'platform' that incorporates all kinds of platform-like activities, allowing users' posts to be sorted according to 'relevance'. One such activity is the editors' recommendation, denoted by a 'Der Spiegel' icon.

Die Zeit stands out in this comparison: they implemented numerous updates to their commenting sections (see Fig. 1 no. 6 and no. 7). In their current version, for example, users need to scroll down through article recommendations and advertisements to access the commenting section. However, after the updates mentioned above—introducing a commenting function linked under the article in 2006 and displaying comments there in 2009—, they essentially maintained the same textual order to the present day: comments remain on-site.

Figure 1. The Historiogram: A visual chronology over on-site user interactions (i.e. mostly comments) and off-site interactions (forums and platforms).


Frankfurter Allgemeine Zeitung faz.net
Der Spiegel spiegel.de
Süddeutsche sz.de
Die Zeit zeit.de
Augsburger Allgemeine augsburger-allgemeine.de
New York Times nytimes.de
Los Angeles Times latimes.de
1996
2000
2005
2010
2015
2020
Legend
Off Site
On Site
Observation
1
2
FAZ 2012
FAZ 2014
3
4
DISKUSSION
Ihre Meinung ist gefragt Diskutieren Sie über diesen Artikel!
Die neuesten Beiträge zum Artikel
ALLE BEITRÄGE ZUM ARTIKEL
Der Spiegel 2010
Der Spiegel 2017
5
Kommentare
SZ 2007
6
7
Die Zeit 2013
Die Zeit 2015
8
9
Donor countries and Haitian leaders should promote self-sufficiency, open the co
Infrastructure and tap the diaspora to rebuild the nation.
Comments are no longer being accepted.
8 of 58 Readers' Comments
ALL COMMENTS
HIGHLIGHTS
READERS' RECOMMENDATIONS
REPLIES
NYT 2010
162 COMMENTS
NYT 2016
10
1332
NYT 2016

5. Discussion

The general pattern observed across all cases studied is firstly that these news websites typically initiate forums, with this trend often emerging in the 1990s. The Augsburger Allgemeine, considered one of the prominent regional newspapers in Germany—that is to say, just one level below the nationwide quality newspapers—, has a forum, first captured in a 2001 snapshot (augsburger-allgemeine.de 2001). Their article pages subsequently linked to a commenting function in 2008 (augsburger-allgemeine.de 2008), eventually evolving into an on-site commenting section in 2010 (augsburger-allgemeine.de 2010).

We interpret this pattern as the adaptation of the traditional media practice of writing 'letters to the editor' in the new context of the (often 1990s) World Wide Web. As mentioned above, even established media such as Die Zeit make iconographic references to letters to the editor, drawing upon a tradition of public discourse that quality newspapers can leverage.

The subsequent move from forums to online comments reflects an influence from blogging on news organizations. This pattern, as described to us in an interview with early blogger and MetaFilter founder Matt Haughey, saw commenting, in the form of a reader's text displayed on the same web page as the text it refers to, established in blogging around 2000. It was later introduced to news websites due to various challenges these platforms faced, especially those of print media, throughout the 2000s. Factors such as the burst of the dot-com bubble, the growing competition from blogs to journalism in the aftermath of the 9/11 terrorist attacks, the economic success of certain blogs following the launch of Google Ads, and, last but not least, a decline in sales of newspapers and printed magazines, exerted pressure on the media and their news websites. As a result, in the late 2000s and early 2010s they began adopting practices from blogging and, subsequently, social media platforms, such as the implementation of comments below their articles.

However, the news organizations could only adopt these new practices because they were following their already established practices of forum-based audience interaction. The continuity of these practices becomes evident through the intermediate steps from forums to comments, such as forums linked under an article, links to commenting subsites that do not display the comments themselves under the articles, and so forth. In this sense, media practices follow the logics of historical practices as reproductions of existing practices in new contexts (Sewell 2005; Schäfer

2016; Bourdieu 1990). Every new media practice is, in that sense, a sequel to an old media practice.

What we found most striking during our research was that initially some news organizations, such as the NYT, FAZ, and Der Spiegel, seemed keen to maintain a distinct separation between articles and comments, as if they wanted to prevent the user content from being read all too clearly as paratexts to journalistic articles. At a certain point, however, in all cases, this demarcation was breached at least once. They all made the step to treat comments as peritexts, allowing them to appear on the same web page, within the same window, and by default (rather than one click away).

Today, however, almost all of them re-introduced at least a new subtle separation. The NYT, FAZ, and Augsburger Allgemeine link on their article web pages to commenting sections, but only reveal the comments upon clicking. Technically, these comments are 'on-site' as the URL remains the same, however, they require a click to be read. In other words, the articles are presented with a commenting *section*, but without *comments*. Even the 'hardest case' of our sample, Die Zeit, the last medium displaying articles with comments, positions them only after a lengthy scroll over recommended articles and advertisements. Loosely following Reagle (see above), one might say they placed their comments on the *very bottom half* of the web page. On the other end of the spectrum, Der Spiegel built its own separate 'platform' with editor's recommendations, gamified counters, and more features akin to social media.

The websites' designs thus operate on the boundaries between paratext and discourse, and it is not the sole actor in this process. Genette refers to these actors, entitled to determine what is considered a paratext, as 'associates': "By definition, something is not a paratext unless the author or one of his associates accepts responsibility for it, although the degree of responsibility may vary" (Genette 2010, 9). By positioning the comments in close proximity to the text, the website and its associates assume a different kind of responsibility for these texts, as opposed to housing them in a forum or relocating them to a platform. This responsibility encompassed not only a spatial dimension, but also a temporal one. For instance, the NYT's commenting section, according to an FAQ answer from April 2016, is closed after 24 hours (nyt.com 2016a).

As argued earlier, many other actors accept or reject this responsibility, including persons Gillespie refers to as the "Custodians of the Internet" (Gillespie 2018), such as moderators. Journalists also negotiate this question of responsibility; some claim they avoid reading the comments due to perceived irrelevance and harm (Paßmann 2023), while others responsible for audience dialogue in newspaper companies assert that journalists typically read most of the comments (ibid.). All these material or semiotic

speech acts contribute to the ongoing discussion of whether or not, and under which conditions, "the author or one of his associates accepts responsibility". This can also be a legal issue, for example when inhumane comments are not deleted for a certain time period, raising the question of whether the author of the main text who refrains from monitoring comments can be held accountable for the content (ibid.).

The websites' design, the positioning of journalists' and users' texts in relation to one another, stands out as a strong, perhaps the most influential, non-human actor in this negotiation process. Being relatively easy to modify, websites make this paratextual perspective not only 'still useful', but arguably even *more* helpful than in the case for which it was developed: books. In the case of websites, textual orders are constantly being changed, and routinely archived by web archives. In contrast, the decisions related to the positioning of texts and their (material) periphery to one another in the mediality of books are often *black boxed* (Latour 2005), with authors and publishers negotiating elements like book titles, cover designs, and preface content. Once these decisions are made, they are reified in the material artifact of the book, potentially allowing literary scholars to reverse-engineer them afterwards. This difference is not categorical for websites, but quantitative: web archives contain a multitude of transformations of textual orders, frequently accompanied by explanations, debates, reversals, and more.

In order to navigate through the wealth of—relevant but especially also irrelevant—data provided by a web archive, the Technograph was of help in the process. In hundreds of thousands of archived web pages, it pointed to a small selection of pages with updates in their commenting sections. It did not find all of the changes in all of the websites, but it proposed points of departure, and in a next step, it helped systematize the findings and make them visible. In an iterative-cyclical process, the visualizations raised new questions, for example by showing gaps in our Historiogram or intervals of uncertainty. Some of these gaps could be closed in the next iteration. However, some intervals remained uncertain because it became evident that certain periods even of large and popular web pages are very poorly archived. In that sense, the Technograph and the Historiograms created with its help also sensitized not only in a general manner for the incompleteness of web archives, but also for the intervals of specific incompleteness, which we could consequently visualize through color gradients (Fig. 1).

6. Conclusion

Taking a web-historical perspective on the newspapers' interaction features reveals a long-ranging negotiation concerning the placement of

users' posts. Paratext theory proves invaluable in understanding this negotiation as a practice of positioning within the 'zone indécise' of texts, which may be perceived as belonging to a main text or be sorted out. This is a fruitful perspective not because we as web historians could (or should) categorize a given text as a paratext, but rather because it helps us understand the conflicts and controversies involving a diverse and heterogeneous group of actors. They engage in continuous negotiations over whether or not users' texts should be construed as paratexts.

The traces preserved in web archives constitute a rich source for this topic, because when conflicts around comments are negotiated, when commenting sections are altered, closed, or shifted to a separate subpage, the implications extend beyond the comments themselves. The cases studied might thus be understood not only as boundary work questioning whether or not and to what extent user comments belong to journalistic articles, but also as an ongoing practice on whose texts a media outlet should take care of. This seems to have changed over the recent decades.

The brief history of online comments outlined above points to a developing mutual selection process between media and their users, readers, or audience. We assume that these practices of *taking care of* differ among different news websites. The sample we analyzed does not include, for example, local newspapers, yellow press, television or radio websites, news portals, and many more. It only focuses on archived web interfaces and the question of how the different texts are positioned in relation to one another. This points to various strands of future research that have yet to be undertaken, but could draw upon the theoretical, methodical, and historiographical directions outlined in the chapter at hand.

In the cases studied here, actors that select participants are manifold: paywalls, valuation via likes, upvotes, etc., or single acts of editors recommending individual comments, are all part of a longer historical process of media outlets finding, selecting, and taking care of readership. Our research shows media outlets experimenting repeatedly, switching from forums to comments and back again, attempting to leverage the societal, economic, or democratic potential of the internet.

If, in a literal sense of the word, this appears to be a history of 're-elitarization' of public discourse, it is anything but a one-sided selection by social elites picking and choosing who may speak. Rather, it appears as a contingent, mutual, and heterogenous selection of technologies and practices by a range of actors that a history of online comments can identify. However, this does not mean that all actors involved in the process are equal or even that they possess a similar agency. Quite the contrary, a website's design is a rather powerful actor that, rather than only selecting who may speak, determines who may speak when and where. Our brief

visual history outlined in this chapter could just point to the fact that controversies about these positionings of user-generated content are ongoing from the early web until today.

What these changes in positioning mean, for example the NYT allowing user comments to appear as peritexts on the article web pages only for a brief period starting in 2016, cannot be answered only on the basis of updates in commenting sections derived from web archive data. However, it directs us to possible controversies worth studying if we want to understand the transformation of public spheres over the recent decades.

Funding

This publication is part of the subproject B03 “Historical Technography of Online Commenting” of the Collaborative Research Centre 1472 “Transformations of the Popular” (438577023), funded by the German Research Foundation (DFG).

Acknowledgement

We thank Lisa Gerzen for her helpful comments. Lisa is also a member of the subproject “Historical Technography of Online Commenting”; the chapter thus also builds upon the research she conducted within that project.

References

augsburger-allgemeine.de. 2001. “Foren. Meinungsaustausch bei Ihrer Augsburger Allgemeinen Zeitung.” December 23, 2001. http://web.archive.org/web/20011223234743/http://www.augsburger-allgemeine.de/Portal/start?pagename=index&sptnid=984043348689.

———. 2008. “550 Briefe: Wirtschaft Protestiert Gegen Stillstand Bei A-8-Ausbau.” May 2, 2008. https://web.archive.org/web/20080502222146/http://www.augsburger-allgemeine.de/Home/Nachrichten/Bayern/Artikel,-Wirtschaft-protestiert-gegen-Stillstand-bei-A-8-Ausbau-_arid,1216708_regid,2_puid,2_pageid,4289.html#kommentar.

———. 2010. “Nachrichten Augsburg Bayern.” October 31, 2010. https://web.archive.org/web/20101031130140/http://www.augsburger-allgemeine.de/Home/Nachrichten/Startseite/regid,2_puid,2_pageid,4288.html.

Böhnke, Alexander. Paratexte Des Films: Über Die Grenzen Des Filmischen Universums. transcript Verlag, 2007. https://doi.org/10.1515/9783839406076.

Bourdieu, Pierre. 1990. *The Logic of Practice*. Translated by Richard Nice. Stanford/CA: Stanford University Press.

Brügger, Niels. 2008. “The Archived Website and Website Philology.” *Nordicom Review* 29 (2): 155–75. https://doi.org/10.1515/nor-2017-0183.

———. 2018. *The Archived Web: Doing History in the Digital Age*. Cambridge (Mass.): The MIT Press.

Consalvo, Mia. 2017. “When Paratexts Become Texts: De-Centering the Game-as-Text.” *Critical Studies in Media Communication* 34 (2): 177–83. https://doi.org/10.1080/15295036.2017.1304648.

Dembeck, Till. 2007. *Texte rahmen. Grenzregionen literarischer Werke im 18. Jahrhundert (Gottsched, Wieland, Moritz, Jean Paul).* Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110975123.

de.wikipedia.org. 2024. “Frankfurter Allgemeine Zeitung.” In Wikipedia. https://de.wikipedia.org/w/index.php?title=Frankfurter_Allgemeine_Zeitung&oldid=242340528#Publizistische_Ableger.

faz.net. 1997. “Frankfurter Allgemeine Online.” 1997. http://web.archive.org/web/19970126031017/http://www.faz.de/.

———. 2001. “F.A.Z. Aktuelle Ausgabe.” 2001. http://web.archive.org/web/20010309172354/http://www.faz.de/IN/INtemplates/faznet/

default.asp?tpl=faz/overview.asp&rub=%7b7113DBFA-05D2-4BA6-94AC-23EE31EEE75B%7d.

———. 2002. “Foren.” January 24, 2002. http://web.archive.org/web/20020124115715/http://www.faz.net/IN/INtemplates/faznet/default.asp?tpl=forum/forum_overview.asp&rub=%7bEEA0DD83-6AE7-47c6-B146-75E5A5F54FB6%7d.

———. 2005. “Aktuell.” June 14, 2005. https://web.archive.org/web/20050614234631/http://www.faz.net/s/homepage.html.

———. 2011. “Die F.A.Z. Im Internet: Übersichtlich, Meinungsstark Und Diskussionsfreudig.” October 4, 2011. https://web.archive.org/web/20111004121255/http://www.faz.net/aktuell/die-f-a-z-im-internet-uebersichtlich-meinungsstark-und-diskussionsfreudig-11447692.html.

———. 2014. “Janukowitschs Macht Zerbröselt: Flucht Aus Kiew.” 2014. http://web.archive.org/web/20140529155827/http://www.faz.net/aktuell/politik/janukowitschs-macht-zerbroeselt-flucht-aus-kiew-12815733.html.

Genette, Gerard. 1997. *Palimpsests: Literature in the Second Degree*. Translated by Channa Newman and Claude Doubinsky. Lincoln: University of Nebraska Press.

———. 2010. *Paratexts: Thresholds of Interpretation*. Cambridge: Cambridge University Press.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press.

Goffman, Erving. 1986. *Frame Analysis: An Essay on the Organization of Experience*. Boston: Northeastern University Press.

Hediger, Vinzenz. 2004. “Trailer online. Der Hypertext als Paratext.” Oder: Das Internet als Vorhof des Films. In *Paratexte in Literatur, Film, Fernsehen*, edited by Klaus Kreimeier and Georg Stanitzek, 283–99. Berlin: Akademie Verlag.

Jones, Steven E., and George K. Thiruvathukal. 2012. *Codename Revolution: The Nintendo Wii Platform*. Cambridge (Mass.): The MIT Press.

Kiechle, Oliver. “Ein gespaltenes Netz? – Das Usenet der 1980er-Jahre zwischen Regulierung und Anarchie.” In *Zur Geschichte des digitalen Zeitalters*, edited by Ricky Wichum and Daniela Zetti, 125–42. Geschichte des digitalen Zeitalters. Wiesbaden: Springer Fachmedien Wiesbaden, 2022. https://doi.org/10.1007/978-3-658-34506-8_7.

latimes.com. 1996. “Los Angeles Times Speak Out.” December 23, 1996. http://web.archive.org/web/19961223084507/http://www.latimes.com:80/HOME/SPEAKOUT/.

———. 1997. “Los Angeles Times Discussions.” August 5, 1997. http://web.archive.org/web/19970805154359/http://www.latimes.com:80/HOME/DISCUSS/.

Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

McLuhan, Marshall. 1994. *Understanding Media: The Extensions of Man*. Edited by Lewis H. Lapham. Cambridge (Mass.): The MIT Press.

nyt.com. 1996. “The New York Times on the Web.” December 19, 1996. https://web.archive.org/web/19961219002950/http://www.nytimes.com/.

———. 2009. “When Bonus Contracts Can Be Broken – Room for Debate Blog.” March 21, 2009. https://web.archive.org/web/20090321143730/http://roomfordebate.blogs.nytimes.com/2009/03/17/when-bonus-contracts-can-be-broken/#comments.

———. 2010a. “Editorial – Thinking About a New Haiti.” February 6, 2010. https://web.archive.org/web/20100206054052/http://www.nytimes.com/2010/02/01/opi

nion/01mon1.html.
———. 2010b. “Thinking About a New Haiti – Readers’ Comments.” February 6, 2010. https://web.archive.org/web/20100206062634/http://community.nytimes.com/comments/www.nytimes.com/2010/02/01/opinion/01mon1.html.
———. 2016a. “Comments.” April 3, 2016. https://web.archive.org/web/20160403071334/http://www.nytimes.com/content/help/site/usercontent/usercontent.html.
———. 2016b. “Hillary Clinton Denounces the ‘Alt-Right,’ and the Alt-Right Is Thrilled.” September 2016. https://web.archive.org/web/20160901001117/http://www.nytimes.com/2016/08/27/us/politics/alt-right-reaction.html.
———. 2024. “Hillary Clinton Denounces the ‘Alt-Right,’ and the Alt-Right Is Thrilled.” January 27, 2024. https://web.archive.org/web/20240127172635/http://www.nytimes.com/2016/08/27/us/politics/alt-right-reaction.html.
Paßmann, Johannes. 2023. “Sorting Texts out (or in). Revisiting Paratexts Praxeologically.” Preprint. https://doi.org/10.13140/RG.2.2.36824.26889.
Paßmann, Johannes, and Lisa Gerzen. 2024. “Follow the Updates! Reconstructing Past Practices with Web Archive Data.” *Internet Histories*, 5 February. https://doi.org/10.1080/24701475.2024.2310405.
Paßmann, Johannes, Anne Helmond, and Robert Jansma. 2023. “From Healthy Communities to Toxic Debates: Disqus’ Changing Ideas about Comment Moderation.” *Internet Histories* 7 (1): 6–26. https://doi.org/10.1080/24701475.2022.2105123.
Reagle, Joseph M. 2015. *Reading the Comments: Likers, Haters, and Manipulators at the Bottom of the Web*. Cambridge (Mass.): The MIT Press.
Schäfer, Hilmar. 2016. “Praxis als Wiederholung. Das Denken der Iterabilität und seine Konsequenzen für die Methodologie praxeologischer Forschung.” In *Praxistheorie. Ein Soziologisches Forschungsprogramm*, edited by Hilmar Schäfer, 137–59. Bielefeld: Transcript.
Sewell, William H. Jr. 2005. *Logics of History: Social Theory and Social Transformation*. Chicago: University of Chicago Press.
spiegel.de. 1997. “Spiegel Online Forum.” June 15, 1997. https://web.archive.org/web/19970615022041/http://www.spiegel.de/forum/.
———. 2001a. “Gespräche Gescheitert: Windows XP Ohne AOL – Netzwelt – SPIEGEL ONLINE.” June 8, 2001. https://web.archive.org/web/20010608201136/http://www.spiegel.de/netzwelt/ebusiness/0,1518,137643,00.html.
———. 2001b. “‘Bianca’s Smut Shack’ – Service – SPIEGEL ONLINE.” June 9, 2001. https://web.archive.org/web/20010609122254/http://www.spiegel.de/service/0,1518,137202,00.html.
———. 2010. “Best of Engadget: Das iPhone Fällt Nicht Weit Vom Ast.” March 4, 2010. http://web.archive.org/web/20100304045925/http://www.spiegel.de/netzwelt/gadgets/0,1518,674791,00.html.
———. 2014a. “Jan Böhmermann: Porträt Vom ‘Neo Magazin’-Moderator.” January 14, 2014. https://web.archive.org/web/20140114144756/http://www.spiegel.de/kultur/tv/jan-boehmermann-portraet-vom-neo-magazin-moderator-a-939862.html.
———. 2014b. “Champions League: Atlético Madrid Gewinnt Gegen Chelsea.” August 18, 2014. https://web.archive.org/web/20140818033329/http://www.spiegel.de/sport/fussball/cha

mpions-league-atletico-madrid-gewinnt-gegen-chelsea-a-967088.html.
———. 2018. “Volker Kauder Warnt Tsipras Vor Ende Des Griechischen Sparkurses.” February 14, 2018. https://web.archive.org/web/20180214125931/http://www.spiegel.de/politik/deutschland/volker-kauder-warnt-tsipras-vor-ende-des-griechischen-sparkurses-a-1016139.html.
———. 2020. “Diskutieren, Was Ist.” January 8, 2020. https://web.archive.org/web/20200108173806/https://www.spiegel.de/backstage/diskutieren-was-ist-a-a633c185-38ef-425b-a1c6-8531f49683c5.
———. 2023. “Neue Community: Freuen Sie sich auf SPIEGEL Debatte.” Der Spiegel. November 7, 2023. https://www.spiegel.de/backstage/neue-community-freuen-sie-sich-auf-spiegel-debatte-a-4a1bae59-84a4-4d71-ab37-922e6a0e5f48.
Stanitzek, Georg. 2005. “Texts and Paratexts in Media.” *Critical Inquiry* 32 (1): 27–42. https://doi.org/10.1086/498002.
sz.de. 2003. “Sueddeutsche.De.” April 29, 2003. https://web.archive.org/web/20030429173428/http://www.sueddeutsche.de/index.php?url=/sz/kommentar&datei=index.php&email=auto-online@sueddeutsche.de&ressort=auto&urlalt=auto/service/64133&dateialt=index.php.
———. 2007. “Videoüberwachung Erste Erfolge Der Virtuellen Streife.” June 18, 2007. https://web.archive.org/web/20070618140929/http://www.sueddeutsche.de/,ra10l1/muenchen/artikel/456/108348/.
———. 2015. “Frankreich: Wehe, Wenn Der Schmerz Nachlässt.” 2015. https://web.archive.org/web/20150110205411/http://www.sueddeutsche.de/politik/frankreich-nach-dem-anschlag-auf-charlie-hebdo-wehe-wenn-der-schmerz-nachlaesst-1.2295265.
———. 2022. “Warum gestaltet die SZ ihre Kommentarfunktion um?” Süddeutsche.de. 2022. https://www.sueddeutsche.de/projekte/artikel/service/sz-transparenz-blog-e120891/.
Werber, Niels, Daniel Stein, Jörg Döring, Veronika Albrecht-Birkner, Carolin Gerlitz, Thomas Hecken, Johannes Paßmann, Jörgen Schäfer, Cornelius Schubert, and Jochen Venus. 2023. “Getting Noticed by Many: On the Transformations of the Popular.” *Arts* 12 (1): 15–36. https://doi.org/10.3390/arts12010039.
Wüllner, Daniel. 2015. “Lassen Sie uns diskutieren.” Süddeutsche.de. January 21, 2015. https://www.sueddeutsche.de/kolumne/ihre-sz-lassen-sie-uns-diskutieren-1.2095271.
zeit.de. 1999. “Debatte.” April 29, 1999. https://web.archive.org/web/19990429121439/http://www.debatte.zeit.de/.
———. 2005. “Halte Was Du Hast, Dass Niemand Deine Krone Nehme.” November 6, 2005. https://web.archive.org/web/20051106094428/http://www.zeit.de/leben/leben_hat_uns/simon_23.
———. 2006. “Sieben Für Deutschland.” July 26, 2006. https://web.archive.org/web/20060726122422/http://www.zeit.de/online/2006/22/Netzwerk_Abgeordnete.
———. 2009. “In Eigener Sache: Die Neuerungen Im Kommentarbereich.” September 9, 2009. https://web.archive.org/web/20090909142221/http://community.zeit.de/user/die-redaktion/beitrag/2009/09/05/eigener-sache-die-neuerungen-im-kommentarbereich.

Multi-level structure of the First Tuesday communities after the 2000 dot-com crash: A social network analysis of economic actors based on web archives

Quentin Lobbé

Abstract: The First Tuesday initiative began in the UK in 1998. This series of monthly meetings between IT entrepreneurs and investors played a key role in the development of the new digital economy. In this chapter, we use First Tuesday meetings as empirical proxies to analyze the social system of the economic actors who survived the 2000 dot-com crash. To this end, we delve into the raw web archives of the firsttuesday.com website in order to reconstruct the social network of First Tuesday attendees. Our analysis reveals that the First Tuesday community was, on one hand, regionally decentralized (both online and offline), but on the other hand, organized in two transnational groups of actors: the financial block and the technological block.

Keywords: first tuesday, financial web, social network analysis, web cernes, stochastic block models.

1. Introduction

The year 1995 marked a turning point in the history of the web. The initial public offering of Netscape in August 1995 heralded the beginning of the 2000 stock market bubble also known as the dot-com bubble. This event reflected the expansion of the New Economy: the idea that the internet and the web could spawn new types of business markets and achieve unprecedented returns on investment (Flichy 2001). The web thus quickly became a source of financial euphoria. The combination of high growth, low inflation, and high employment transformed investments into a gold rush. Venture capital became readily available and valuations in startups related to information and communication technologies (ICT) experienced exponential growth (Ofek and Richardson 2003). Indeed, in the late 90s, starting an online business required minimal capital, leading to a proliferation of startups across the USA and Canada before extending into Europe (Abélès 2002). Stock options further inflated the capitalization of young companies and venture capitalists liberally invested in pursuit of short-term profits. The tech market eventually reached a point of no return.

The dot-com bubble burst in March 2000, dragging the global valuation of tech markets down with it (Griffin, et al. 2011). It would take nearly ten years—and the success of Facebook—for confidence in the digital market to recover. In the wake of this crash, the digital economy had to reinvent itself by pivoting towards new business models and new areas of investment. The period from 2000 to 2004 can thus be considered as a

Quentin Lobbé, CSS Team Marc Bloch Center, France, quentin.lobbe@gmail.com, 0000-0003-2691-5615

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Quentin Lobbé, *Multi-level structure of the First Tuesday communities after the 2000 dot-com crash: A social network analysis of economic actors based on web archives*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.21, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 247-258, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

pivotal moment in web history (Lobbé 2018), marking the end of the first golden age of e-commerce and the dawn of the mobile web era. In this chapter, we aim to study this historical moment.

1.1 What historical traces remain from the post dot-com crash period?

From a historiographic perspective, the growth or decline of financial markets are typically analyzed through aggregated economic indicators such as business turnovers, recruitment dynamics, and the like (Luo and Mann 2011; Mann and Luo 2010). However, focusing solely on such indicators overlooks the underlying social interactions. Indeed, tech or financial markets are complex social worlds (Becker 2008), comprising various agents interacting with one another, including entrepreneurs, investors, and business lawyers. But given the intrinsic opacity of modern finance, accessing public historical records of these social interactions is challenging. Who was financing whom? Who was meeting with whom and under what circumstances? Who sponsored these meetings? And so forth.

Nonetheless, it was precisely as a community that the tech market redefined itself between 2000 and 2004. Therefore, the question arises: is there a way to reconstruct this social network? How can contemporary historians investigate the social history of economic actors who weathered the 2000 dot-com crash?

1.2 The First Tuesday meetings

Our research has led us to uncover the significance of the First Tuesday meetings. Emerging in the late 1990s, these offline social events played a key role in fostering communities within the burgeoning digital economy. Created in 1998 in Great Britain, First Tuesday events were monthly gatherings held in major technology hubs across the Western world (Evans 2002). On the first Tuesday evening of each month, these gatherings brought together hundreds of investors and entrepreneurs in prestigious venues such as luxury hotels, corporate headquarters, and government ministries. Renowned startup founders delivered keynote lectures, while multinational tech and finance companies sponsored the events. However, beyond the formal presentations, attending a First Tuesday event allowed entrepreneurs (identified by yellow badges) to connect with investors (identified by green badges), present their business plans, and potentially secure funding. For a few hours, these events transformed into giant ephemeral offline social networks. In the early 2000s, the concept of First Tuesday spread throughout North America and Europe through the establishment of regional and local chapters. First Tuesday events peaked in 2001–2002, gradually declining after 2003 and becoming more exclusive as born-digital professional networking platforms such as LinkedIn emerged and eventually replaced the offline First Tuesday meetings.

1.3 From archived meeting descriptions to offline social interactions

In this chapter, we propose to use the First Tuesday meetings as empirical proxies to analyze the social network of economic actors who weathered the 2000 dot-com crash. These meetings served as spaces for socialization where new relationships were created and forged among participants. Each meeting can thus be modeled as a network of encounters between economic actors: actor A met actor B during the meeting M_i such as $(A \leftrightarrow B)_{M_i}$. By aggregating the offline networks generated from each meeting, we aim to approximate the social structure of the 2000–2004 tech market.

However, as far as we know, there are no existing records of these offline meetings apart from the pre-meeting descriptions that were published on a dedicated website called firsttuesday.com. Unfortunately, firsttuesday.com disappeared from the web more than a decade ago, around 2010.

But nothing is truly lost on the web. Indeed, the automated collections carried out by the Internet Archive initiative (Kahle 1997) preserve the memory of past websites and substantial portions of firsttuesday.com were archived between 1999 and 2010. Therefore, in this chapter, we will delve into information published twenty years ago on firsttuesday.com by using raw web archive data. How can we transform raw snapshots of meeting descriptions into a viable archive of offline social interactions?

This chapter represents a methodological contribution to the field of digital humanities and will be valuable to scholars interested in extracting reliable historical sources from raw web archive materials (see section 2).

1.4 Previous work and research questions

This chapter builds upon previous web archive research conducted on the digital strategy behind the organization of the First Tuesday meetings, as documented in Lobbé 2023:

- The organization of the meetings was decentralized, with the First Tuesday initiative divided into regional (state-level) or local (city-level) chapters. Each chapter had its own dedicated local/regional website to support a common offline/online strategy.
- The main website, firsttuesday.com, served as a hub for the entire First Tuesday initiative. It advertised upcoming events from local chapters, reported on past meetings, and hosted a discussion forum.
- The focus of the meetings evolved over time, covering three main topics: e-commerce and e-business in 1999–2000, mobile web and telecoms in 2000–2002, and biotechnology after 2003.

- Economic actors who spoke during meetings were often influenced by the myth of the self-made man (Galluzzo 2023) and presented a ‘rise and fall and rise again’ narrative to motivate their audience following the dot-com crash.

Building on this foundation, our current chapter will focus on the analysis of offline interactions recorded in the firsttuesday.com web archives. How can we investigate offline interactions that are twenty years old starting from online archived traces? To what extent can we reconstruct the social system of the tech market after the 2000 dot-com crash? What was the structure of this system? Was it decentralized and organized into chapter-like communities? Were there higher levels of organization? Were there global actors? Can we say that the post dot-com crash tech market was made up of a unique community or scattered clusters of actors? etc.

1.5 Static or dynamic social network analysis?

To address these research questions, we will use a social network analysis approach called stochastic block models (SBM). The benefits of using SBM will be addressed in section 3. However, at this juncture, it is important to clarify that this method enables the study of the structure of a given social system from both static and dynamic perspectives. Unfortunately, the temporal quality of the firsttuesday.com archives fluctuates significantly, limiting our ability to conduct a consistent diachronic analysis of the period 2000–2004. Therefore, we have decided to narrow the scope of our study to the single year 2001 (from January 2001 to January 2002). This decision is based on 2001 being the busiest year in terms of meeting frequency and marking the renewal of the digital economy, as explained in subsection 1.4. Furthermore, the temporal coverage of 2001 is the highest in the entire raw corpus of firsttuesday.com web archives. Henceforth, we will consider the year 2001 as a static moment with no temporal evolution. In section 5, we will explore possibilities to extend our analysis towards a more dynamic approach.

2. Reconstructing social interactions from raw web archives

Our initial objective is to build a reliable collection of social interactions extracted from the raw web archives of firsttuesday.com. We define two economic actors as being in interaction if they participated together in at least one First Tuesday meeting. The resulting social network will be built upon this premise. To visualize the evolution of the firsttuesday.com website and filter its most relevant archived pages, we will employ the web cernes approach (Lobbé 2023). Additionally, we will use the web fragments framework (Lobbé 2018) to extract the actors from the meeting descriptions. Our data mining protocol comprises five steps:

- Visualize the temporal evolution of firsttuesday.com.

- Identify the archived pages relevant to our study.
- Extract the meeting descriptions from the filtered archives.
- Extract economic actors from the meeting descriptions.
- Reconstruct a network of interactions between economic actors.

Initially, we extract all archived pages related to the firsttuesday.com website from the Internet Archive database using the Wayback CDX Server API. We harvest a collection of 8,280 snapshots, which are then reduced to 3,670 deduplicated snapshots, representing 1,507 unique archived pages. Next, we visualize the temporal evolution of the firsttuesday.com structure between 1999 and 2010 using web cernes[1]. Figure 1[2] illustrates the evolution of firsttuesday.com over the years into sub-parts and sub-sections. Dark lines represent single pages that remained consistent over time, while pages belonging to the same section are displayed nearby. Upon examining Figure 1, we observe that all meeting descriptions are housed within three dedicated sections: the blue, green, and orange parts. These sub-sections contain 593 distinct meeting descriptions, each following a similar pattern as illustrated in Figure 2[3]:

- A title describing the main subject of the meeting
- A date (day, month, year) indicating when the meeting is scheduled to occur
- A location (building, city) indicating where the meeting will take place
- An extended abstract detailing the subject of the meeting

Among other information, the abstracts contain the names of the speakers and sponsors invited to participate in the meetings. These speakers and sponsors typically represented companies or institutions. In the subsequent analysis, we aggregate speakers and sponsors under the umbrella of their respective companies or institutions, considering them as key players within the First Tuesday social network. It is important to note that these sets of actors are merely subsets of a larger list. Twenty years ago, attendance at meetings was not necessarily reported on firsttuesday.com, and, regrettably, not all events published on the platform have been archived. Therefore, this compilation of actors serves as a reconstructed approximation of historical reality. Nevertheless, these actors likely held

[1] An interactive version of the evolving structure of firsttuesday.com can be explored at http://maps .gargantext.org/unpublished maps phylo/web archives/firsttuesday.html

[2] Figure 1: The temporal evolution of the firsttuesday.com website reconstructed from a collection of web archives by using the web cernes approach (Lobbé 2023). The website grows from the center of the figure in 1999, then splits into sub-sections. It was gradually abandoned after 2004 before being erased in 2010. The blue, green, and orange sections represent the sections where the First Tuesday meetings were announced (see: https://doi.org/10.5281/zenodo.11066424).

[3] Figure 2: An example of a First Tuesday meeting held in Riga in December 2001 (see: https://doi.org/10.5281/zenodo.11066438).

significant influence, as we can assume their positions within the social fabric of the First Tuesday initiative facilitated their invitation as speakers.

Next, as explained in subsection 1.5, we have chosen to focus our analysis solely on meetings occurring between January 2001 and January 2002, totaling 213 meetings. Subsequently, we delve into the textual content of each selected meeting, using the web fragments framework to extract dates, locations, and actors, resulting in the identification of 438 unique economic actors. These actors were manually categorized into 8 types: tech and IT companies (43%); investment, finance, and law firms (15%); press (11%); consulting firms (10%); non-tech trade companies (8%); public and governmental entities (6%); research and educational institutions (4%); and health-related companies (1%). The resulting social interaction network encompasses 438 unique actors and 3,364 unique interactions. The interactions are weighted by counting the meetings in which each pair of actors jointly participated. We denote this network G.

3. Detecting multi-level social blocks of economic actors

Our chapter now moves into computational social sciences to analyze the social network G. Network science has been instrumental for historians in reconstructing evolving networks of social groups based on time-stamped interactions (Gardin and Garelli 1961). Within these networks, communities, clusters, or blocks often emerge, representing groupings of entities sharing common interaction patterns. Detecting such local structures within larger networks can offer precise insights into the organization of an economy by illuminating real and de facto historical associations among economic actors.

The field of community detection methods can be broadly categorized into two families: descriptive methods and inferential methods (Peixoto 2021). Descriptive methods rely on context-dependent notions, such as modularity (Blondel et al. 2008), to define a reasonable division of the network into groups. While intuitive, these approaches often yield outputs open to uncertain interpretations and lack explanation. In contrast, inferential methods aim to identify latent partitions of nodes (called blocks) that are more likely to explain the network under study. These Bayesian approaches, particularly stochastic block models techniques (SBM) (Karrer and Newman 2011), originating from the field of statistical sociology in the 1970s (Lorrain and White 1971), focus on explaining structures within observed networks, making them well suited for interpreting empirical observations. SBM not only helps to understand the role of each block within the network and the mechanisms behind their genesis, but also reveals multi-level organizations in the form of meta-blocks of blocks.

In this chapter, we find the SBM approach highly relevant for testing both the decentralized and multi-level hypotheses formulated in subsection 1.4. We thus use the ‘Graph Tool’ Python library (Peixoto 2014) to detect

possible multi-level blocks within the network G. The process reveals a first level of organization consisting of 29 blocks and a second level of organization comprising 2 meta-blocks. Figure 3 illustrates the resulting network of interactions, with economic actors represented by dots. ctor types are depicted by a dedicated color: orange for tech and IT companies; blue for investment, finance, and law firms; purple for press; green for consulting firms; brown for non-tech trade companies; dark purple for public and governmental entities; dark blue for research and educational institutions; and dark green for health-related companies.

4. Results

Figure 3 provides insights into the structure of economic actors into two distinct interlocking levels of organization.

The first level is depicted by areas of strong interactions surrounded by fine dotted lines. Our initial observation is that these zones are all associated with specific geographical areas, such as major metropolises in the United States or capital cities in Europe. These areas predominantly involve local actors who tend to interact with their own communities. This local organization aligns closely with the digital strategy outlined in our previous contribution (Lobbé 2023), showing a decentralized system within the social organization of the tech market after the 2000 dot-com crash, operating both online and offline.

A second notable observation pertains to the diversity among actors within the first-level blocks. Contrary to simplistic assumptions suggesting that these groups consist solely of entrepreneurs and investors, our field data demonstrate a much more nuanced composition of these blocks. Each first-level block comprises at least three different types of actors, with an average of four types per block across the entire network. While entrepreneurs and investors play central roles, they collaborate with other actors, including those from the press and media, the public sector, and the world of education and universities, to fuel and promote local tech markets.

A recent study by Chiapello and Roth (2024) revealed similar complex social interactions at a local level while analyzing the evolution of the Impact Investing community using Twitter data. The authors draw parallels between the local structures observed in the Impact Investing community and the concept of social worlds as defined by H.S. Becker in 1982 for arts worlds (Becker 2008). According to Becker, a social world represents a collective process involving various actors whose activities are necessary for the production of works within that social structure.

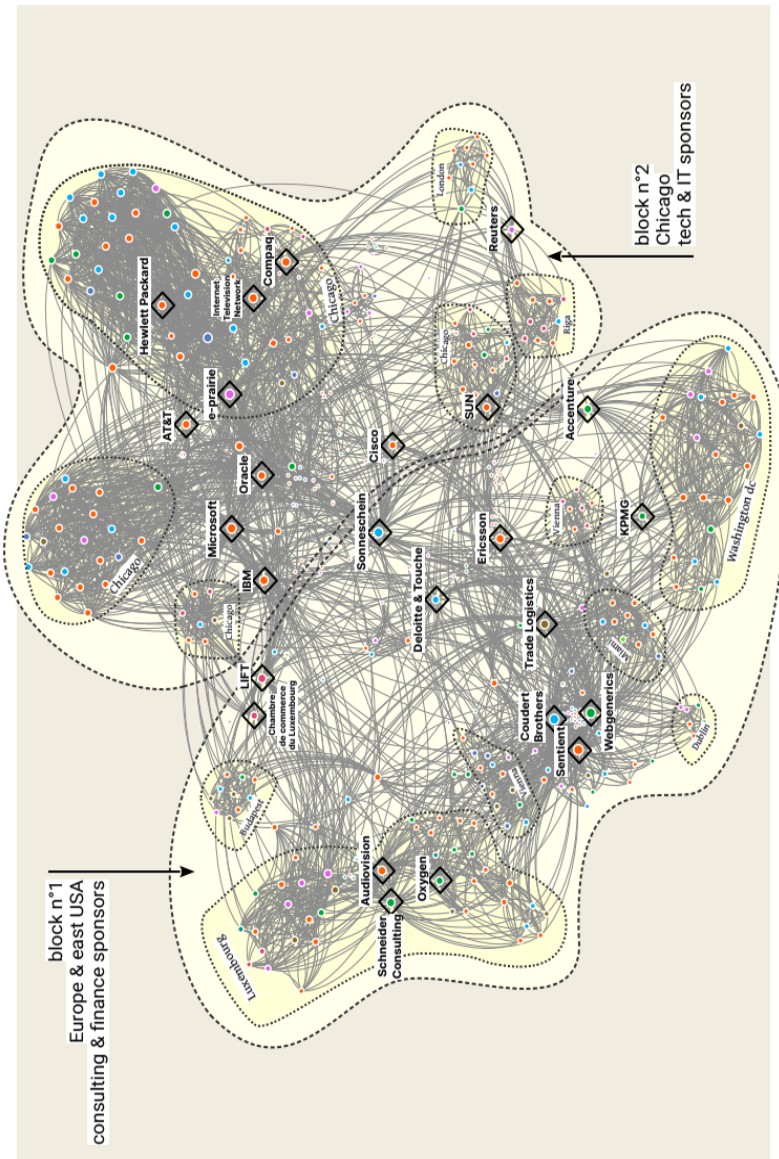


Figure 3. Network of social interactions between economic actors extracted from the descriptions of 213 First Tuesday meetings

Building upon Chiapello and Roth's insights, we view the communities participating in the First Tuesday meetings as akin to Becker's social worlds. In the social worlds of the post dot-com crash era, the circulation of business ideas, subjects, narratives, and myths was not centralized around

entrepreneurs and investors. The press and media actors also played a vital role in promoting topics and disseminating information across local sub-communities. Universities and scientific societies facilitated connections between academia and startups, often hosting meetings in prestigious institutions like MIT or Harvard. Public and state actors from local governments like the city of Riga to the EU commission promoted the local establishment of companies and bridged the gap between the financial and tech sectors. The 'bridge role' played by entities like the Chamber of Commerce of Luxembourg and the EU agency LIFT (Linking, Innovation, Finance, and Technology) in Figure 3 thus speaks for itself.

Our analysis also considers the highest level of organization depicted in Figure 3. This global level is represented by broad dotted lines, revealing that first-level blocks were contained within two larger meta-blocks structuring the First Tuesday initiative at an international level. Although we cannot definitively link these two meta-blocks to specific geographical areas, a rough approximation suggests that meta-block no. 1 centers around Europe and the east coast of the United States, while meta-block no. 2 centers around Chicago, a prominent technology hub in the USA before the rise of Silicon Valley in 2006 (Abélès 2002).

If geography is not the primary factor, then the nature of actors/sponsors connecting each local community may help explain the existence of the two meta-blocks. In Figure 3, central sponsor-actors are symbolized by diamonds, identified using the betweenness centrality measure (Brandes 2001). Block no. 1 includes sponsors from finance, law, and consultancy (e.g., Deloitte & Touche, Accenture, Coudert Brothers), while block no. 2 comprises sponsors from new digital technologies (e.g., Microsoft, IBM, Oracle). This higher level of organization reveals a dichotomy between finance and digital technologies, between entrepreneurs and investors, that we expected to observe at the local level. Based on this criterion, global meta-blocks delineated local geographical boundaries; for instance, the Washington DC community was linked to finance, while Chicago was more connected to digital technologies. Nevertheless, actors outside these realms, such as the Chamber of Commerce of Luxembourg and the EU Agency LIFT, acted as bridges between tech and finance sponsors.

5. Conclusion

This chapter has delved into the study of the social organization of economic actors who weathered the 2000 dot-com crash. Our approach followed two distinct avenues of inquiry. Firstly, within the realm of digital humanities, we curated a collection of offline social interactions extracted from the raw web archives of firsttuesday.com, enabling the reconstruction of a global social network based on descriptions of First Tuesday meetings dating back two decades. Secondly, drawing upon computational social

science, we used the stochastic block models approach to analyze the structure of this global social network. Through our analysis, we have validated the hypothesis of a tech market characterized by both online and offline decentralization. Additionally, we have studied the complexity and heterogeneity of local sub-communities, uncovering the central role played by major sponsors from finance and digital technologies in shaping higher transnational organizational levels.

However, due to the temporal limitations of the firsttuesday.com web archives, our analysis was restricted to a static review of the year 2001. To conduct a comprehensive dynamic analysis of the social history of the First Tuesday initiative, two avenues for improvement are proposed:

- Delve into the web archives of regional First Tuesday websites to supplement the list of meetings documented on the main firsttuesday.com website.
- Use the list of economic actors compiled in section 2 as a foundation for conducting interviews and accessing personal archives of actors who participated in First Tuesday meetings.

With the decentralized nature of the First Tuesday initiative established, future research should focus on understanding the unique attributes of each local chapter and community. For instance, our analysis in section 4 highlighted the distinctive role played by actors from Luxembourg, positioned at the border between meta-blocks. They seem to have acted as a permeable and global interface between the worlds of finance and digital technologies. To what extent can we refine this observation by exploring the web archives of firsttuesday.lu?

Lastly, the question of the decline of the First Tuesday communities warrants exploration. To which other platforms did these economic actors gravitate once the influence of First Tuesday meetings waned? Can we trace their “digital migrations” (Lobbé 2018) through web archives to platforms like LinkedIn or xing.com?

References

Abélès, M. 2002. *Nouveaux riches (les) : Un ethnologue dans la silicon valley*. Odile Jacob.

Becker, H. S. 2008. *Art worlds: updated and expanded*. University of California Press.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. 2008. "Fast unfolding of communities in large networks." *Journal of statistical mechanics: theory and experiment* 10, https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008.

Brandes, U. 2001. "A faster algorithm for betweenness centrality." *Journal of mathematical sociology* 25 (2), 163–177. https://doi.org/10.1080/0022250X.2001.9990249

Chiapello, E., and Roth, C. 2025. "Socio genesis of the impact investing world in France." In *Varieties of impact investing: Creating and translating a label in local contexts*.

Evans, R. 2002. "E-commerce, competitiveness and local and regional governance in greater Manchester and Merseyside: A preliminary assessment." *Urban Studies* 39 (5–6): 947–975. https://doi.org/10.1080/00420980220128390

Flichy, P. 2001. "Genèse du discours sur la nouvelle économie aux Etats-Unis." *Revue économique* 52 (7): 379–399. https://doi.org/10.3917/reco.527.0379

Galluzzo, Anthony. 2023. *Le Mythe de l'entrepreneur: Défaire l'imaginaire de La Silicon Valley*. Paris: Zones.

Gardin, J.-C., and Garelli, P. 1961. "Étude par ordinateurs des établissements assyriensen cappadoce." *Annales*, 16 (5): 837–876. https://www.persee.fr/doc/ahess_0395-2649_1961_num_16_5_420758

Griffin, J. M., Harris, J. H., Shu, T., and Topaloglu, S. 2011. "Who drove and burst the tech bubble?" *The Journal of Finance* 66 (4):1251–1290. https://ssrn.com/abstract=459803

Kahle, B. 1997. "Preserving the internet." Scientific American 276 (3): 82–83. https://www.scientificamerican.com/article/preserving-the-internet/

Karrer, B., and Newman, M. E. 2011. "Stochastic blockmodels and community structure in networks." *Physical review E*, 83 (1), https://doi.org/10.1103/PhysRevE.83.016107

Lobbé, Q. 2018. "Where the dead blogs are." In *International Conference on Asian Digital Libraries*, 112–123.

Lobbé, Quentin. 2023. "Continuity and Discontinuity in Web Archives: A Multi-Level Reconstruction of the Firsttuesday Community through Persistences, Continuity Spaces and Web Cernes." *Internet Histories* 7 (4): 354–85. https://doi.org/10.1080/24701475.2023.2254050.

Lorrain, François, and Harrison C. White. 1971. "Structural Equivalence of Individuals in Social Networks." *The Journal of Mathematical Sociology* 1 (1): 49–80. https://doi.org/10.1080/0022250X.1971.9989788.

Luo, T., and Mann, A. 2011. “Survival and growth of silicon valley high-tech businesses born in 2000.” *Monthly Lab. Rev.,* 134, 16. https://www.bls.gov/opub/mlr/2011/09/art2full.pdf

Mann, A., and Luo, T. 2010. “Crash and reboot: Silicon valley high-tech employment and wages, 2000–08.” *Monthly Lab. Rev.,* 133, 59. https://www.bls.gov/opub/mlr/2010/01/art3full.pdf

Ofek, E., and Richardson, M. 2003. “Dotcom mania: The rise and fall of internet stock prices.” *The Journal of Finance* 58 (3): 1113–1137. https://doi.org/10.1111/1540-6261.00560

Peixoto, T. P. 2014. “The graph-tool python library.” figshare. Retrieved 2014–09–10, https://doi.org/10.6084/m9.figshare.1164194.v14

Peixoto, T. P. 2021. “Descriptive vs. inferential community detection: pitfalls, myths and half-truths.” *arXiv preprint* https://doi.org/10.48550/arXiv.2112.00183

Semantic analysis of web archive historical data: 1983 “Marche pour l’égalité et contre le racisme”

Davide Rendina, Sophie Gebeil, Mathieu Génois, Patrice Bellot

Abstract: Based on a corpus composed by data obtained from the web archive of the French National Audiovisual Institute, including web pages referencing the history of the 1983 March for Equality and Against Racism, we explored how the memory of a historical event is built through the recounting of web media and the possibilities afforded by computational text analysis methods for the study of large corpuses of historical data from the archived web. This chapter presents the methodology and results of Davide Rendina's master's thesis in computer sciences under the supervision of Sophie Gebeil, Mathieu Génois, and Patrice Bellot. The objective is to demonstrate how historians can utilize archived HTML pages to study the media coverage of historical subjects on the web.

Keywords: anti-racism, media web archive, memory studies, topic modeling, 1983.

Introduction

The chapter delves into the historiographical challenges and digital humanities methodologies encountered in examining the mediatization of the March for Equality and Against Racism that took place in Paris on December 3, 1983, through the French media web archives. This event, a pivotal moment in contemporary French history, marked the emergence of second-generation post-colonial immigrants who confronted racial prejudices and demanded societal recognition. The chapter situates the research at the interdisciplinary intersection of historical studies, digital humanities, and computational sciences[1].

Initially labeled as the “Marche des beurs” by the media, the March holds significance not only as a post-colonial event but also as a manifestation of anti-racist and immigrant social movements. Its historical context is deeply rooted in the aftermath of the Algerian War and the rise of xenophobic sentiments, notably exemplified by the electoral success of the National Front. This event's complex narrative encompasses themes of identity, social justice, and political activism, making it fertile ground for

[1] This chapter is based on the Master Thesis Report: Rendina, D., Gebeil, S., Génois, M., and Bellot, P. (2024). Semantic analysis of web archive historical data the 1983 "Marche pour l'égalité et contre le racisme" [Zenodo]. https://doi.org/10.5281/zenodo.11199667

Davide Rendina, Aix-Marseille University, France, davide.rendina@gmail.com, 0009-0001-3001-8864
Sophie Gebeil, Aix-Marseille University, France, sophie.gebeil@univ-amu.fr, 0000-0002-9883-733X
Mathieu Génois, Aix-Marseille University, France, mathieu.genois@univ-amu.fr, 0000-0001-5492-8750
Patrice Bellot, Aix-Marseille University, France, patrice.bellot@univ-amu.fr, 0000-0001-8698-5055

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Davide Rendina, Sophie Gebeil, Mathieu Génois, Patrice Bellot, *Semantic analysis of web archive historical data: 1983 “Marche pour l’égalité et contre le racisme”*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.22, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 259-273, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

scholarly inquiry. The study aims to explore representations of the March in audiovisual media and on the web, with a particular focus on its semantic treatment online. This interdisciplinary research intersects fields such as memory studies, the history of representations of the past, digital humanities, and computational sciences.

The chapter delves into digital media representations of the March, particularly on the web, probing how these evolve amidst immigration debates and colonial legacies, notably the Algerian War. Employing distant reading, the first section aims to dissect online semantic treatment, uncover recurring themes, and decode underlying narratives. Methodologically, the second section integrates natural language processing methods and network analysis, facilitating systematic exploration of a vast web archive corpus from the INA (French National Audiovisual Institute). Subsequent sections detail data retrieval and methodology, addressing challenges through innovative strategies such as automatic indexing and community detection. The chapter culminates by presenting results and discussions, illuminating temporal trends, identifying entities and topics, and unveiling structural patterns. These insights offer nuanced understanding of media narratives surrounding the March, reflecting societal dynamics and political discourses over time.

1. Historiographical issues and digital humanities challenges

This interdisciplinary research intersects multiple fields: the history of representations of the past, digital humanities, and computational sciences.

Historiographically, the triumphant arrival of the March for Equality and Against Racism on December 3, 1983, in Paris is, in many respects, a significant event in contemporary France. Studying it enables a better understanding of the identity tensions that agitate present society. From a media perspective, it signifies the emergence of the second generation of post-colonial immigration, previously considered a temporary phenomenon. The press and cameras focused particularly on these children of North African immigrant workers born in large housing estates, who had become young adults denouncing racist crimes and, more broadly, the mechanisms of exclusion they faced. The Maghreb-focused lens led journalists to label this unprecedented anti-racist initiative as the “beurs’ march”, a designation imbued with colonial heritage and reductionism.

Indeed, the March is also a post-colonial event in the sense that the violence induced by the Algerian War (1954–1962) still lingers. The process of memory work only began in the 1990s, with the French state officially considering the war as a mere “law enforcement operation” until 1999 (Branche 2005, 24–44). France’s defeat in 1962 left its mark on public opinion and fueled racism towards those perceived as Algerian (Gastaut

2000). Nostalgic groups for French Algeria carried out racist attacks, such as the one in Marseille in 1973. The National Front, a xenophobic far-right party founded by Jean-Marie Le Pen in 1972, himself a former soldier who served in Algeria, began to achieve electoral success in the cantonal elections of 1982 and municipal elections of 1983.

Furthermore, the March represents a significant moment in the immigrant and anti-racist social movement. It originated in the working-class neighborhood of Minguettes in Lyon against a backdrop of tensions between the local youth and law enforcement, fueled by a surge in racist crimes in France (Hajjat 2013). Toumi Djaïdja, then 19 years old and president of the association Avenir Minguette, was injured by arbitrary police gunfire and hospitalized. This incident sparked the idea of crossing France to denounce racist violence and demand better treatment for immigrants and their children. Inspired by figures such as Gandhi (1930), Martin Luther King (1963), and the Larzac farmers (1978), the March garnered support from Father Christian Delorme and the CIMADE network (Inter-Movement Committee for Evacuees) from its inception. The first seventeen marchers gathered in Marseille on October 15, 2023, welcomed by local support committees. They journeyed across France until reaching Paris, where they were received at the Élysée Palace by President François Mitterrand, who pledged to grant a 10-year residency permit for immigrant workers (Hajjat 2013). The following year saw the establishment of the SOS Racisme association, following the lead of the Socialist Party, one of the historical anti-racism associations behind the memorable 1985 concert. The "Don't Touch My Buddy" badge marked an entire generation, but for the marchers and anti-racist activists, SOS Racisme was criticized as a political exploitation of the March, embodying the unfulfilled promises made by the Socialist Party to the inhabitants of working-class neighborhoods.

A complex and divisive event, the March remained largely overlooked in the 1990s and 2000s. It was not until its thirtieth anniversary that celebrations began to emerge, including several exhibitions and, notably, the 2013 film *La Marche* by Nabil Ben Yedir. Following an initial qualitative study highlighting the ambivalences of the memory of this event, which oscillated between nostalgia and bitterness (Gebeil 2013), we aimed to study its representations in audiovisual media and on the web within the PICCH project[2]. This involves understanding: How do representations of

[2] Polyvocal Interpretations of Contested Colonial Heritage (PICCH 2022–2024) is a European project involving five national partners and coordinated by Prof. Daniela Petrelli (Sheffield Hallam University). It explores how archival material created in a colonial mindset can be re-appropriated and re-interpreted to become an effective source for decolonization and the basis for a future inclusive society. The French team at Aix-Marseille University, coordinated by Sophie Gebeil (PI), comprising Véronique Ginouvès (archivist), Christine Mussard (historian), Pauline Savéant (Ph.D.

the March transform with the debates related to immigration in France since the early 2000s? What portrayal is given of the marchers themselves? What role do references to the colonial past play in media coverage of the March? Especially the Algerian War? These questions draw upon corpora composed of audiovisual archives, web videos, and web pages collected by the INA. In this chapter, we focus on the media coverage of the March on the web, particularly the semantic treatment of the event online. Through distant reading, we aim to understand how the event is described within online media and identify recurring names and themes associated with its depiction.

In addition to these inquiries within the fields of memory studies and the history of representations of the past (Gebeil 2021), this research also falls within the realm of digital humanities. Indeed, it raises questions about leveraging the analysis of archived HTML pages from the web to study the media coverage of a subject in web history. This implies an interdisciplinary reflection as it also addresses significant questions in the study of semantic networks and computer science.

2. Corpus and data from the INA web archive

To retrieve data related to the event, the web archive was queried for three different expressions: “Marche des Beurs” OR “Marche pour l’égalité et contre le racisme” OR “Marche de 1983”, with OR being the standard union operator. The web pages were scraped using Boilerpipe, a library designed to process HTML files and extract the main content, thereby filtering out text associated with navigation links and other extraneous elements. It is important to note that since websites are archived every time a single byte changes, the same web page may appear multiple times in the archive. Therefore, a deduplication was performed, retaining only the first (chronologically) URL of each web page. The resulting data formed a JSON file. In this project, only some of the available information from each webpage was kept and extracted in a CSV file:

- id: the unique identifier of the web page.
- url: the original URL.
- title: the HTML title.
- date: the extraction date of the web page, from which the year was used for the deduplication, which can be further used as a proxy for the publication date for a diachronic analysis of the corpora.

---

student) and Mara Bertelsen (research assistant), in partnership with the INA (Institut National de l’Audiovisuel), mobilize diverse archives to study colonial and postcolonial narratives.

- webpage_text: the text scraped by the boilerpipe algorithm from the web page.

In total, the final corpus contains 12,688 entries, whose distribution across the years is highly skewed, with 44% being from 2013[3]. This is most likely due to the 30th anniversary of the March, coinciding with the release of a movie and a documentary. The temporal heterogeneity in the corpus is important to consider, as it may affect comparisons in diachronic analysis. Additionally, the data includes 558 different domains, ranging from radio (franceinfo.fr) and television (non-stop-people.fr) websites to blogs and forums. Therefore, the data is expected to be heterogeneous in both content and form as well as in time.

3. Methodology: pipeline

The main challenge in handling such a corpus is its size (12,688 documents), which makes manual investigation impossible. Furthermore, the documents are 'raw' texts, potentially containing multiple independent sections (e.g., a blog page with several articles, a media home page with titles and excerpts, a forum page with different messages, etc.). The first step in exploring the corpus is thus to make it browsable, i.e. enabling targeted retrieval of documents related to specific questions. Given the prohibitive number of documents in the corpus for manual tagging, we tested whether Natural Language Processing (NLP) methods could facilitate automatic indexing of the documents. Specifically, we focused on two approaches: identifying entities and identifying topics.

3.1 Named Entity Recognition

Named Entity Recognition (NER) enables the identification of entities—words referring to objects that can be denoted with a proper name—in a text, and classifying them into categories (Ehrmann 2021). Among the different techniques that can be used for NER, we relied on a pre-trained deep learning model developed by Babelscape (Tedeschi 2021). Deep learning models hold significant advantages, as they are based on transformers that are able to capture semantic relationships and contextual information in texts without the need for manual feature engineering. Babelscape[4] as a further advantage of being usable on a multilingual dataset.

Though performant, automatic methods are not perfect. The model

[3] Web page distribution across the years, available online: Figure 5.1: Web page distribution across the years, Zenodo. https://doi.org/10.5281/zenodo.11203104

[4] Babelscape, https://web.archive.org/web/20240324020754/https://babelscape.com/

initially identified 127,195 unique entities, a significant portion of which proved unusable due to detection errors. Manual filtering was performed to select only readable entities (34,467 entities). This filtering process, conducted by individuals familiar with the 1983 March, also allowed to tag the entities based on their relevance to the corpus topic. Interestingly, only 2,134 entities were clearly linked to the March, illustrating the 'broad capture' of the initial query. Table 1 provides further insights, showing the number per category, as classified automatically by the Babelscape model.

Table 1. Number of usable named entities extracted per category.

	Relevant	Other
Person	781	23,155
Location	1171	3,184
Organization	73	2,857
Misc.	109	3,317

Named entities provide an initial entry point into the corpus. They enable targeted searches for documents mentioning specific individuals, locations, etc., which can then be manually analyzed. Moreover, they enable cross-searches for documents mentioning multiple entities simultaneously.

More interestingly, the analysis of the list of entities itself opens a window for studying the corpus in its globality. It provides a broader and more exhaustive list of entities related to the main subject, minimizing the risk of overlooking relevant items. Investigating why entities that were *a priori* unrelated to the March appear in the corpus can yield new insights on the subject. For example, focusing on the 20 most frequent persons appearing in the corpus. Remarkably, while political figures and actors from the movie about the March are present, the most frequent person is a journalist, presumably due to her prolific contributions to articles on the subject[5].

3.2 Topic Modeling

The second approach we used is Topic Modeling (TM). In essence, a TM algorithm identifies words that co-occur and groups them into clusters,

[5] See figure online: Figure 5.2, Frequency of top 20 PER NE extracted, p. 32, Zenodo. https://doi.org/10.5281/zenodo.11203104

which are then labeled as ‘topics’. Among the various methods available, we tested two pre-trained deep learning models: BERTopic and Top2Vec. Deep learning approaches based on large language models offer several advantages: they can capture semantic relationships between words and phrases, they do not require extensive text preprocessing tasks, and they can process raw text. More importantly, they automatically detect the number of topics within a given corpus. After testing the coherence of topics generated by both methods (Röder 2015), we focused on BERTopic, as it yielded better results, identifying 106 topics. Figure 3 shows the 20 most frequent topics.

However, these two methods have limitations regarding the size of the text they can analyze. Since text extracted from documents can sometimes be very long, these were split into non-overlapping ‘chunks’ of a maximum size of 512 tokens. Another limitation is that only one topic can be assigned to each chunk. To address the fact that a document may often mention several topics, we assigned to each webpage all identified topics from its chunks.

Table 2. Top 10 topics with their final labels.

Topic	Count	Label	Top 10 Keywords
0	2517	Islam and Muslim Culture	islam, musulmane, ramadan, islamique, mosquée, signes, communauté, école, autres, islamophobie
1	2134	Anti-Racism Activism in Marseille in 1983	égalité, racisme, 1983, octobre, jeunes, immigrés, marseille, collectif, mouvement, violences
2	1766	Music, Television and Entertainment	détails, musique, stop, 25, légales, partenaires, twitter, hanouna, réagissez, adolescence
3	1719	Documentary and Film Festivals	documentaire, films, blog, commentez, festival, caméra, production, réel, email, changer
4	1699	Youth Education	jeunesse, indifférence, éducation, générale, publicité, changé, hôtellerie, sports, trente, guides

5	1443	Democracy and Politics	démocratie, croissance, modèle, social, question, élus, travail, délégation, amendement, politiques
6	1393	Police Violence	police, article, violence, intérieur, amendement, justice, ministère, rumeur, jeunes, sécurité
7	1187	Immigration and Integration	immigration, immigrés, intégration, travailleurs, immigré, migratoires, immigre, familles, population, africains
8	1144	Anti-Semitism	juifs, antisémitisme, palestiniens, israéliens, humoriste, paix, antisémites, football, sionisme, allemands
9	1005	Commemoration of Algerian War	guerre, 1962, algérien, indépendance, histoire, mémoire, peuple, colonial, française, nationalisme

As for entities, labeling documents with their topics enables targeted exploration of those mentioning specific subjects. Cross-searches are also possible, both between topics and between topics and entities. Similar to entities, investigating the list of topics may offer new insights. For example, while the 1983 March initially addressed equal rights and racism, the most prominent topic in the corpus is Islam, which suggests a potential bias or specific presentation of the subject within the corpus[6].

3.3 Network approaches

While automatic labeling of documents with NER and TM yielded compelling results, we went one step further in exploring the corpus. We used a network-based approach to discern whether entities and topics could unveil an underlying structure. We defined four networks:

1. the **document-document network based on topics**: Each document is a node, and a link exists between two documents sharing at least one topic. The weight of a link is then the number of shared topics.
2. the **document-document network based on entities**: Each document is a node,

[6] Top 20 topics extracted, with for each the top 5 most relevant words. Figure available online: Rendina_5.7.png, Zenodo. https://doi.org/10.5281/zenodo.11203104

and a link exists between two documents with at least one common entity. The weight of a link is then the number of common entities.

3. the **topic-topic network**: Each node is a topic, and a link exists between two topics if they appear together in at least one document. The weight of a link is either the number of documents in which they co-appear, or the Pointwise Mutual Information (PMI) score, which adjusts for differing topic frequencies.
4. the **entity-entity network**: Each node is an entity, and a link exists between two entities if they appear together in at least one document. The weight of a link is the number of documents in which they co-appear.

We then used community detection algorithms to ascertain whether these networks were structured, i.e. if nodes form groups. For simplicity, we used the Louvain algorithm (Fortunato 2009).

In all four networks, we observed that their structure is far from random or uniform. Both topics and entities group themselves into communities, suggesting correlations in their occurrences within documents: some topics (respectively entities) are related to each other. In particular, topics appear to form two distinct communities[7].

Document-document networks also exhibit non-trivial community structures (see Figure 1). This indicates that shared topics or shared entities are indeed related to the existence of sub-corpora within the corpus, which may share a focus on a specific aspect of the main subject, a particular discourse, etc.

[7] Topic-topic network with PMI. Zenodo. https://doi.org/10.5281/zenodo.11203104

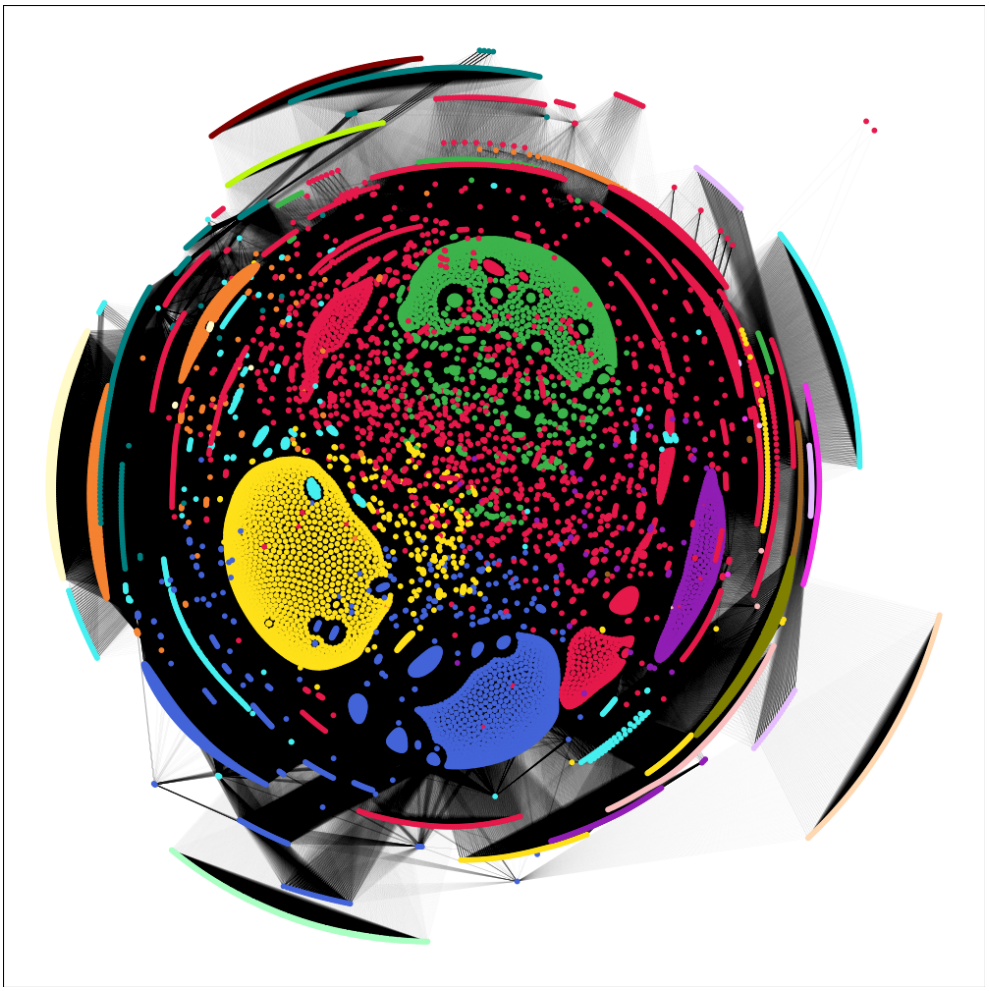


Figure 1. Document-document network based on shared topics. Colors indicate the main communities found, with red as ‘other’.

Exploring the underlying reasons for the existence of these communities can provide profound insights into the organization of discourse around the 1983 March. By using communities as additional labels for documents, we can specifically target them for analysis. The presence of topic and entity communities can be instrumental in selecting documents within the corpus associated with particular groups, and studying why and how these documents form a consistent object.

4. Results and discussion

Overall, the implemented pipeline provides several insights into the media coverage of the March through a distant reading.

The diachronic approach first illustrates the peak of media coverage in 2013 regarding the March online, corresponding to the event's 30th

anniversary (Fig. 2). While this spike in 2013 also reflects corpus bias due to the abundance of data from that year, the ability to visualize the evolution of domains by year also shows that 2009, 2010, and 2018 are peaks as well. This corroborates findings from a qualitative survey conducted in 2014, which highlighted a resurgence of interest in the March due to the publication of testimonies or the political engagement of former marchers. Comparing themes per year also reveals an intensified association with anti-Semitism in 2014, which constitutes a particularly interesting avenue for further exploration. This demonstrates the significant role of media coverage in shaping perceptions of the March and its participants, for better or for worse. Indeed, in 2014, F. Belghoul, one of the oldest marchers, launched a crusade against the ABCD of equality program initiated by the Ministry of National Education, with the support of essayist Alain Soral consistently linked with anti-Semitic views. This radical shift from anti-racism to far-right ideology is extensively covered in the press and on television. The March, once seen as a symbol of anti-racism in the 1980s, is now invoked to underscore this radical political change.

Exploring web entities provides an overview of the media narratives of the March. First and foremost, it is evident that the term “Marche des beurs” persists, overshadowing the original name of the March. However, upon closer examination of the data, the term “Marche des beurs” is often enclosed in quotation marks, suggesting a potential avenue for further exploration.

Secondly, the corpus also highlights the main stages of the March, with Paris being the most represented, followed by Marseille, the starting point, and Lyon, along with references to the Minguettes and the name of Vénissieux (Rendina et al. 2023, 54). However, while Marseille is frequently mentioned, it is uncertain whether manual archive consultation would yield substantial documentary evidence about the departure from Marseille, as it received little media attention in 1983: the fact that Marseille is prominently mentioned does not necessarily mean that the events that unfolded in the city on October 15, 1983, are as well-documented as the history of the Minguettes).

The word cloud of organizations identified as entities serves as a reminder that the March remained, even in the 2000s, a focal point for various political parties (Rendina et al. 2023, 54–55). Occurring while the Socialist Party was in power, the names of leftist parties, along with SOS Racisme, are among the most frequently cited organizations. The Green Party is also mentioned. Interestingly, far-left parties such as the LCR, the NPA (Nouveau Parti Anticaptialiste), and Lutte Ouvrière are absent from the list of most-named entities. On the right, while the UMP garners mentions, it is the National Front that stands out prominently concerning the March, possibly reflecting the far-right’s early adoption of web-based

strategies (Gimenez andVoirol 2017; Mudde 2007).

Another significant contribution of this semantic analysis lies in the identification of the key players in the media coverage of the March within the audiovisual sector. Among the prominently featured occurrences are the public group France Info (radio), France 2 (TV), and Canal+. While Canal+ is now owned by the Bolloré group, whose editorial stance aligns with conservative right-wing ideologies, marked by recurrent anti-immigrant and anti-diversity discourses, its involvement in the media coverage of the March in 2013 is not surprising. On one hand, the channel played a pivotal role in fostering comedic and anti-racist immigrant narratives, notably through personalities like Djamel Debbouze. On the other hand, journalist Maxime Musca, a member of the show "Le Petit Journal", spearheaded the "Refaire la Marche" initiative, with each stage broadcasted during daily shows until its culmination in Paris in December 2013.

Topic Modeling offers valuable insights into the thematic underpinnings surrounding the narratives of the 1983 March within the web milieu of audiovisual media in 2013. The commemorative events notably feature the promotion of Yabil Ben Yedir's film *La Marche*, starring Djamel Debbouze, as prominently evident in topics 3, 6, and 16.

Delving into textual data further confirms a pronounced focus on Islam and Muslims, reflecting the media's enduring preoccupation with the societal positioning of Muslims in France since the early 2000s. While religious demands were present among the marchers in 1983, they were not foregrounded or explicitly articulated. Rather, they symbolized a broader call for the full assimilation of post-colonial immigrant children into the national fabric, amid fervent expectations of equal rights in the face of racial violence and social marginalization. The figure of the young Maghrebi captivated attention in the 1980s, effectively reducing the March to the narrative of the 'beurs'. Three decades later, references to the offspring of the original marchers remain entwined with inquiries into the integration prospects of Muslims and their descendants, despite their longstanding French citizenship spanning multiple generations. This thematic discourse is intricately linked to the substantial presence within the reference corpus of Eric Zemmour, a right-wing provocateur convicted in 2011 for inciting racial discrimination. Zemmour's discourse, disseminated through media platforms such as "On n'est pas couché" (2006–2011) and later on RTL, pervades online discussions. Consequently, celebrations of the March are overshadowed by persistent controversies concerning French Muslims, jihadist terrorism, and the Israeli-Palestinian conflict.

Lastly, topic 9 warrants particular attention due to its alignment with the objectives of the PICCH project, which scrutinizes representations of the colonial past in contemporary media. It corroborates an associative linkage between the narratives of the March and the Algerian War, the French

defeat that culminated in Algeria's independence in 1962. This data analysis not only facilitates the identification of contentious issues but also underscores the need to delve deeper into the nuanced significance attributed to the invocation of this event, which unfolded 21 years preceding the March.

Conclusion

This interdisciplinary study, drawing upon text extracted from HTML pages covering the media portrayal of the March, sourced from a vast corpus of the INA web archives, provides valuable insights into Memory Studies, as well as broader fields such as digital history and web archive studies. By harnessing NLP techniques and network analysis, it demonstrates the efficacy of integrating these methodologies for the exploration of large historical corpora. By leveraging the capabilities of NER, topic modeling, and network analysis, we have provided digital historians with valuable tools to navigate and extract meaningful insights from vast amounts of historical data.

The study first sheds light on the ambivalences surrounding the thirtieth anniversary of the March. This milestone prompted an unprecedented surge in content production, fueled in part by the release of Yabil Ben Yedir's film and its cast. However, the mention of the actors in this march, predominantly children of the first generation of immigrants perceived as Black and especially ‘beurs’ in the 1980s, sparked myriad controversies, prejudices, and hateful discourse, illustrating the enduring prevalence of the catch-all media figure variously labeled as ‘beur’, ‘Maghrebi’, ‘Algerian’, or ‘Muslim’. Consequently, the event appears saturated, drowned in a torrent of controversies that obscure the marchers’ initial message as they attempt to share their own vision of the March. This study thus opens numerous avenues for further exploration and inquiry, which will require examining discourse related to Islam and the Algerian War within the corpus.

Furthermore, this interdisciplinary approach, combining history and computer science, also contributes to the renewal of historical methods facilitated by the use of archived web pages as inherently digital sources, beyond the case of the media coverage of the March. It offers a distinct approach to historical sources using deep learning models, commonly termed AI-driven automated tools: not only for data processing but also for source compilation and analysis. Semantic data processing is enhanced by tools such as BERT, and the developed pipeline can be replicated across any web corpus. Additionally, the work on named entities provides researchers with a repertoire of words related to the media coverage of the March, which can be used to search for other web pages or to further analyze the

corpus. This repository, available online, serves to document the event from a fresh perspective[8]. The exploration of semantic networks is also a novel approach, revealing clusters that remain to be interpreted. This reflects an important historiographical consideration for researchers working with data sourced from both the live and archived web, an area of study set to evolve within the WebLab created at the MMSH (Maison Méditerranéenne des Sciences de l'Homme) in April 2024 from a multidisciplinary perspective.

The interpretation of this distant approach stands to benefit from qualitative analysis of sources, focusing on content related to Islam and the Algerian War within the corpus. Lastly, the study's limitations prompt reflections for future research on corpora derived from archived web data, beginning with the case of the 1983 March. Efforts are underway to explore other semantic analysis tools in collaboration with the WebLab and the *CEntre de formation et de soutien aux Données de la REcherche* (CEDRE) at Aix-Marseille University[9]. Using the same corpus, comparisons between results obtained from different programs will be conducted. While diachrony remains a cherished aspect of historical research, it has been relatively overlooked here. Thus, this project marks the initial phase for expanding and refining the corpus to reproduce the pipeline on a television and web corpus spanning from 1983 to 2023, enabling an exploration of semantic network and topic evolution across successive commemorations.

[8] Named Entities Topics Analysis https://public.tableau.com/app/profile/davidearendina/viz/NE_Topics_analysis/NE_Topics_Analysis

[9] WebLab https://pba.mmsh.fr/?page_id=1465; CEDRE, https://www.univ-amu.fr/en/node/7879

References

Davide Rendina, Sophie Gebeil, Mathieu Génois, Patrice Bellot. 2023. “Semantic analysis of web archive historical data: the 1983 ‘Marche pour l'égalité et contre le racisme’.” Master Thesis. Erasmus Mundus Joint Master's Degree in Big Data Management and Analytics (BDMA). Data Analysis, Statistics and Probability [physics.data-an]. ⟨dumas-04541382⟩

Davide Rendina, Sophie Gebeil, Mathieu Génois, Patrice Bellot. 2023. “Master Thesis Report – Semantic Analysis of Web Archive Historical Data the 1983 ‘Marche pour l'égalité et contre le racisme’.” Zenodo, 10 August 2023. https://doi.org/10.5281/zenodo.10972646

De Lange, Sarah L. and Mudde Cas. 2005. “Political extremism in Europe.” *European Political Science* 4(4): 476–88. http://www.cambridge.org/9780521850810

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2024. “Named Entity Recognition and Classification on Historical Documents: A Survey.” *ACM Computing Surveys* 56 (2): 1–47. https://doi.org/10.1145/3604931.

Fortunato, S. 2009. “Community detection in graphs.” *Physics Reports*, 486 (3–5), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002.

Gimenez, Elsa, and Voirol Olivier. 2017. “Les agitateurs de la toile. L’Internet des droites extrêmes. Présentation du numéro.” *Réseaux* 202–203, no. 2–3: 9–37. https://doi.org/10.3917/res.202.0009

Pippa, Noris. 2003. “Preaching to the converted?: Pluralism, participation and party websites.” *Party Politics* 9(1): 21–45.

Röder, M., Both, A., and Hinneburg, A. 2015. “Exploring the space of topic coherence measures.” In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM 2015, Shanghai, China, February 2–6, 2015, X. Cheng, H. Li, E. Gabrilovich, and J. Tang, Eds., ACM, 399–408.

Tedeschi, S., Maiorca, V., Campolungo, N., Cecconi, F., and Navigli, R.. 2021. “Wikineural: Combined neural and knowledge-based silver data creation for multilingual NER.” In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Virtual Event / Punta Cana, Dominican Republic, 16–20 November, 2021, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2521–2533.

SECTION 6
Body and health studies in a digital context

Food, cooking, and health in a selected corpus of websites and connected YouTube channels in France. Collecting and archiving the audiovisual web

Christian Bonah, Solène Lellinger, Caroline Sala

Abstract: Based on a collaborative effort between the research project BodyCapital and the Bibliothèque Nationale de France (BnF), we present a two-step archiving process and analysis of audiovisual web content related to food and health history, investigating how audiovisuals have contributed to shaping our eating habits. The first step involved a web crawl with Heritrix, targeting 158 identified seed URLs compiled based on BnF science & technology lists and URLs identified by the research group. The crawl harvested 1,067,159 URLs. A content analysis identified 1,718 videos in our corpus. Content mapping and the identification of links to YouTube videos were performed, leading to the second step involving a focused collection of 34 YouTube channels harvesting 24,427 videos (2.4 TB) to be analyzed..

Keywords: audiovisuals, web archive videos, health, food history, YouTube.

Introduction

In her concluding chapter of the *Sage Handbook of Web History,* Jane Winters issues two calls. Firstly, the author advocates for a promising future "in which many different types of historian, not just those with an interest in contemporary politics or digital methods, can integrate web archives into their research" (Winters 2019, 596). Furthermore, Jane Winters argues that web archives and born-digital data might require or engender a radical reframing of avenues for historical research. She proposes a shift from focusing solely on textual elements of the archived web to incorporating a more comprehensive analysis of sound, still and moving images evermore present on the web. In doing so, historians "might engage with the history of art and design, media and communication studies, the history of technology, linguistics, film studies" (Winters 2019, 600).

However, historians working with web archives often experience disappointment when accessing reconstructed web pages that display the message "The video content cannot be read". This is undoubtedly one of the reasons why the above-mentioned handbook includes a chapter on the sonic web (Wage Morris 2019), but not for web audiovisuals, digital-born or not. Yet audiovisual web content has become significant since the advent of

Christian Bonah, University of Strasbourg, France, bonah@unistra.fr, 0000-0003-4756-1844
Solène Lellinger, University of Paris, France, solene.lellinger@u-paris.fr, 0000-0002-9384-7360
Caroline Sala, University of Strasbourg, France, csala@unistra.fr, 0009-0007-1730-2135

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Christian Bonah, Solène Lellinger, Caroline Sala, *Food, cooking and health in a selected corpus of websites and connected YouTube channels in France. Collecting and archiving the audiovisual web*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.24, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 277-294, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

photo-sharing tools like Flickr (in 2004) and video-sharing sites such as Vimeo (in 2004) and YouTube (in 2005).

Our research project lies at the intersection of working with web archives without being web historians, a specific interest in audiovisual content in web archives, and the intention to include web audiovisuals in a long-term audiovisual history spanning from traditional film to TV to the web. Our contribution will present a collaborative effort between the ERC research project BodyCapital at the University of Strasbourg and the Bibliothèque Nationale de France (BnF). It involves a two-step process aimed at archiving the audiovisual web related to food history and providing contextualization. Originating from health history and grappling with the methodological challenges of working from and writing through audiovisuals as central archival material, our research question seeks to understand how audiovisuals have contributed to shaping our eating habits and their connection to concerns about our individual health and healthy eating[1].

Methodological questions setting the stage: History through audiovisuals

Why audiovisuals? Exploring an entangled visual history (Lepenies 2003; Werner and Zimmermann 2004; Werner and Zimmermann 2006), our BodyCapital project investigates the impact of modern mass media, and especially visuals, throughout the twentieth century. Our focus is on understanding the role these media play in transforming our bodies into a form of capital and shaping individual receptiveness to the economization of health, reaching a point where individuals internalize the adoption of specific health practices and devices. In return, health practices and health-related goods emerge as a particularly stable and valuable vantage point from which to address twentieth-century changes in health conceptions and practices, national health policies and politics, and liberalizing market economies in Europe.

Three images from three audiovisuals in three media periods—the film *Mechanics of the brain* (Pudovkin, Russia, 1926), the TV program *Localisations cérébrales* (Jacques Rutman, France ORTF, 1964) and the internet video ego-document of American graduate student Ian Eslick (United States, 2011)—illustrate our point. Taken from a Russian educational documentary, the first image shows an experimental subject in Ivan Pavlov's research laboratory with a primitive brain wave detector, focusing on the localization and analysis of the central nervous system. The second image, from a French TV film designed for teaching, presents a

[1] To illustrate the chapter, additional visuals (screenshots, data visualization, etc.) and audiovisual materials are available at: https://medfilm.unistra.fr/wiki/Crp:Food_cooking_and_health

patient with an electroencephalography headset in a hospital where doctors study the central nervous system for diagnosis and treatment purposes. The third one, sourced from an ego-document, shows an American student facing the camera in a staged shot. The quantifying self-activist registers his brain activity using an electrical sensor headband connected to a bedside base-station alarm clock (the commercialized Zeo REM sleep monitor). These three visuals suggest avenues for historical analysis of the incorporation of commodification, hybridization, evaluation, and the internalization of body techniques and devices, into individual, autonomous health practices throughout the twentieth century. The images indicate the fluctuating trajectories of visuals across social spaces (laboratories, schools, and the public sphere), different national spaces (Russia, France, and the United States), and different media formats with distinct aesthetics (experimental film, educational film, and internet video) over time (1926, 1964, and 2011). The three visuals juxtaposed here are evidently not equivalent in their original (visual) context. The first one was projected on a large screen, the second on a small school television set, and the third can be consulted on a computer or a smartphone. The body as capital in this comparison of visual sources shifts from exploitation to transformation and then self-investment, suggesting the analytical and heuristic potential of our audiovisual archive-based approach. Beyond their integration in visual sequences in time and space, images require contextualization through non-visual materials to situate their use and utility.

We conceive audiovisuals not merely as reflections or expressions of observed phenomena, but as media forms possessing their own distinct, interactive, and performative power. We consider them essential for several key reasons. Firstly, their distribution has experienced considerable expansion; secondly they transcend professional and social groups, thirdly because of their utilitarian character, and fourthly their complementarity with economic market principles enhances their role in the promotion/communication. More explicitly, they are essential for our approach, which seeks to track and understand changes of bodily health perceptions and practices, since (1) visuals at the age of mechanical reproduction (Benjamin 1935) have become missionary communication tools extending public relations to hard-to-reach segments of society, such as illiterates, rural populations, and lower classes (see, for example, lantern shows and early cinema as fair attractions and genuinely popular culture) (Gunning 1990; Strauven 2006; Elsaesser 2009). (2) Visuals have traversed social worlds and time, evolving from clinical and laboratory settings to various forms such as health education brochures, posters, films, television programs, and video. They have even found a place in consumer goods, transforming into communication tools like X-ray images printed on cigarette packages to convey cancer risks. (3) New (audio)visual media—

from motion pictures to television and the internet—have been revolutionary technologies throughout the twentieth century. They have entertained, documented, instructed, and transformed mass audiences, similar to how moveable type transformed medieval Europe (McLuhan and Lapham 1994). Yet, visuals are more than an ideal vantage point for research; they represent conditions and conditioning that have transformed word-based health politics—akin to knowledge and scientific objectivity with and without words (Daston and Galison 2007)—into less verbal and more visual communication vectors essential to the enactment of health beliefs and practices in the twentieth-century communication society. Communication involving visuals requires methodological and analytical reflection. This involves considering how visuals, beyond their symbolic functioning from pictograms to smileys, relate to pre-existing discourse, accompanying lectures or interpretative words. Analysis of visual and non-visual material is therefore needed to account for the fact that visuals in social contexts are generally associated with or interpreted by printed, screened, or spoken words. (4) Since the interwar period, visuals have been conceived as indispensable tools for the "invisible government" (Bernays 1928), acting as an alter ego to the "invisible hand" of the market and taking the form of promotion-communication and corporate public relations.

Building on this commitment, our research project investigates how audiovisuals have changed body politics and influenced the self-perceptions and practices related to health in individuals within market-based societies in twentieth-century Europe. Following on from film and television, our contribution conveys a systematic audiovisual web analysis focusing on the subject of food studies health, eating, and nutrition.

Crawling the audiovisual web # 1: Creating a food, eating, nutrition and health web archive corpus

In collaboration with the Digital Legal Deposit Unit of the Bibliothèque nationale de France (BnF) and the Bibliothèque Nationale Universitaire de Strasbourg (BNUS), we initiated a web crawl using the Heritrix robot covering 158 identified seed URLs compiled based on BnF science and technology lists and URLs identified by the research group.[2] The crawl conducted between March 9, 2021 and March 12, 2021 harvested 1,067,159 URLs at a level n+2. A content analysis using SolrWayback identified 1,718 videos within our food, eating, nutrition, and health corpus (FENH).

[2] We are grateful to Alexandre Faye, Sara Aubry, Isabelle Degrange, and Leslie Bellony from the BnF, and Jérôme Schweitzer and Madeleine Hubert from the BNUS, for their help and guidance.

The BnF, and the BNUS in delegation, are two libraries in France that hold the responsibility of the legal deposit of the web. The rationale behind this collaboration was to benefit from their extensive experience of over ten years in collecting archives, their capacity to archive the data permanently, their proficiency in indexing and the availability of indexing tools, along with their capability to provide critical review with web archive specialists throughout the research process. Monthly work meetings covered remote access in Strasbourg, indexing advancements, problems encountered, and critical feedback from both parties. This rationale seemed especially useful for us as historians working with web archives without being web historians.

In more detail, the first crawl targeted French food-related websites, with a focus on scrutinizing representations of food on the web. This encompassed themes such as food and health, food producers and processing, cooking, agriculture, diets, and other relevant topics including forms of intermediality and audiovisuals. Seed URLs were selected from pre-existing lists of the BnF Science and Technologies Department, which has constructed a decade-long thematic collection. These were further validated by the expertise of researchers from the BodyCapital group. A few websites related to recent trends absent from the pre-existing list were added to enrich the selection. In total, 84 website homepages (primary URLs) were gathered, complemented by an additional 88 targeted tabs (secondary URLs). The complete list is recorded in the BnF curator tool called *Collecte du web* (BCWeb). A general description and keywords (topics covered, producer type, media included) were entered for each item, providing an overview of the entire selection. The crawl settings were defined as follows: Page +2, allowing the bot to explore content with a maximum depth of 50,000 URLs per website. The crawl, conducted by the BnF in March 2021, produced a relatively small but representative, controlled, and curated corpus. The uncompressed data size is 88.17 GiB and the WARC files are preserved by the BnF, with a copy delivered to the BNUS. In compliance with French legal deposit legislation, a dedicated workstation at the BNUS was made available for our use.

SolrWayback[3] software was chosen as the tool to initiate the analysis of the specifically collected FENH web archive corpus. This allowed us to quickly and efficiently index the WARC files, including images and video, and to test numerous features provided by the tool. Our first analysis involved producing comprehensive textual and visual mind maps. These were organized by themes and image typologies, highlighting the use of

[3] SolrWayback, developed by the Royal Library of Denmark, is quickly becoming an important tool for the archivist community. For further information about SolrWayback see: Anders Klindt Myrvoll's presentation of the tool: IIPC WAC 2022: SESSION 1 #2: SolrWayback at the Royal Danish Library https://youtu.be/-q4a-edVP5E?si=EmxMXrhpQ_wG4WY9.

archetypes in the representation of diets and food types. Our case study focused specifically on color and gesture. However, our video analysis encountered several challenges, including superficial content descriptions, difficulties in reading or retrieving videos for close visual analysis, struggles with diverse video file formats, and, notably, the absence of video content in the collected archives due to their hosting outside the website itself.

To address and complement the limitations of our fragmented video corpus produced through an URL-based crawl, we opted for a second crawl, once again with the BnF, but this time centered on YouTube channels initially identified by the first crawl and further complemented by lists compiled by web archiving specialists from the BnF. The supplementary YouTube crawl, the second step in our approach, was initiated 24 months after the initial URL crawl.

Crawling the audiovisual web # 2: Complementing the FENH corpus with a targeted YouTube channel crawl

YouTube hosts videos from all the content producers identified in step #1 (URL crawl), underscoring the importance of video communication for all players (companies, associations, and individuals). Following discussions with the BnF, we decided to initiate a second crawl focused on preserving YouTube channels related to the topic of food. Leveraging the BnF web archivists' experience crawling YouTube and our familiarity with the platform as researchers and users, we collected data from 34 French channels representing 24,427 videos (MP4 format: 2.4 TB) in June 2022, thus constituting our FENH video corpus. This second collection includes data for each channel retrieved from the YouTube API, stored in JSON format by the BnF. These data are also used to reconstruct a YouTube page needed for accessing the video collection through the BnF public application. A dedicated *Guided Tour* (*parcours guidé* in French) allows readers to browse all the preserved channels.

Our methodology for selecting YouTube channels resembles the approach to URL selection, combining information collected beforehand by the BnF web archivists and cross-referencing input from the BodyCapital team. Detailed channel descriptions were incorporated, allowing researchers (and the audience of the guided BnF presentation) to understand the nature of the content they encounter and why it was considered relevant to our research project. In collaboration with the BnF, we produced a detailed description of 36 harvested YouTube channels, including NGOs, national organizations, individual blogs, food processing professionals, and industrial actors. These descriptions are based on the analysis of channels and individual videos within each YouTube channel. Providing data on the

channel (name, author, content, goals) and highlighting the video's importance for the history of food and health, these descriptions also include links to websites and blogs harvested in step 1, or links to other social media platforms such as Instagram.[4] Working with video files posed major challenges, including managing data volume and file size, the need for operating media readers, and the frequent absence of connected descriptions of the file content.

The research conducted on the URL crawl, taking into account the time lapse between the two crawls, allowed us to analyze data from step #1 and to complete the second YouTube crawl, particularly addressing missing types of actors or producers and issues of under- and over-representation.

Content providers range from industrial actors to culinary chefs, (health)food-related associations to cooking brands, and private blogs to agri-food institutions. The categorization of content producers was predetermined in accordance with BnF guidelines. In a preliminary analysis of channels to produce the highest number of videos, notable channels included *Marmiton* (2,465 videos, online recipes from a cooking brand); *Ptitchef.com* (1,122 videos, online recipes); *Chef Michel Dumas* (681 videos, professional chef); *Les fruits et légumes frais* (479 videos, health and nutrition association); *Chez Jigmé* (469 videos, recipes by an individual YouTuber). The most active channels predominantly focus on recipe postings. To prevent saturation of the corpus with dominant content such as recipes, the number of providers was limited. As a result, our corpus does not claim to represent the entire food video landscape on the web. Rather, our interest lies in the variety and multiplicity of video genres and types. Instead of studying hundreds of similar recipe videos with minimal variations beyond the dish prepared, our goal is to conduct a visual analysis of short-format videos, examining their directing style, settings, camera angles, protagonists, plots, intertwined textual elements, sound, and music. Ultimately, our analysis aims to yield a better understanding of their role as contemporary mass media in the internet era and highlight their similarities and differences with traditional, prevailing audiovisual media forms like film and television. At the same time, our research targets a long-term twentieth-century analysis of food habits, cooking practices, and perceptions of healthy nutrition. This extends into the larger process of transforming the act of nourishing our bodies into investing in our body capital, generating individual receptiveness to the economization of food

[4] The YouTube channels involved in the project are described here: https://www.bnf.fr/sites/default/files/2021-03/videos_parcours.pdf. In the 'Internet Archive' consultation application, a facet has been introduced to limit the search to selections made as part of the research project. With this link: http://archivesinternet.bnf.fr/parcours/18-videos/projetrecherche/bodycapital

and eating, and internalizing changing paradigms for the adoption of healthy food consumption.

The subsequent presentation will delve into some of the results of the two-step crawl, both in quantitative and qualitative terms, and will propose analytical approaches for the analysis of the collected corpus. Furthermore, the thematic analysis will investigate a comparison between the results of the two-step crawl and the audiovisual formats and themes of a television program corpus that has been systematically collected and analyzed in parallel. The trans-media analysis will provide insights into the transformations between television and the web, particularly regarding specific thematic sub-corpuses.

Cross-sectional analysis of a two-step crawl: Content providers and opinion leaders, a case study

A first step towards analyzing our FENH corpus involved identifying highly visible content producers and their contributions. Given the overlap between the two crawls, our emphasis was on producers present in both collections. These producers were then grouped into several categories[5] (Figure 1).

[5] Aprifel – Agency for research & information on fruit & vegetables; Anses – French Agency for Food, Environmental and Occupational Health & Safety; Ameliore ta santé – Enhance your health; CrudiVegan – Raw and vegan food; Cuisine saine – Healthy food, gluten-free, lactose-free; Ptitchef.com – Recipes; Vitagora – Agri-food innovation cluster; Cultures Sucre – Association of French sugar beet growers and sugar manufacturers; Nestlé France; ANIA – French National Association of Food Industries; CERIN nutrition – Nutritional Resources and Information Center, health department of the dairy products interprofession; LaBananeInfo – Interprofessional Banana Association; Produits laitiers – Consumer-oriented information of French dairy interprofessional organization.

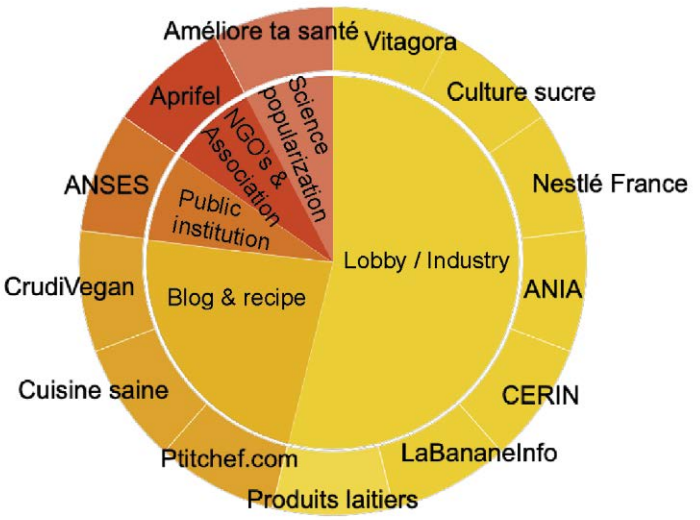


Figure 1 : Typology of content producers and their presence at the intersection of URL and YouTube channel crawls (presence in both crawls).

Five main categories of producers and contents emerge both from the collection of URLs and from the YouTube channels. Firstly, there is a group of producers from the agri-food industry, professional groups, and lobbies. Secondly, a major category revolves around cooking recipes, often originating as blogs and evolving into professional websites and media platforms. A third category concerns public institutions. A fourth category consolidates content produced by NGOs and associations, while the final category focuses on different forms of science popularization. It is important to note that this overview is not representative of the entire web, but is influenced by the collection of choices described above. Nevertheless, the overall distribution and typology indicate a substantial presence of agri-food industry players and media groups investing in food and cooking activities. Agri-food industries and nutrition/cooking web media account for 75% of the audiovisual media content in our FENH corpus. In some cases, such as *Marmiton*, the content producer began as a private website and blog for exchanging recipes and, due to its success, transformed into a professional web media entity acquired by media groups. At times, agri-food industries with longstanding commitments to audiovisual media communication expanded their activities to the new media landscape of the web in the late 2000s.

Our previous work on healthcare lobbying (Lellinger and Bonah 2021)—in particular, the commercial practices of the pharmaceutical industry—led us to select a specific case study to explore in greater detail. Our choice

concerns a producer from the agri-food category, more specifically the case of the dairy industry. One notable example is the YouTube videos produced by the French Dairy Interbranch organization (CNIEL), such as "Dairy products are our friends for life Mister V", which garnered over 7 million views within its first five years on the web.[6] Concurrently, the CNIEL has a longstanding commitment to public communication, with highly successful television spots promoting milk and dairy products dating back to the 1980s. Given our integrative long-term audiovisual approach, the "Les Produits Laitiers" (Dairy products) website and YouTube channel are of particular interest due to their extensive prehistory, including television, poster, and even film campaigns (Nourrisson 2020; Zimmermann 2021).[7] The CNIEL's public engagement with broad and specifically targeted audiences has been studied and the numerous television commercials produced since the 1980s promoting dairy products in France coined the famous and enduring slogan "Les produits laitiers sont nos amis pour la vie" (dairy products are our friends for life) (Nourrisson 2021). The CNIEL's dairy production promotion through television spots in the 1980s invested and engaged with specific audiences. Animation and pop music catered to entertainment forms specifically aimed at the younger generations. Different spots addressed nutritional questions at a more scientific level, targeting adult audiences with specific expectations regarding healthy eating. These spots functioned as a hybrid of promotion and education hyphens, showcasing the CNIEL's commitment to social responsibility for promoting milk and dairy products as integral components of a healthy diet. "Milky way didactics" incorporated incentives for "good consumerism", emphasizing nutritional value and offering advice that intertwines food education with product promotion (Nourrisson 2002). Through web archives network analysis, we can track and analyze these trends following the shift of the CNIEL's communication and promotion to the web sphere. Research into the CNIEL web organization and culture reveals an intricate constellation of websites affiliated with the dairy industry. The French Dairy Interbranch organization, CNIEL, made separate websites for each interaction, promotion, and lobbying theme structured around at least 16

[6] *Produits Laitiers* MISTER V - LES COPAINS AU LAIT - EP.1 LE LOTO, 2021-06-06-1 145 308 views. Page archived on: 22 June 2022 at 19:50 GMT in the videos collection http://archivesinternet.bnf.fr/20220622195046/http://www.youtube.com/watch?v=6A1qrXnrtWM. Accessible from BnF Internet Archives.

[7] For a comparative perspective: Tricia Close-Koenig's unpublished work on milk and dairy product promotion in Canada, the USA, and the UK. Different films are available at MedFilm: *Three for Health* (1950, Canada, https://medfilm.unistra.fr/wiki/Three_for_Health), *The Milky Way* (1939, Canada, https://medfilm.unistra.fr/wiki/The_Milky_Way) , *Grandpa's Party* (Canada, https://medfilm.unistra.fr/wiki/Grandpa%27s_Party), *The Milky Way* (1948, UK, and MedFilm analysis at: https://medfilm.unistra.fr/wiki/The_Milky_Way_UK Accessed February 5th, 2024.

related websites with content tailored to specific objectives and diverse audiences.

A more detailed network analysis of the "Produits Laitiers" website and YouTube channel's outgoing links confirms the CNIEL web sphere described above. The first 20 nodes of outgoing links from their website, generated with SolrWayback, notably highlights social media (YouTube, Facebook, Twitter, Instagram), cooking-oriented social media (Youmiam), traditional media (cheriefm.fr), public health-related websites (mangerbouger.fr, santepubliquefrance.fr), and professionally related websites of the French dairy interprofessional organization, and food industry associations. More surprisingly, the analysis also reveals connections with a museum institution (Quai Branly)[8].

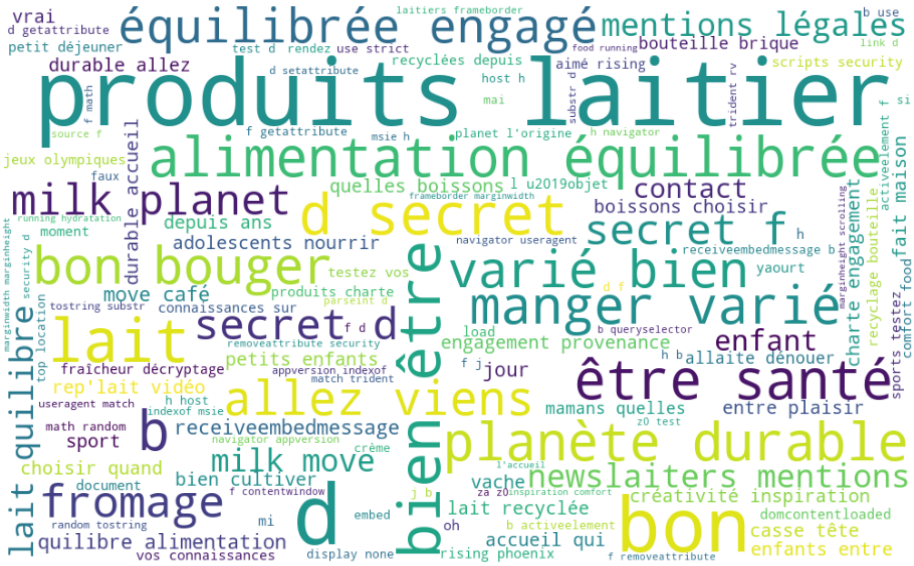
The positioning of *Produits-laitiers.com* as a significant player in food education is highlighted by its outlinks to state-controlled public health websites dedicated to health education, including *santepubliquefrance.fr* and *mangerbouger.fr*. The implied responsibility for health education aligns with dairy product promotion, turning health education into a persuasive argument for the milk industry and its web presence and publications. The *Produits-laitiers.com* website sits at the crossroads of public health issues, on the one hand, and professional promotion and lobbying on the other (Kratz 2024).

The word occurrence and association analysis based on page content (Figure 2) after cleaning our crawl data[9] produces a wordcloud illustrating that the main topic categories are, of course, dairy products (milk, cheese, etc.), but subjects such as healthy eating, well-being, eco-responsibility/ethics, and considerations for children and teenagers also feature prominently. The centrality of well-being, balanced diet, and dairy products encapsulate the website's central message.

[8] See the figure "Outgoing link analysis of *Produits-laitiers.com*. Based on the FENH corpus, SolrWayback" on the companion website:
https://medfilm.unistra.fr/wiki/Crp:Food_cooking_and_health

[9] The cleanup consisted of removing linking words, certain HTML tags, and certain terms that generated noise, such as 'Cookies'. We used JupyterNotebook for data analysis.

Figure 2: Word occurrence analysis of *Produits-laitiers.com*. Based on the FENH corpus, SolrWayback, page content, wordcloud



Beyond health references, the word content analysis of the website suggests that *Produits-laitiers.com* aims to establish a connection with the web user (*allez viens*/come on; *testez*, try it...), directed at the general public and evident through word games and direct interpellation. Geared towards promoting a positive image of dairy products, whether in terms of their effects on health (maintaining a balanced diet, for example)—with a focus on babies, children, and adolescents—or their environmental impact (in terms of the planet, sustainability, recycling, etc.) the 'milk planet' abounds in word play ranging from 'newslaiters' to 'milk move'.

From data to visual analysis

Our initial quantitative analysis of textual elements in the URL corpus was complemented by a qualitative examination of the audiovisuals in the YouTube channel corpus. The FENH corpus contains 110 videos in various formats directly associated with the promotion of the dairy industry. The channel's self-presentation emphasizes the longstanding reputation of the CNIEL's audiovisual promotional spots and highlights its recent collaboration with influential media figures, including television presenters and YouTubers. The wordplay continues with one of the entertainment series being aptly named 'Lait'Flix'[10].

[10] See http://archivesinternet.bnf.fr/20221106173359/http://www.produits-laitiers.com/laitflix/divertissement/ Accessible from BnF Internet Archives.

The YouTube channel content exhibits a diverse range of productions targeting various audiences, encompassing agriculture reports, cooking recipes, educational content, pure slapstick entertainment, and straightforward advertising. Health, education and social responsibility remain high on the *Produits-Laitiers* YouTube channel agenda. Societal concerns, such as lactose intolerance, are tackled head on. The few-minutes long video from 2014 "Lactose intolerance", which mimics a purely educational format, reassures viewers that "90-95% of the French population can digest a bowl of milk without any problem" and that everyone produces the enzymes necessary for milk digestion. Educating and promoting dairy products within the context of societal questioning is one of the key video production lines of *Produits-Laitiers*.com.

Sophisticated entertainment remains the second pivotal axis for *Produits Laitiers*. Shifting promotional content to the web, the CNIEL has significantly invested in commercial and media collaborations facilitated by social networks, engaging highly visible journalists, influencers, and web videographers. The *Produits Laitiers* YouTube channel even reintroduces older successful television concepts, playing on the nostalgia of older-generation 'webspectators'.

Illustrating this practice, the video "Les différents laits, présentés par Jamy" (Different forms of milks, presented by Jamy)[11] features Jamy Gourmand, a popular science journalist who gained recognition in the 1990s–2000s with his TV show "It's not rocket science" (*C'est pas sorcier*), broadcasted on the public television channel *France 3*. The concept of the broadcast involved duplexing from his science truck, with journalists Frédéric Courant and Sabine Quindou, explaining science to children. In 2020, Jamy Gourmand established his own YouTube channel, *Epicurieux,* dedicated to science popularization, currently boasting over 1.7 million subscribers. In October 2021, the renowned journalist produced a video for the *Produits Laitiers* YouTube channel, explaining the different forms of milk, leveraging both his past TV presentations and his recent involvement in YouTube video production.

The CNIEL website ecosystem finds its audiovisual translation in an imitation of the reputed Netflix video streaming platform, aptly named "100% milk videos". Playing on words and slogans once again, Lait'Flix *X* hosts dairy promotion videos grouped under headings like entertainment, nutrition & health, agriculture and animal well-being, or sports activities and milk. Lait'Flix incorporates humorous videos from YouTubers, such as

[11] *Produits Laitiers : Les différents laits, présentés par Jamy*, 2021-10-28-363 242 views, Page archived on 22 juin 2022 at 19:50 GMT in the videos collection http://archivesinternet.bnf.fr/20220622195058/http://www.youtube.com/watch?v=6nTPXX72JkQ Accessible from BnF Internet Archives.

one featuring the YouTuber *Mister V,* who boasts 6.3 million subscribers, or offers cooking recipes centered around… dairy products.

Back to the web archives at the BnF

The primary and summary analysis of our two-step audiovisual web crawl and the CNIEL case study suggests several directions for further investigation. On the one hand, we could broaden and deepen the analysis in terms of the collection scope whether spatial and/or temporal. It appears beneficial to study the interconnections between the various CNIEL sites in greater detail, investigating audience-specific translations operating from one *Produits Laitiers* website to another. This would facilitate the analysis of the thematic environment specific to each information and communication space.

As demonstrated by the analysis of outgoing links, we could, by extension, study the textual and especially audiovisual contents of sites to which the *produits-laitiers.com* site links, beyond the related CNIEL sites.

Another avenue of exploration is of a more 'archeological' nature. BnF legal deposit web archive collections allow us to go back in time. The historical depth of BnF website archives, collected over many years, permits consultation of archives from 2006. At that time, *Produits Laitiers* had already initiated its web presence and featured the "le Blog des Produits Laitiers".

Original, sophisticated, and fostering consumer bonding, the *Produits Laitiers* website had been offering Flash content, including games for both children and adults, since 2010. While this type of content has now disappeared from the current web, it underscores the ongoing commitment of the CNIEL to pioneering promotional practices in the most recent media spaces.

Pitched to teenagers and school children, the "three milk powers (growth, strength, and balance)" present challenges to be met. 3D films, quizzes, interactive games, and puzzles captivate first-generation web users. Once again, entertainment and health education intertwine when the headline announces a "challenge in between sensory pleasure and health".

Web sphere ecologies can be traced over time through outgoing link analysis. Studying referrals from the *Produits-laitiers.com* website to other sites in 2015, for example, indicates referrals made to the Parisian spring attraction *Salon de l'Agriculture*, an annual agriculture show in Paris. This referral also features the 2021 outgoing link analysis presented above.

The historical depth and interconnections explored in this way involve examining audiovisual content and beyond to understand how it adapted to

the communications challenges and expectations of different eras and the 'trends' of the moment, such as blogging, flash games[12], and streaming.

For the historian, our case study highlights the ability to move back and forth between different types of archives. It allows historians to link them together, extending beyond the simple web archive consultation, thereby reconstructing long-term dynamics in promotional-educational practices that condition consumer behavior at a lifelong generational level. The methodological choice of using a public institution for web legal deposit not only guarantees the continuity of the collection, but also empowers historians to navigate between various archival collections, be they web archives or archives of prior audiovisual media.

Conclusion

Our two-step web crawl-based approach, assembling a food, eating, nutrition, and health web video corpus (FENH corpus), represents an effort to systematically analyze audiovisual sources on the web and their archives. This approach aims to overcome challenges posed by the often-difficult accessibility and analysis of audiovisual web content. Collaborating with the BnF as a legal deposit institution not only ensures the continuity of data hosting, but also provides opportunities to (1) transition from data to visual studies of archived material; (2) investigate genealogies of audiovisuals on the web and beyond-before; (3) shift from outgoing link analysis to relational audiovisual content analysis; and eventually (4) address ephemeral audiovisual web content**.**

The *Produits Laitiers* case study inevitably prompts the question of whether the case represents more than itself. Given its extensive size and distribution of the French web, coupled with the fact that 75% of audiovisual media content in our FENH corpus originates from agri-food industries and nutrition and cooking web media, the analysis raises the question of whether the commercial web has significantly marginalized public actors, NGOs, and associations, and if so, when did this occur? Was the utopian egalitarian vision of the early web merely a chimera from the beginning, as the *Produits Laitiers* analysis may suggest? In any case, the personalized content approach of web promotion catering to individual users has led to a multiplication of audience-specific targeting. Interactive strategies, involving forms to be filled out for inquiries, advice, or exchanges, point in the same direction. This direction becomes evident in our wordcloud analysis, which names specific audiences (mothers, teenagers, user match, etc.)

[12] See, for example, this site, which offers flash games and a challenge for schools. http://archivesinternet.bnf.fr/20100422153616/http://www.les3pouvoirslaitierstedefient.com

In our broader research goal to integrate web audiovisual productions into a century-long history written through successive and overlapping audiovisual media forms, our analysis positions web audiovisuals as a third-era source for long-term studies of the role of audiovisuals in shaping our eating habits. A transmedial and intermedial approach opens up new avenues for *long-durée* research based on cross-analysis of audiovisual formats from archives. These formats portray and question continuities and changes in dairy promotion over a century. From *The Milky Way* (1948), a central galaxy of food products in Western modernity, agriculture, and their promotion to *Milk products are our friends for life* (1981), agricultural lobbying and food industries have consistently invested in their product promotion. Perhaps part of their success lies precisely in their generational long-term and life-long promotional-educational impact.

Milk films and videos have spanned the 20th century, evolving from depictions of milk processing and distribution (*Milky Way*, UK, 1948) to portraying milk as a lifestyle dancing companion (*Milk products our friends for life*, TV spot, 1981), and eventually to comical media self-reflection and consumer persiflage (*Lait'Flix: Mister V*, web video, 2021/2019). From school lunch campaigns in the 1950s to critical appraisals in the 2010s, the image of the milk and dairy products has undergone significant changes and challenges (Souccar 2008). However, milk promotional campaigns have endured throughout the century, transitioning from arguments about reconstructing the nation after WWII to promoting personal nutritional emancipation and self-enhancement through milk consumption. Adopting recent video and YouTube trends, web archives reveal the continued engagement of milk promotional efforts in the third-media era, adapting to changes in media and society.

References

Benjamin, Walter. 1935. *The Work of Art in the Age of Mechanical Reproduction*. https://web.mit.edu/allanmc/www/benjamin.pdf.

Bernays, Edward. 1928. *Propaganda*. New York: Horace Liveright.

Daston, Lorraine, and Peter Louis Galison. 2007. *Objectivity*. New York: Zone Books.

Elsaesser, Thomas. 2009. "Archives and Archaeologies. The Place of Non-Fiction Film in Contemporary Media." In *Films That Work: Industrial Film and the Productivity of Media*, edited by Vinzenz Hediger and Patrick Vonderau, 19–34. Film Culture in Transition. Amsterdam: Amsterdam University Press.

Gunning, Tom. 1990. "The Cinema of Attractions. Early Film, Its Spectator and Avant-Garde." In *Early Cinema: Space, Frame, Narrative*, edited by Thomas Elsaesser and Adam Barker, 56–63. London: BFI Pub.

Kratz, Amélie. 2024. "Au-Delà de l'éducation Nutritionnelle : L'éducation Au Manger Des Enfants à Travers Les Audiovisuels En France et En République Fédérale d'Allemagne (1950–1980)." PhD thesis, University of Strasbourg.

Lellinger, Solène, and Christian Bonah. 2021. "'This Corporation Has "Anesthetized" the Actors in the Drug Chain'. Influence Peddling and the Normality of Conflicts of Interest in the Mediator® Scandal." In *Conflict of Interest and Medicine*, 181–200. London: Routledge.

Lepenies, Wolf, ed. 2003. *Entangled Histories and Negotiated Universals: Centers and Peripheries in a Changing World*. Frankfurt: Campus Verlag.

McLuhan, Marshall, and Lewis Henry Lapham. 1994. *Understanding Media the Extensions of Man*. Cambridge (Mass.): The MIT Press.

Nourrisson, Didier, ed. 2002. *A votre santé !: éducation et santé sous la IVe République*. Saint-Etienne, France: Publications de l'Université de Saint-Etienne.

Nourrisson, Didier. 2020. "Faire-savoir et savoir-faire à l'École : l'alimentation à l'écran." *Les Enjeux de l'information et de la communication* 20/3A (S1): 117–130. doi:https://doi.org/10.3917/enic.hs9.0117.

Nourrisson, Didier. 2021. *Du lait et des hommes: histoire d'un breuvage nourricier de la Renaissance à nos jours*. Paris: Éditions Vendémiaire.

Souccar, Thierry. 2008. *Lait, mensonges et propagande*. Vergèze: T. Souccar.

Strauven, Wanda, ed. 2006. *The Cinema of Attractions Reloaded. Film Culture in Transition*. Amsterdam: Amsterdam University Press.

Wage Morris, Jeremy. 2019. "Hearing the Past: The Sonic Web from MIDI to Music Streaming." In *The SAGE Handbook of Web History*, edited by Niels Brügger and Ian

Milligan. London; Thousand Oaks, California: SAGE Publications.
Werner, Michael, and Bénédicte Zimmermann, eds. 2004. *De la comparaison à l'histoire croisée*. Paris: Seuil.
Werner, Michael, and Bénédicte Zimmermann. 2006. "Beyond Comparison: Histoire Croisée and the Challenge of Reflexivity." *History and Theory* 45 (1): 30–50.
Winters, Jane. 2019. "Web Archives and (Digital) History: A Troubled Past and a Promising Future?" In *The SAGE Handbook of Web History*, edited by Niels Brügger and Ian Milligan. London; Thousand Oaks, California: SAGE Publications.
Zimmermann, Yvonne. 2021. "Early Cinema, Process Films, and Screen Advertising." In *Advertising and the Transformation of Screen Cultures*, edited by Bo Florin, Patrick Vonderau, and Yvonne Zimmermann, 21–48. Amsterdam: Amsterdam University Press.

We're all experts now? Archiving public health discourse in the UK Web Archive

Alice Austin, Leontien Talboom

Abstract: Emerging from COVID-19 collecting initiatives that underscored the fragility of online health discourse, the Archive of Tomorrow was an ambitious collaborative project that set out to curate a representative and diverse collection of public health websites in the UK. The project encountered a number of challenges, such as technical barriers in capturing interactive and dynamic sites, ethical considerations concerning how disputed or outdated information might be responsibly made available to researchers, and philosophical questions about how 'health information' is to be defined. This chapter reports on the outcomes of the project and discusses future directions for improving the production and use of large-scale archived web collections.

Keywords: collection development, metadata, legal deposit, health information, misinformation.

As other chapters in this volume have reflected, the Covid-19 pandemic catalyzed a rising effort to archive information from the web, with libraries and archives rushing to document the traces of a 'new normal' that saw life for many move online. The speed at which information entered the public realm and was subsequently discussed, debated, and debunked served to crystalize some of the key issues that heritage organizations encounter when trying to capture a nebulous and rapidly-evolving information landscape: How do we capture history when it is still happening? How do we respectfully and responsibly reflect the dissent and divisions in a moment without a single, unifying narrative? And what preparations can we make now to meet the needs of the researchers of the future?

In an attempt to document the 'unprecedented' and historic events that unfolded at the start of the decade, numerous institutions and community groups turned to web archiving as a means of ensuring collecting could continue remotely. As Amanda Greenwood (2022) has detailed in a thorough literature review, the scope of these initiatives varied greatly—from global to local and community to individual—serving to reflect the myriad ways in which the pandemic's impacts were felt (Greenwood 2022). The potential (and limitations) of the web and (by extension) web archiving as a means of producing a historical record that serves as both correlative and counterpart to institutional and state narratives has long been recognized in the literature (Milligan, Ruest, and Lin 2016; Barrowcliffe 2021) and many curators experimented with participatory archiving practices to this end, inviting contributions of personal narratives, reflections, and experiences in a manner that Kerrie M. Davies has termed

Alice Austin, University of Edinburgh, Scotland, United Kingdom, alice.austin@ed.ac.uk, 0009-0007-5586-2571
Leontien Talboom, University of University of Cambridge, United Kingdom, lkt39@cam.ac.uk, 0000-0001-7408-5471

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Alice Austin, Leontien Talboom, *We're all experts now? Archiving public health discourse in the UK Web Archive*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.25, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 295-307, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

'crowd-coaxing' (Davies 2023). As Tizian Zumthurm and Stefan Krebs (2022) have reflected, however, creating a web archive that comprehensively represents the divergences of experience in a moment like the pandemic is a complex endeavor, as underscored by recent efforts to explore how web archives are understood and used in online discourse (Odgen, Summers, and Walker 2021) and even deployed and weaponized in the service of misinformation (Acker and Chaiet 2020). Such work has prompted greater consideration on the part of archivists, curators, and other memory workers as to the role of web archive collections in an ever-evolving information nexus.

Evolving from these attempts to document the pandemic and its attendant 'infodemic', the Archive of Tomorrow (AoT) was an ambitious pilot project that sought to build a test-bed collection in the UK Web Archive (UKWA) through which such questions could be explored. Led by the National Library of Scotland in partnership with Cambridge University Library, Bodleian Libraries Oxford, and the University of Edinburgh, the project was funded by a Wellcome Research Resources Award in Humanities and Social Science and ran from February 2022 to April 2023. The technical team comprised three web archivists (each based within a university library), a project manager, a metadata analyst, and a rights officer (all based within the National Library of Scotland). A research data engineer was initially engaged for the project but resigned from the post at an early date. Throughout the project the team benefitted from the support and expertise of colleagues at the British Library, and was guided by an Advisory Board comprising academic researchers and industry experts from a diverse range of relevant disciplines.

The project aims were both tangible and exploratory. The primary goal was to curate a 'research-ready' collection of websites within the UKWA around the theme of Health Information and Misinformation. This curated collection would then serve as a test-bed around which options for metadata, computational analysis, ethics, and rights issues could be explored. The team further aimed to build a network of researchers across relevant disciplines in order to involve potential users in the process of building, evaluating, and using collections. It was anticipated that the project would serve to concretize recommendations for future work and provide a focus for advocacy for change to make web archives more representative, inclusive, and open for health research. After setting out the legal and technical contexts in which the project operated, this chapter will outline the processes and deliberations involved in the production of a large-scale web archive collection, describe the challenges encountered when trying to capture such a hotly debated area, and outline areas for future work.

2. Capture

2.1 Legislative background of the UK Web Archive

The UKWA is a partnership of the six Legal Deposit Libraries (LDLs) that performs the web function of the LDLs' legislative responsibility to collect and preserve a copy of all material published in the UK and Ireland. The UKWA has been systematically collecting non-print material since 2013, with the majority of material being captured through an annual domain crawl that attempts to make a copy of any content published to a website with a recognizable UK top-level domain (e.g., .uk, .scot), or hosted on a server physically located in the UK (identified via a GeoIP lookup). The yearly crawl is supplemented by curated collecting, achieved by manually adding targets to the Annotation and Curation Tool (W3ACT), a web-based interface that allows a user to create an entry for a specified URL, establish parameters such as depth or frequency of a crawl and record metadata for description and rights-management purposes. Access to content archived in the UKWA is by default restricted to users at computer terminals onsite in Legal Deposit Libraries, unless open access permission has been explicitly granted by the website owner. The project's dedicated Rights Officer was responsible for corresponding with site owners and issuing these permission requests. If significant concerns about making a site's content accessible were identified, the Rights Officer may choose not to pursue access permission.

2.2 Technical specification

The UKWA uses Heritrix, an open-source web crawler written in Java to perform crawls. Crawled content is written to a WARC file and stored alongside metadata necessary for interpreting the crawl. There are technical and legislative limitations to what the UKWA is able to capture. For example, dynamic pages (pages in which user interaction initiates server-side scripts) present difficulties as the crawler is unable to perform any such interactions—and so any content retrieved by querying a database cannot be captured. Material that requires login credentials to access is excluded on two fronts, as the crawler cannot input such credentials and the legislation that enables non-print legal deposit only covers material that is made publicly available. The legislation also does not extend to broadcast material that is primarily audio-visual, excluding content on video-centric platforms such as YouTube or TikTok.

2.3 Scope and focus

Establishing the boundaries and scope of any collecting activity is always a challenge, and the nature of the subject in question only compounded this

difficulty. The scoping process was a continuous one. Initial efforts used the Collection Development Framework created by the Web Archiving Team at the University of North Texas. This comprehensive document was particularly useful in encouraging an holistic view of the collection throughout the curation lifecycle, and was invaluable in guiding discussions about the kinds of material that it was anticipated could be encountered during the project and the potential issues we might have to navigate.

Health is an exceedingly broad term and area of study, and the project team were keen to ensure that collecting did not focus solely on the biomedical but also reflected how debates around physical, mental, and social wellbeing intersect with other issues, such as those of politics, economics, and technology. Early in the project web archivists met with academics and students at their respective institutions to gauge areas of interest for current health-related research, and surveyed published research in these areas in order to consider how existing studies might be extended into the digital realm. It was initially anticipated that the project would focus predominantly on the issues of misinformation and disinformation that had accompanied the Covid-19 pandemic, however as the project unfolded it became clear that that was just one aspect of a much larger picture of how the internet is used to find, share, and debate issues of health. One particular observation to emerge from these conversations concerned the value of social media data for research; however this also emphasized the challenges of using archived social media content for research within the confines of the legal deposit legislation.

2.4 Identification of material

A number of different methods were employed for the identification of material, both systematic and serendipitous. Existing directories (for example, a list of health-related charities exported from the Charity Commission Register) were used to locate content, and participatory collecting methods—engaging with health researchers to determine particular areas of interest—was also a valuable approach. Web Archivists also experimented with the use of various search engines to explore how the results differed, and noted that it was challenging to record the impacts of such tools on the resulting resources that were found as many of these impacts were obscured through algorithms designed to promote and suppress particular types of content.

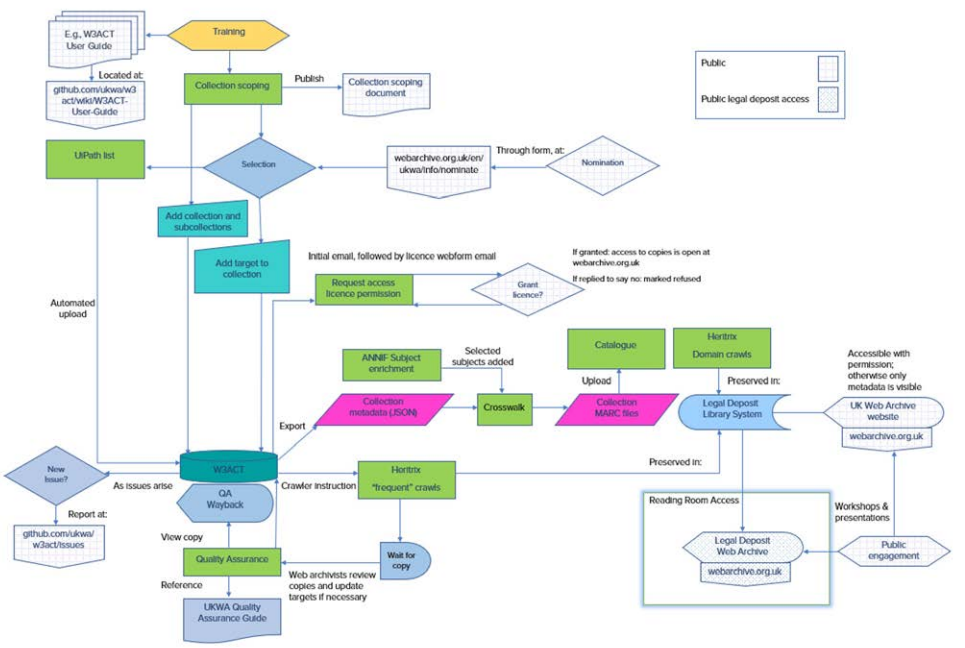
The collection of social media was an important objective for the project. Web archiving technologies allow for greater representation of ‘everyday’ or ‘street-level’ opinions through collection of social media resources, and collecting from platforms such as Twitter, Reddit, and Mumsnet provided a balance to the commercial algorithms that tend to promote commercial

enterprises. Again, there are technical and legislative limits to what can be captured from these platforms, and in order to comply with non-print legal deposit legislation, selection had to be meticulous: for example, search terms for specific topics on Twitter were combined with location tags (e.g. '(#wellbeing) near:edinburgh'), and collection from Reddit was limited to forums that explicitly located themselves in the UK (r/UKNurses, r/UKMCPatientCommittee/, r/UKantilockdown).

2.5 Workflow

The illustration below provides an overview of the team's workflow. The Web Archivists selected sites for inclusion, and records were created in W3ACT for each target URL either manually (URL-by-URL) or by means of a UiPath sheet for bulk creation. Each targeted URL was recorded in a shared-access spreadsheet, which also served as a mechanism for the Web Archivists to flag sites with potential access concerns to the Rights Officer. After an initial crawl had been performed the site was checked for quality, however as the collection grew the available capacity for quality checking diminished.

Figure 1: Project workflow



3. Collection overview

The resulting collection—named "Talking About Health" (TAH)—comprises around 3,500 individual targets. The TAH collection is subdivided along various lines through the use of W3ACT's tagging feature, whereby a target can be 'tagged' or marked for inclusion into any number of different collections or subcollections. As collecting progressed a number of areas that warranted a deeper level of collection emerged, and the tagging function has been used to group targets on focused areas such as dental health, substance abuse, menopause, and nutrition.

As a matter of course, sites were tagged into the main collection at the highest level appropriate, but in some cases more specific pages have also been targeted and tagged into lower-level sub collections. For example, the British Heart Foundation's website[1] has been included in the main collection, and in addition, a child page dedicated to 'heart-healthy recipes' has been tagged into the Nutrition sub-collection. This provides a more direct entry point to the specific content of that page, but also ensures the wider context of the site as a whole is preserved.

4. Access

4.1 Structure

Navigation of the collection is achieved through W3ACT's tagging feature which, as described above, provides some high-level description of targets. The project team was keenly aware of the narrative power of archival description and the potential for the terms used to convey a judgment of value or otherwise reflect curatorial bias. As collecting progressed it became apparent that the initial framework developed by the project team was not sufficient for representing the breadth of information and subjects covered, and that flexibility in the collecting framework was needed to allow current and future users to add areas or topics of interest beyond project completion to help ensure the collection remains relevant over time.

As the group experimented with different structures and descriptive frameworks it was interesting to observe how the diverse discipline backgrounds of the team members influenced these discussions: those who worked primarily in a library setting were inclined towards using tags to provide a description of the content and focus of a site, whereas those colleagues operating in a more archival context generally attempted to use terms to describe the creator as best as possible. As none of the team have a medical background and are therefore unqualified to make such

[1] https://web.archive.org/web/20240515150135/https://www.bhf.org.uk/

assessments, a decision was made to avoid the use of any labels that could be read as a judgment on the efficacy or suitability of a practice or its proponents. This appeared to be a straightforward approach in the context of medical information, but became a more complex challenge as the foci moved towards areas where medicine intersects with the legal, social, and discursive contexts in which it operates. The team combined these approaches to develop a multifaceted structure that provides a broad descriptive overview of the content, but also invites critical deliberation as to the circumstances and context of a site as a publication.

This process was significantly informed by the efforts of the Rights Officer to gather feedback from site owners and publishers on the hesitations, concerns, and barriers that held them back from providing access permission. A number of site owners requested information on how their site would be categorized within the collection, and voiced concern that their site might be misrepresented. The initial name of the project ('Health Information and Misinformation') was found to be causing particular consternation, and the collection name was subsequently changed to 'Talking About Health' in order to reflect the discursive nature of the content. Such insights into the concerns of site owners and creators were also influential in prompting the project team to consider the responsibilities that the UKWA has as a 'secondary publisher' of material, and how those responsibilities extend to both content creators and potential users. With regard to content creators, it is important that we fully consider the extent to which archiving practices are understood and anticipated by content creators, and what recourse those who find their content included in an archive collection have to challenge the preservation and republication of material about them. Such considerations are necessary for any kind of collecting from social media, but the team felt this was particularly crucial given the sensitive nature of much content about health, and the probability that content could have been made at a moment of crisis on the part of the poster.

4.2 Rights

The most significant factor affecting access to content in the UKWA is the legal deposit legislation that enables capture: a site owner must provide explicit approval in order for archived copies of their site to be viewed from outside legal deposit library premises. Standard procedure is for this permission to be requested through an automated email sent via the curation tool but a large proportion of these requests usually go unheeded, with the result that less than 1% of UKWA content was accessible offsite at the outset of the AoT project.

In the course of the project the dedicated Rights Officer issued offsite

access permission requests to 1,840 email addresses against 58% of the targets collected. Permission was granted for 7% of those requests and refused for 3%. The Rights Officer reported that reasons for refusing access permission were varied: some website owners saw no benefit of granting permission, and some creators were concerned that archiving itself may ultimately be damaging, either in terms of increased reputational risk from information remaining in publicly accessible online spaces beyond the original intention, risk to data subjects or risk of copyright infringement, or in terms of diverting website traffic to the archived resource and away from a live site.

At the close of the AoT project, 21% of the collected targets were accessible from outwith a legal deposit library, with the table below providing an overview of the levels of open access achieved across the subcollections:

Table 1. Breakdown of license status by category.

Heading	Remote access permission granted (%)
Blogs & social media	10.6
Charities & non-profit organizations	44.3
Commercial/industrial health sector	6.7
Focus	20.2
Government	58.6
Health professions	20.6
NHS	90.2
News & commentary	9.5
Politics & health	33.7
Research	23

While this is a vast improvement on the UKWA average, it still leaves a significant proportion of the collection inaccessible to the majority of researchers. In order to explore options for improving access a review of existing license agreements and approaches was conducted, and it was noted that the non-print legal deposit framework already recognized and respected where content had been published under an Open Government License. A paper was submitted proposing that a similar approach be adopted for content published under other open licenses such as Creative Commons licenses. This proposal was approved by the Legal Deposit Libraries Committee, and implementation is now underway.

4.3 Metadata

A key goal of the project was to investigate making use of the collection metadata as a means of ‘surrogate’ access to closed captures: that is, while full computational access to captured content is not possible within the framework of the legal deposit legislation, it was hoped that making

collection metadata available for use would encourage experimentation with this dataset and the identification of new tools for 'reading' and interpreting large-scale collections such as this. In collaboration with colleagues at the British Library, a tool was developed for obtaining robust, structured JSON exports of metadata from the UKWA curation tool via API. This contains technical information such as the URL(s) targeted, the date a target was created, the frequency and depth of a crawl, and rights status, along with any descriptive metadata, and significantly improves access to (and supports reuse of) UKWA metadata.

This metadata work was also valuable in helping to improve the representation of the archived web in library catalogs and to encourage a critical consideration of the material as archival source. A frequent hesitation voiced by creators was the concern that preservation in the archive might inadvertently lead a user to access outdated or inaccurate information, and the project team were conscious of the need to ensure archived web content was properly contextualized in order to protect creators and potential users. Archived copies appear with a blue 'UK Web Archive' banner at the top, marked with the capture date in order to clearly indicate that they are not live pages, however some website owners expressed concern that this branding may not be enough to discourage misunderstanding. Led by the Metadata Analyst and Rights Officer, the team explored options for stewarding access to archived captures and communicating the nature of archived web content to users. In response, the following short statement was included in the catalog records for subcollections identified as containing a higher incidence of broadly sensitive material:

> Please take care when accessing, using and sharing information from this collection, which may contain outdated or offensive language, sensitive information about living individuals, or otherwise sensitive material.

The UKWA-derived metadata was also used to produce library catalog records for each target. The Metadata Analyst developed a crosswalk to prepare metadata for ingest into the ALMA cataloging system used by the National Library of Scotland. This repurposed descriptions generated by the web archivists and technical metadata to populate a MARC record for each target, with the intention that such surrogate records would not only facilitate a more user-friendly means of accessing the collection, but also serve as a public-facing source of information about the existence of the UK Web Archive and the archive web more generally. Additionally, it was hoped that representing the archived websites alongside other more 'traditional' sources would serve to impress that they should be understood with the same level of critical deliberation as to the circumstances and context of a site as a publication, and of the collection as an entity.

4.4 Transparency

In an effort to further shed light on the archival processes that contribute to a collection, the exported metadata is accompanied by documentation that provides further information on the technical and non-technical circumstances of the collection's creation. This is based on the 'datasheets for datasets' framework adapted for web archive collections by Emily Maemura, and members of the project team met with Maemura in November 2022 and began drafting a datasheet based on Maemura's questions about a dataset's provenance, parameters, omissions. The resulting document provides context for the Talking about Health collection data made available to researchers, communicating technical details and outlining potential uses of the data. Similar efforts were made to find routes to 'open up' the collection development process to potential researchers, and the team participated in a number of researcher-focused workshops and seminars discussing questions around what contextual and collection-development information researchers need access to and the best means for delivering these. A Discourse channel was established as a space in which documentation about the project's background, aims, and approaches could be shared, and the project team also used this space as a means to share relevant reading materials and other resources.

5. Lessons learned

The AoT project offered a valuable opportunity to study the various processes, workflows, deliberations, and interactions that contribute to the development of a large-scale archived web collection, and a number of useful lessons can be observed. Firstly, the AoT project reinforced the need for continued and sustained resourcing for web archiving work. All web archivists were employed on a part-time basis, and this may have contributed to the shortfall in targets collected—the total sits around 3,500, rather than the 10,000 aimed for, reflecting the level of work necessary to curate a collection around such a sensitive and potentially fraught topic. The presence of the dedicated Rights Officer on the project was particularly illustrative of how transformative adequate resourcing can be: the feedback gathered from website and content creators was invaluable in helping to guide the collection development process, and as rights management work is usually performed by a web archivist alongside their selection, curation, collection, and quality assurance duties, the majority of web archiving staff simply do not have the capacity or resources necessary for such discussions. The value of these conversations can be seen in the final statistics: 21% of this collection can be accessed from outwith a legal deposit library, compared with 1% of the UKWA collections as a whole.

A related (if not unexpected) finding concerns the value of collaboration between individuals and institutions when approaching topics of this scale. The final collection is broad in its coverage with over 70 different subtopics represented. Such broad coverage could not have been achieved by a single institution, and would not have been possible without the input of a range of stakeholders from a variety of disciplines and backgrounds. Scoping the collection was an iterative and discursive process, and having web archivists located within different academic environments meant the team was able to discuss the collection development process with an established group of researchers from health and health-adjacent disciplines. Not only was this extremely valuable in pushing collection efforts beyond the strict categories of 'medical' or 'health' information and in encouraging the team to consider the relative value of different types of information sources, but was also instrumental in helping the team to better understand the needs, concerns, and ambitions of potential users.

Recognizing the subjective and discursive nature of the collection development process, the team considered how curatorial decision-making might be better communicated to potential users and content creators. This could be achieved through access to further technical metadata (for example, information on which collections a target appears in, when it was added to a collection, or how the crawl parameters for a target have changed over time) or reflective documents that describe and communicate how a term or subject has been interpreted. Similarly, while the ethical issues that arise when capturing personal narratives and discourse cannot be avoided, improved documentation of how such concerns have influenced and shaped collecting may go some way to strengthening not only the research value of this material, but the relationships with creators and users.

6. Conclusion

The AoT project offered a rare opportunity for observing the processes and deliberations involved in collaboratively building large-scale collections of archived web content. The collaborative nature of the project allowed for better identification of the points where decisions were based on assumptions and expectations that required probing. This in turn impressed the need for digital historical representations like web archives to provide clear contextual and provenancial information alongside records, and to integrate the mechanisms of archival representation into our understanding of context. Such information will allow users to fully interrogate the relevance, integrity, and reliability of records for themselves.

The project also demonstrated the impact of dedicated resourcing for web archiving, particularly with regard to rights work. By engaging with web creators and site owners, the Rights Officer has been able to articulate

the concerns and hesitations that prevent rights-holders from granting access, and the next step will be to seek means of addressing and alleviating these concerns. First among these will be efforts to increase the visibility and understanding of web archives amongst the general public, and continued advocacy of web archives as a key part of UK research infrastructure.

Finally, while the experiences of the AoT project found that there is researcher interest in and appetite for archived web resources, it also further illustrated the hurdles that must be overcome in order for collections like this to be more widely utilized in research. As has been noted, the biggest barrier to use of the collections are the conditions of the legislation under which they are created, and it is anticipated that expanding the recognition of open licenses to include creative commons licensing will significantly increase the proportion of the archived web that is accessible for research. Improving the ease of the export of data from W3ACT is a constructive step towards an alternative means for access, and providing a route to datasets that can be used for computational analysis while still respecting the boundaries of legal deposit legislation represents significant progress towards increasing the visibility—and with hope, the research use—of the archived web.

With thanks to the Archive of Tomorrow project team: Eddie Boyle (Research Data Engineer, National Library of Scotland); Cui Cui (Web Archivist, Bodleian Library, University of Oxford); Mark Simon Haydn (Metadata Analyst, National Library of Scotland); Mary Garner (Project Manager, National Library of Scotland); Jasmine Hide (Rights Officer, National Library of Scotland); Agnieszka Kurzeja (Metadata Coordinator, University of Cambridge); Eilidh MacGlone (Web Archivist, National Library of Scotland). The full project report is available via the British Library's Research Repository https://doi.org/10.23636/6q6k-8369

References

Acker, Amelia and Mitch Chaiet. 2020. “The weaponization of web archives: Data craft and COVID-19 publics.” *Harvard Kennedy School (HKS) Misinformation Review* 1, no. 3. https://doi.org/10.37016/mr-2020-41

Barrowcliffe, Rose. 2021. “Closing the Narrative Gap: Social Media as a Tool to Reconcile Institutional Archival Narratives with Indigenous Counter-Narratives.” *Archives and Manuscripts* 49, no. 3: 151–66. http://www.doi.org/10.1080/01576895.2021.1883074

Davies, Kerrie M. 2023. “Crowd Coaxing and Citizen Storytelling in Archives of Crisis.” Life Writing 20, no. 2: 351–365. https://doi.org/10.1080/14484528.2022.2106611

Greenwood, Amanda. 2022. “Archiving COVID-19: A Historical Literature Review.” *The American Archivist* 85, no. 1: 288–311. https://doi.org/10.17723/2327-9702-85.1.288

Milligan, Ian, Nick Ruest, and Jimmy Lin. 2016. “Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses.” In JCDL '16: *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 107–110. https://doi.org/10.1145/2910896.2910913

Ogden, Jessica, Ed Summers, and Shawn Walker. 2021.“Patterns of Use: Conceptualising the role of web archives in online discourse.” Paper presented at *Fourth Research Infrastructure for the Study of Archived Web Materials (RESAW) Conference: Mainstream vs. marginal content in Web history and Web archives*, University of Luxembourg, Luxembourg, June 17–18. https://hdl.handle.net/1983/59169b00-10ac-435b-8179-f6b88cff9c1c

Zumthurm, Tizian and Stefan Krebs. 2022. “Collecting Middle-Class Memories? The Pandemic, Technology and Crowdsourced Archives.” *Technology and Culture* 63 no. 2: 483–493. https://doi.org/10.1353/tech.2022.0059

SECTION 7
Studying mediatized memories

Websites as historical sources? The benefits and limitations of using the websites of former repatriates for the history of schooling in colonial Algeria

Christine Mussard

Abstract: Since the 1990s, former repatriates from Algeria, now independent, have used the web as their favorite space to post and share their memories. The school memory plays an important part in these online stories, documented by class photos, testimonies of teachers and students, monographs of schools and, more rarely, personal experiences and institutional documents. Based on a small corpus of websites considered to be born-digital sources, this paper will question the different ways the memories of colonial Algeria are mediated in the internet era. It will look at how colonial experiences are recounted on the web, and also focus on the methodological issues social scientists face when working with these materials.

Keywords: Algeria, methodology, schooling, memories, websites.

This article expresses my opinion and describes my experience of using websites for historical research. They were the default solution I had to resort to because of a lack of iconographic documentation and, more significantly, the difficulty of getting to the country itself. I consulted the websites to gain a concrete understanding of the school environment and to enable me to better illustrate my study of the history of the schooling of Algerian pupils within French primary education between 1944 and 1962 (Mussard 2024). My observations focused primarily on Blida, a region to the south of Algiers in the heart of the well-known Mitidja plain, held up as a prime example of the success of French colonization (Côte 2014). My medium of choice was administrative archive documents, which were available in vast quantities, but I also used a few websites created and updated by former repatriates and their descendants. I had already used the regional websites to seek out testimonies from an area in the East of the country for a previous project researching the mixed commune (*commune mixte*) of El Kala (Mussard 2018).

I used two websites in particular for this analysis. They pertain to a memory of the *pieds-noirs* [French settlers in Algeria] which initially came about with the creation of clubs and associations, one of the iconic ones being *Le Cercle Algérianiste*, founded in 1973 (Moumen 2020). In the early 2000s, the web gradually enabled these groups to raise awareness of a mythicized past and to try to attract new members who could contribute to documentation aimed at publicizing and "safeguarding an endangered

Christine Mussard, Aix-Marseille University-Marseille University, CNRS, IREMAM, France, christine.mussard@univ-amu.fr, 0000-0003-1722-1566

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Christine Mussard, *Websites as historical sources? The benefits and limitations of using the websites of former repatriates for the history of schooling in colonial Algeria*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.27, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 311-320, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

culture"[1]. The *Cercle Algérianiste* website went live in 2000, and another prominent website—*Alger-roi*—was available from 2002[2]. The websites I consulted for the purposes of this article have a much more local focus than the 'heavyweights' of online *pied-noir* memory but there are similarities in the topics of the documentation and the amount of space given over to schooling. It featured prominently in the online archives and in my search for people from the world of education and the school environment. I found quite a few incredible leads among the class photos and the accompanying anecdotes on two websites dedicated to the city of Blida and the surrounding region. So, my approach primarily involved an opportunistic understanding of these websites as sources. However, in addition to the odd photo and memory I took from these websites, I was also able to see the different ways of presenting the school memories which occupied a large part of these community sharing platforms. They revealed how the contributors were attempting to reunite classes in today's very different French Algeria, providing insights into the social connections of the past and the form they take today. The aim of this article is, therefore, to investigate the different ways of using these websites which tell memory-packed stories, sorting the real from the fake in terms of the sources and understanding the practice of memory expression as a research topic.

I will begin with an overview of these popular spaces for posting regional memories and then look at the main content of the website, focusing on how it presents opportunities and limitations for research. Finally, I will investigate the incomplete nature of a community memory and how the omissions hint at the inequalities of colonial domination and are important for historical research.

1. Local history and regional memory

1.1 Regional memories of a colonized village

It was when conducting my doctoral research that I first consulted a website dedicated to *pied-noir* memories, essentially to find, and make contact, with people who had lived through the era and, as the website said, were returning to the places they spent their childhood and reengaging with the Algerian populations[3]. Alongside the old photos and postcards, there were also a large number of videos of members' recent trips to Algeria. As such, the website was both showcasing the trips and also raising the profile of an association that was still seeking new contributors. Today, there is also a very active, closed Facebook group with daily posts.

---

[1] Manifesto of the Cercle Algérianiste association, in *Le Cercle Algérianiste*, March 1985, cited in Moumen 2020.

[2] This information has been taken from the *Wayback Machine*, an online digital archive.

[3] https://amicaledescallois.forumactif.com/. This website was created in 2003.

Photos and short films were posted of the groups of Algerian French members of the association *Amicale de Callois* returning to the East of the country. Most documented the different stages of the journey, starting from the airport, and showed them enjoying spending time together. The once-renowned seaside city of La Calle—today known as El Kala—is some 40 kilometers from the Tunisian border. Back in 1550, the small port was besieged by Tomasino Lenche who built the Bastion de France, known for fishing and coral, which was traded for spices from the Levant. Talking about the event on March 29, 1929, the eve of the centenary of French presence in Algeria, Deputy Mayor of Algiers Louis Filippi mentioned "the early pioneers of civilization in the Barbary states [who founded] a first colonization village, the heroic forerunner of all our Algerian villages" (Martini 2002). During the colonial period, this port was known as a fully-fledged commune (*commune de plein exercise*), in other words its administrative functions were similar to those of mainland France. The area I am investigating—the mixed commune (*commune mixte*) of the same name—borders the south of this port and comprises a few small villages founded during French occupation, some of which appear in the photos posted online. In the first instance, I viewed them on the web alongside the vast map resources available at the national archives of the French overseas territories (Archives Nationale d'Outre-Mer), thus familiarizing myself with them from afar before actually seeing them in person.

1.2 A plethora of memory traces from an influential city

For this new local history research project, I focused on the regional websites which had content relating to the city of Blida and the surrounding area. Others, primarily those related to Algeria-Algiers-Roi, provided information about Blida's schools but were more rudimentary. The city of Blida sits in the heart of a central region of Algeria, an economic hub that was densely populated during the French occupation and the city itself had around 300,000 inhabitants in the mid-1950s. With its large estates and farms, the area had a thriving grape and orange growing industry. It was the birthplace of bastions of French industry, the most famous being Orangina, which was established in Boufarik in 1930, and Bastos, which expanded there from Oran. There was also the Sainte-Marguerite estate where plants were grown for the perfume industry. Many French people lived there and ran businesses alongside the Algerians, forming a vast, heterogeneous population. From the plethora of information on the websites created by former repatriates, it is clear that Blida was an important city whose influence extended to the surrounding region. I was much more interested in the documentary resources on these websites than the potential contacts I could make through them.

The names of the websites I used the most—Blida Nostalgie[4], website A, and Blida Rose de mes 15–20 ans[5], website B—gave clear indications of the creators' intentions and the emotions and nostalgia for times gone by. There was an obvious sense of community and potential associations with modern day Algeria on the El Kala site that was absent from the websites created by individuals who had collected documents and information from their own personal resources or had had them sent to them by other contributors.

Both refer back to a past to be documented in images and words, the main topic being schooling. Land on the homepage of website A, and you are very quickly redirected to the newsletters of the Bilda schools' former pupils' association and the first three tabs of website B are dedicated to the schools of Blida and the surrounding area. Website A, created in 2005, is still active and updated by its founder who is now 80 years old. He has contributed to the success of the participative web which has helped democratize the sharing of memories online (Gebeil 2015). By contrast, nothing has been added to website B since 2009. In both cases, the generations coming after those who lived through Algeria's independence have not taken over the administration of these websites. Moreover, they have no inclination to join member associations.

2. The opportunities

2.1 Making contacts and visiting the country

The official *Amicale des Callois* website is a discussion platform for a vibrant and dynamic community. Visitors are initially invited to join the group by completing a membership form and are also given the opportunity to purchase works and publications produced by members.

Through this website I was able to make contact with one of its most active members who gave me permission to post a request for testimonies and, more importantly, to join a trip organized by the association in 2010. This visit to a rather rural Algeria, my first contact with the country, was difficult. Meeting members of the association and also Algerians living there was hugely beneficial for my research. The testimonies I gathered during my stay and the private archives I then collected from descendants of colonists of the village El Tarf were invaluable sources.

2.2 The website as a source

[4] https://blidanostalgie.fr/index.html
[5] http://michelgast.mathieu.free.fr/algerie/blida/index.html

2.2.1 Identifying segregated school areas

From the two websites I consulted as part of my research into Blida, I obtained a wide variety of documents that depicted the different aspects of schooling in the region. Taken together with the archives and testimonies, I was able to build an in-depth knowledge of the school establishments in the city and the surrounding area. Their names, urban or rural locations, and the history of their creation were important for my research into the schooling issues faced by the Algerian pupils. Through postcards, local press clippings, descriptions of the building and classrooms, I was able to sketch an initial outline of where the young Algerian pupils went to school and identify the sectorization criteria in the city of Blida. A visit to the region itself then confirmed the location of these places which are still present on the various websites I consulted. The information on the website therefore paved the way for my understanding of schooling as one of the strong markers of the different levels of segmentation in the colonial village. By pulling together the class photos and the map I found on the website with the press cuttings and testimonies I gathered in the country, I was able to identify the diversity of the school environment in the city of Blida and, to a lesser extent, the smaller surrounding communities. I was thus able to draw a distinction between the schools which were almost exclusively attended by Algerian pupils and those frequented by the Europeans and to map the logics of a school landscape which the families can tend to overlook. I was also able to use the descriptions of some of the establishments, confirmed by maps recovered from archives, to understand how, in the 1950s, new classrooms and annexes were added to create a reconfigured school complex in order to accommodate some of the Algerian pupils who had until then been excluded from French public establishments.

2.2.2 An almost partial picture of life in the classroom

There were a vast number of class photos, a common feature of former pupil websites. Annotations varied but some were accompanied by lists of almost exclusively European names, suggesting that it was only former repatriates who visited the website. Website A had no schools for Algerian nationals only and there were very few of them on website B.

Stories about these establishments, which were even more rare, gave the reader an in-depth insight into everyday school life, the real-life experience of the “black box ” (Caspard 1990) of the classroom. On website A, the “memories of lessons at Bonnier school” pages were written by former pupils who recounted in detail the quirks of their former teachers, the school’s practices, and the games they played at break time. There was in

particular a lot of information about this boys' school, described as the "very best school in Mitidja and an incubator for talent"[6], posted by former pupils. They described in detail the layout and use of the buildings, some of the teachers and schooling during World War 2 but these accounts were difficult to confirm from other sources. This school featured prominently but there was no further input about others mentioned on a list, the "école de la cité musulmane" girls' school for example. The fact that not all schools and classes are represented equally on the website, and those with the most coverage were mainly attended by Europeans, is evidence that the contributors' aim was not to replicate the entire school landscape at the time of French Algeria but to share only part of the history that showcased the French schooling policy at local level, reflected in the quality of the buildings, the expertise of the staff and also the conviviality of the inaugurations and prize giving ceremonies. Despite the fact they were very much in the minority in this region, almost all of the photos and comments are about the Europeans. A rare exception to this 'blinkered' portrayal of schooling at the time is the page devoted to the *école-ouvreur* Gallieni, a school-cum-workshop opened in 1913 and initially for Algerian girls only. The article about this school contains no testimonies or photos of pupils but rather a cutting from a local newspaper explaining its inauguration, the success which led to its expansion, and the importance of its role within a Muslim population where "mothers of families sent their daughters to school so they didn't have to worry about looking after them rather than as a concern for what they were doing there[7]". The local press cutting did however mention the name of the headmistress, which led me to want to find out more about her and, more generally, to involve the local press which reported on the school inaugurations.

Several links in the list of schools led to detailed testimonies of school life. Learning, a punishment received, and a surprise interrogation are some of the topics which interspersed the very few accounts, and all within the very specific context of the classroom. There was very little about break times and what happened outside the school, but when mentioned, there was reference to the games and childhood practices in the 1950s. Although girls were in the classroom and appeared on some of the photos uploaded, they were rarely mentioned in the memories which were almost exclusively posted by men.

[6] https://blidanostalgie.fr/bulletin-anciens-eleves/Scan%20des%20bulletins/Bulletin_2012/bonnier.pdf
[7] Le Tell journal des intérêts coloniaux, March 18, 1918.

2.2.3 Capturing the rare voice of the teachers

There were fewer memories posted by former teachers, especially concerning the lessons given to Algerians. A female teacher who had worked at the Tirman school for Algerian boys talks about her time in the classroom teaching hard-working pupils who arrived at school soaked after a long walk which took them across a ravine. This testimony is one of the only ones which mentions these local boys and is more about looking after them after the ordeal of their journey than teaching them. However, the teachers were frequently mentioned and described by former pupils, who sent the website administrator group photos or told them about a strict math teacher or an elegant art teacher.

3. The shadowy side of the source

3.1 Perpetrating a closed group

As with all materials used for historical research, a website gives only a partial and one-sided insight into the topic so it is vitally important that the online data is compared with other sources. Website A, which is particularly well populated and updated, and thus the main one I consulted, did as I mentioned contain individual testimonies, photos, postcards, and also press cuttings. These press cuttings, and also extracts from the newsletters of former pupil associations,[8] are different in that they contain personal accounts from the contributors and are more akin to 'documentary evidence', which gives the website greater credibility. However, hardly any of the accounts given refer to the schools for local Algerians which are sometimes mentioned but not backed up by any documents. The names of the pupils in the class photos are almost exclusively of European origin. Indeed, the main school featured on the website is a school for French pupils, or rather Europeans in Algeria. The bias is such that the content could be French, or more specifically about mainland France. This could be explained by the willingness on the part of former repatriates to reunite their classes and the website has thus become a place primarily for members of that community to chat and share documents and information. As such, it encourages and perpetrates a closed group, while at the same time questioning the widespread notion of schooling for childhood friends from all backgrounds that broke down social and cultural barriers. There is absolutely no doubt that this type of social connection did exist between the populations of these colonial societies but there is no mention of it in the documents and accounts, thus excluding it from the past that is being disclosed and preserved.

[8] https://blidanostalgie.fr/journaux/Bulletin%20%20juin%201932/bulletin-juin%201932.htm

So where is Algeria in these school-based community recollections? What route is it taking? The names of the schools, pupils and teachers, the subjects taught, and the classroom processes could easily be those of a primary or secondary school in a North Mediterranean country. This particular portrayal of the school institution in French Algeria contradicts the perceived notion that schooling was an exception to the generally negative 'record' of French domination. Class photos are often seen as tangible proof of what was learnt at the French schools and the childhood and adolescent friendships forged between the different populations. There are indeed Algerians in the photos on the websites consulted but they only account for a very small portion of the indigenous school-age population. Despite being in the majority at the time, this group did not have access to schooling; on the eve of independence, only 31% of them were schooled by the French (Desvages 1972). The mediocre education record of French colonization is representative of the whole Empire and the vast numbers of theses and publications which appeared in the 1970s dispelled any illusion of the success of a "civilizing mission" in both the Maghreb and Sub-Saharan Africa (Barthelemy 2010). By contrast, all the European pupils in Algeria received schooling, just like their counterparts living on the other side of the Mediterranean. As and when French families settled in the mixed communities of colonial villages, small towns, and coastal areas, drawn by public initiatives or the many congregations of different religious denominations, first primary schools and then secondary schools appeared. Areas where there were little or no French inhabitants, on the other hand, had very few schools. The reason for the lack of Algerian pupils on the website is therefore because only a small number of them received schooling compared to the pupils of European origin in Algeria, who received the same level of education as those on mainland France. So behind the story of a steadfast French educational institution lies a schooling divide and a lack of Algerian pupils in the classroom.

Moreover, the various anecdotes and illustrations say nothing or very little about the tensions that were brewing some years or months before the start of the Algerian war of independence, and yet school in the throws of the second world war, when some teachers were called up to fight and flag raising became a daily practice, does get a mention. There is one single reference to murmurings and unrest on May 8, 1945, of "Arabs armed with sticks", but the pupils had come out of school to celebrate the Allies victory. They caused the pupils to flee and suggest that the independence demonstrations in the East of the country also took place in Blida (Rey-Goldzeiguer 2006).

So, if these websites are to be used as potential sources to understand the

history of schooling in colonized Algeria, the gaps and omissions in the school memory cannot be ignored. This means that historians should also have conducted a preliminary quantitative study, backed by statistical sources, of the schooling of these populations.

4. Conclusion

Producing a history of schooling in colonized Algeria using the websites of former repatriates calls for additional methodological precautions. Historians using the web might be delighted to find such a lot of diverse documentary sources, which is of course what the contributors intended. They should however stick to their strict research procedures and compare all types of traces with other sources because the website created, and the content uploaded, is very much influenced by the agenda behind the memory and community recollection. Every source is to a certain extent subjective but these, at the conjuncture of history and memory and packed with truly unique recollections, require particular attention, but this does not mean that they are any less valuable.

The absence of Algerians, but also reminiscences of the normal day-to-day of a French education system in an Algeria where tensions and divides are emerging, lead to gaps and cover-ups in a truncated past, a topic for investigation by historians. Actually, this selective and incomplete account of schooling in Algeria confirms the unequal access to education, and yet this right had been legalized by the introduction of compulsory schooling in 1944 and the end of segregated schooling brought in by the decree of March 1949. The regional recollections uploaded to the web, therefore, confirm the discrepancies and contravention of the rules experienced in the country itself, fuelled in particular by silence and omission.

References

Barthelemy, Pascale. 2010. “L’enseignement dans l’Empire colonial français : une vieille histoire?” *Histoire de l’éducation* 128: 5–27. https://doi.org/10.4000/histoire-education.2252

Caspard, Pierre. 1990. “Introduction.” *Histoire de l’éducation* 46: 1–3. https://www.persee.fr/doc/hedu_0221-6280_1990_num_46_1_3332

Côte, Marc. 2014. “L’exploitation de la Mitidja, vitrine de l'entreprise coloniale?” *Histoire de l'Algérie à la période coloniale*. *1830–1962. La Découverte*, edited by Abderrahmane Bouchène. https://doi.org/10.3917/dec.bouch.2013.01.0269

Desvages, Hubert. 1972. “La scolarisation des musulmans en Algérie (1882–1962) dans l'enseignement primaire public français. Étude statistique.” *Cahiers de la Méditerranée* 4: 109–137. https://doi.org/10.3406/camed.1972.1678

Gebeil, Sophie. 2015. “Le Web, nouvel espace de mobilisation des mémoires marginales. Les mémoires de l’immigration maghrébine sur l’internet français (2000–2013).” *Cahiers Mémoire et Politique* 2. https://doi.org/10.25518/2295-0311.115

Martini, L. 2002. “Bastion de France. La costa che guardano li Francesi in Barbaria.” Corsicans and the French Overseas Territories, *Ultramarines* 22.

Moumen, Abderahmen. 2020. “Les pieds-noirs en 1973, le tournant mémoriel?” *Hommes & migrations* 1330: 55–57. https://doi.org/10.4000/hommesmigrations.11438

Mussard, Christine. 2018. *L’obsession communale : La Calle, un territoire de colonisation dans l’Est algérien. 1884–1957*. Aix-en-Provence: Presses universitaires de Provence. https://doi.org/10.4000/books.pup.46035.

Mussard, Christine. 2024. “L’école empêchée. Blida et sa région, 1944–1962.” Thèse d’habilitation à diriger des recherches, Université de Lille.

Rey-Goldzeiguer, Annie. 2006. “Aux origines de la guerre d’Algérie. De Mers-El Kébir aux massacres du Constantinois.” Paris: *La Découverte*. https://doi.org/10.3917/dec.goldz.2006.01

A Social media archive for digital memory research

Costis Dallas, Ingrida Kelpšienė

Abstract: Social media is an important social and cultural interaction arena, and a growing field of social research. Acknowledging the limitations of social media platforms and institutional web archiving initiatives to fully support the needs of researchers, this chapter makes the case for a reorientation of social media archiving, drawing from critical digital curation and archival theory to define specifications for a data architecture applying knowledge graphs, and aspects of the Open Archive Information System standard, to support research on Lithuanian memory, heritage, and identity interactions on social media. Based on this experience, it discusses broader implications for web archiving and digital curation in the context of research data infrastructures.

Keywords: social media, semantic modeling, web archiving, research data archives, digital curation.

1. Introduction

Social media platforms have become central to contemporary social and cultural practices, profoundly informing the way individuals and communities communicate, share information, construct their identities (Papacharissi 2011), establish affiliative relationships (Baym 2010) and social networks (Garton, Haythornthwaite, and Wellman 1997). Social media practice is central in phenomena such as the memory wars in Eastern and Central Europe (Rutten 2013), political protest in the Arab spring (Tufekci 2017), Holocaust memory (Manca 2020), fake news in nationalist narratives (Bonacchi 2022), and contestations on the difficult past and heritage (Kelpšienė et al. 2023). Such practices are distinct in their reliance on the 'logic' of social media platforms.While these platforms support identity, presence, relationships, reputation, group membership, conversations, and content sharing functions for their users (Kietzmann et al. 2011), they do so by activating mechanisms of datafication, commodification, and algorithmic selection (van Dijck et al. 2018). These mechanisms pose challenges to the autonomy and agency of individuals and communities, simultaneously subverting truth, trust, and the public sphere.

The ubiquitous nature of social media, coupled with its dynamic and interactive capabilities, renders it an invaluable resource for information, communication, and social science research (Garton, Haythornthwaite, and Wellman 1997; Snelson 2016; Stoycheff et al. 2017; Stieglitz et al. 2018; Shibuya, Hamm, and Pargman 2022). However, the ephemeral nature of social media content, combined with the proprietary algorithms and platform policies governing data access, poses significant challenges for researchers. As Ben-David (2020) contends, the task of researching social media is fraught with complexities not least because of the "unarchivable" nature of platforms like Facebook, which claims the role of the "archon" of

Cotis Dallas, Vilnius University, Lithuania, konstantinos.dallas@kf.vu.lt, 0000-0001-9462-0478
Ingrida Kelpšienė, Vilnius University, Lithuania, ingrida.vosyliute@kf.vu.lt, 0000-0003-3741-9510

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Cotis Dallas, Ingrida Kelpšienė, *A Social media archive for digital memory research*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.28, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 321-340, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

contemporary records of communicative social action. Platform architecture, which privileges the constant accumulation and flow of new content, complicates efforts to capture and maintain a stable record of digital interactions for scholarly analysis. Moreover, legal and ethical considerations surrounding user privacy and data access, and alarming efforts of platforms to curtail data access for scholarly research on social media practices and their political and ethical implications (Bruns 2019) further complicate the landscape for researchers seeking to investigate social media data.

Establishing a social media archiving infrastructure that can serve the needs of scholarly research beyond the affordances of platforms thus emerges as a crucial challenge for social media researchers. In this chapter, we share our insights from establishing a data archive suitable for investigating digital memory, heritage, and identity practices on Lithuanian social media within the context of the Connective Digital Memory in the Borderlands research project.[1] We discuss how this endeavor addresses theoretical and methodological issues relevant to web archiving, digital curation, and social media research. Specifically, we explore our understanding of the challenge in designing a social media archive suitable for scholarly research; our knowledge graph approach to the semantic representation of social media data to tackle issues of intelligibility, analytical power, and theory building; how our social media open archives architecture and workflow addresses the complementary methodological challenge of reliability of collected data, and the related digital preservation requirements of integrity and authenticity. Finally, we reflect on the lessons learned from this experience regarding the specification and design of social media archives for scholarly use.

2. The Challenge of social media archiving for scholarly research

Numerous studies in the field of digital preservation and research data infrastructures have sought to address the challenges inherent in social media data archives: data capture and collection (Littman et al. 2018; Lomborg and Bechmann 2014; Marres and Weltevrede 2013; Pehlivan, Thièvre, and Drugeon 2021); long-term preservation and data management (Thomson 2017; Voss, Lvov, and Thomson 2017; Hemphill, Leonard, and Hedstrom 2018); ethical and legal constraints (Zimmer and Kinder-

[1] "Connective Digital Memory in the Borderlands: A Mixed-Methods Study of Cultural Identity, Heritage Communication and Digital Curation on Social Networks" is a three-and-a-half-year project undertaken by the Connective Research Group at the Faculty of Communication, Vilnius University, aiming to explore heritage, memory, and identity-related interactions on Lithuanian social media. The project received funding from the European Social Fund (project No. 09.3.3-LMT-K-712-17-0027) under a grant agreement with the Lithuanian Science Council (LMT).

Kurlanda 2017; Bruns 2019; Franzke et al. 2020; Fiesler and Proferes 2018); and, last but not least, access and use of social media archives, and their impact on scholarly research methods and practices (Thomson and Kilbride 2015; Weller 2014). Concurrently, several social media archiving tools have emerged (Borji, Asnafi, and Naeini 2022; Social Media Lab 2024), and some national libraries, national archives, and other bodies and initiatives in the field of digital preservation and curation, have been actively setting requirements, developing guidelines, and initiating wide-scope projects for institutional social media archiving on both national and international scale (Hockx-Yu 2014; Thomson and Kilbride 2015; Pehlivan, Thièvre, and Drugeon 2021; Acker and Kriesberg 2017). However, while the scholarly value of web archives in general is widely acknowledged (Brügger and Schroeder 2017; Brügger and Laursen 2019; Vlassenroot et al. 2019), there is little evidence that institutionally-based *social media* web archives are actively used by researchers. This suggests that such initiatives, unlike focused, narrowly targeted data archives produced by researchers to support their own investigations, aim for the broad objective of long-term digital preservation, rather than active research use. Indeed, as noted by Vlassenroot et al., "the use of publicly accessible national archives and large-scale social media archives in scientific studies [is] only just emerging", with the focus primarily on social media use by government entities, and in the context of natural and health emergencies (Vlassenroot et al. 2021, 118; cf. Michel et al. 2021; Milligan, Ruest, and Lin 2016).

The risk of "epistemic failure—the inability to account for diverse theoretical, substantive and methodological perspectives in particular disciplinary traditions which require access to digital resources" is a concern for digital data repositories in general. This prompts a call for "a radical re-examination of current notions of *context* ... so that it encompasses the structure and evolution of the pragmatic references of such objects in the real world" as well as "semantic representations of the *epistemic content* of curated information objects … account[ing] for dynamically evolving semantic representations of 'things in the world' at the instance (occurrence) level as well" (Dallas 2007). Acknowledging the tension between researcher expectations and affordances of national web archives in Denmark and the UK, Jessica Ogden and Emily Maemura also underscore the need to consider "the relationship between records and the surrogate material objects they represent." To adequately access records, they had to move beyond document-centric and collection-centric views of WARC data in the archive, relying "on other representations and views into web archives at different junctures, including the use of Solr search indices and query interfaces, faceted and free text search interfaces under development, curation tools (which provided different seed- and collection-

centric views), and numerous ad-hoc and incomplete data" (Ogden and Maemura 2021, 59–60).

These broader challenges are further exacerbated by the unique nature of social media. The three complementary foci of social media research—content, interactions, and network structure—are constituted differently within the context of different platform affordances, disciplinary traditions, theoretical frameworks, and methodologies endorsed by specific research projects, casting doubts on the possibility of a 'one size fits all' approach to social media web archiving. The unit of inquiry in social media research is extremely variable, ranging from individual posts to the global graph of social media interactions. "Where to cut the network", to paraphrase Latour, is very much a theory- and researcher-laden question, and crucial contextual elements such as reactions, re-posts (retweets, shares), and comments are often absent from social media archives. Social media interactions are ephemeral and dynamic, accumulating interactions and "layers of context" after they have been collected (Acker and Kriesberg 2017), and retrieving information about deleted or altered data is often impossible (Ben-David 2016). Besides, on platforms such as Facebook, the record of each social media interaction is inherently plural and context-dependent, as different users may experience the same conversation differently based on their profile, history, and network of 'friends'—a reality that sits uncomfortably with the notion that web archives can capture a singular, fixed, and objective representation of social media.

Overall, this situation points to an onto-epistemological entanglement between the conceptualization of social media as a field of interactional communicative practice (Lomborg 2012) and almost all facets of establishing a research-capable social media archive. These facets include adopting appropriate capture strategies, ensuring authenticity and integrity of social media records, providing sufficient scope and expressiveness in social media data representation, and supporting digital methods of access and analysis. This entanglement is especially relevant within the context of datafication (Rogers 2013; Schäfer and van Es 2017), alongside the emergence of digital methods and practices tailored specifically for social media research (Edwards et al. 2013; Winters 2017; Perriam, Birkbak, and Freeman 2020; Wilson 2022). Recognizing the legitimacy of "a distinction between archiving social media data for a specific research purpose (scholar uses) and institutional archiving" (Pehlivan, Thièvre, and Drugeon 2021, 44), and drawing inspiration, among other factors, from advancements in the research use of web archives and the maturation of web archiving theory and practices over the past two decades (Brügger 2018; Helmond and van der Vlist 2019), we aim to challenge the notion that research-capable social media archives are still confined today to a form of micro-archiving previously characterized as "small scale", "here-and-now", and undertaken

by “individuals … whose technical knowledge of archiving or of the subsequent treatment is either lacking or on an amateur level” (Brügger 2005, 11).

In what follows, we address aspects of this onto-epistemological entanglement by sharing the challenges faced and decisions made in creating a social media archive capable of supporting data access, mixed-methods analysis, and theory-building within the framework of the Connective project.

3. Representing social media in a research-capable data archive

The Connective project is based on multiple analyses of a corpus comprising over 30,000 conversations (more than 250,000 posts and comments, authored by over 90,000 unique users) between 2016 and 2023, and sourced from Lithuanian Facebook accounts, pages, and groups, Instagram hashtag collections and accounts, and VKontakte communities. While it is not possible to assess reliably how much of the overall social media activity in Lithuania focuses on conversations about heritage and the past, such conversations clearly establish an active arena for negotiating important aspects of collective identity, values, and attitudes towards contemporary issues. Facebook data were identified and collected in summer 2022 by means of several hundred online queries, conducted under five different Facebook user accounts. Lexical expressions used in the queries are connected to 72 topics established by the research team after an initial qualitative scoping of relevant conversations on Facebook. These topics belong to nine broader themes: history, ethnoculture and language, 90s culture, religion, minorities, everyday life, war in Ukraine, objects and memory wars, and contemporary concerns. This corpus was further enriched by 75 interviews with users actively engaged in contested heritage meaning-making and identity work. Research in the Connective project, conducted by a transdisciplinary team of eleven researchers includes an integrative analysis of communicative repertoires, interactional patterns, and affiliative network structures with specific case studies of post-memory and identity construction on Lithuanian social media. These case studies explore themes such as: the remembrance of childhood in the late Soviet period and the challenges faced by youth subcultures in the 1990s; the post-memory of WWII Polish Home Army partisans in Lithuania; the self-identification of Russian Lithuanians, drawing on conceptions of the past; how invented traditions and ethnographic performance shape creative ethnicity among Lithuanians; the troubled memories of events related to the resistance and dual occupation by the Nazis and Soviets from 1939 to 1953; the ‘monument wars’ raging against public monuments and memory institutions focusing on historical figures from the Soviet era; conflicts around the use

of the 'foreign' letters *w*, *q*, and *x* in Lithuanian passports; and the complexities of remembering and the silencing of the Roma Holocaust. A second data collection season, focusing on additional queries representing topics relevant to these case studies, took place in fall and winter 2023. The research team employs diverse approaches, including corpus linguistics methods, narrative analysis, critical discourse analysis, metaphor analysis, visual methods, qualitative content analysis, and social network analysis, to identify cultural meanings embedded within social media content, group affiliations and influence in user interactions, and cultural schemas (such as stereotypes, conceptual metaphors, and deep narratives) that shape identity construction.

The Connective project leverages social media as a data infrastructure to facilitate multifaceted data-intensive investigations into various aspects of encounters with the past and participatory practices on social networking sites (Dallas 2018; Kelpšienė 2021). We have been working with digital data to reveal how memory practices on Lithuanian Social Network Sites (SNS), mediated by contested heritage, shape cultural identities. Based on identifying "a tentative proposed set of relationships, which can then be tested for validity [and] can often help in working through one's thinking about a subject of interest" (Bates 2009, 3), we drew from activity theory (Engeström 1999) and cultural semiotics (Lotman 2005) to establish an event-centric ontology of SNS semiotic activity on heritage, memory, and identity (Kirtiklis et al., 2023), viewed as a practice of digital curation "in the wild" (Dallas 2016). Our research was empirical and data-driven, based on analyzing a large number of conversations on the history and difficult past of Lithuania on Facebook, Instagram, and vKontakte. As we grappled with establishing a corpus of social media data needed for our research, a key question emerged: what are the properties of archived information objects, and their relationships, that enable their use as epistemic objects suitable for evidence-based scholarly research?

Our data infrastructure embodies an approach to social media archiving designed to capture the multifaceted nature of online interactions. At its core is a commitment to preserving the integrity, authenticity, and intelligibility of social media data, thereby ensuring its utility for scholarly research. This section presents the design decisions that underpin our data infrastructure, focusing on two aspects: the capture of the 'experienced' manifestation of social media content, and the representation of its conceptual properties through knowledge graphs.

Recognizing that the value of social media content often resides in its presentation and the context of user interactions, the Connective project adopted a web archiving approach that accounts for capturing social media content as it was experienced by users. This entails collecting facsimile scrolling screenshots of expanded social media conversations, ensuring the

preservation of visual layout, interactive elements, and temporal sequencing of posts. This method mirrors the user's perspective, retaining the contextual cues and platform-specific nuances crucial for interpreting social media discourse. Such an approach not only maintains the visual and interactive integrity of the data, but also respects its original context, addressing a critical need for authenticity in digital archiving (Ben-David 2020).

In addition to preserving experienced data, we place emphasis on the ontological, structural, functional, and semantic aspects of social media interactions. These elements are re-envisioned through a redefinition of the notion of "significant properties" as those aspects that can warrant "the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record" (Grace, Knight, and Montague 2009, 3), essential for the archive's intelligibility and informational value in a research context. It has been noted that platforms such as Facebook, designed to serve business purposes, are not suitable to account for rich representations of social media useful for social research (Helmond and van der Vlist 2019, 18). Our approach therefore was to extract structured representations of social media data into an external research data archive for further analysis. To capture significant properties based on the content and context of social media interactions in our data we adopted the property graph data model supported by the Cypher query language as applied in the popular Neo4j graph database management system (Robinson, Webber, and Eifrem 2015). This approach allows for the mapping of a semantic schema onto neo4j labels representing key entity types involved in social media conversations, as well as their properties and relationships.

The Connective property graph schema provides for the semantic representation of three types of nodes accounting for social media interactions: Message, representing posts, comments or replies; Thread, representing the sequence of a post and following comments and/or replies; and Actor, representing social media users who post, react, or share Messages and collectivities (such as Facebook groups and vKontakte communities) where such interactions are displayed. An additional type of nodes, Topic, is used to represent descriptive thematic categorizations. Topics are organized in a taxonomy capturing the hierarchical relationships between concepts, and are connected to relevant Messages, Threads, and Actors in a way that allows plural characterizations of any node based on notions in the Topics taxonomy. The taxonomy is faceted, and therefore different Topic hierarchies can accommodate not just subject-laden thematic categorizations, such as people, events, and places mentioned in a Message, but also different kinds of analytical categories and discursive constructs identified by our analysis of the data. In addition, the schema includes a Deposit node type aimed to represent aspects of the process of data capture

of different sets of Threads, providing for provenance and preservation metadata in the archive. Different types of relationships (PART_OF, RESPONDED_TO, FOLLOWED_BY, SHARED_FROM, POSTED, REACTED_TO, CONNECTED_AS, DISPLAYS, CONTAINS, IS_ABOUT) are established to capture the compositional, presentation, and discursive structures of social media interactions as well as group membership, semiotic activities, and reactions of users.

The adoption of knowledge graphs using the Connective schema facilitates a dynamic representation of social media data, enabling the articulation of complex relationships and attributes inherent in online interactions. For instance, thematic connections between posts, the network dynamics of user interactions, and the temporal and spatial context of conversations are all encoded within the graph structure. This method aligns with an agency-oriented approach to digital curation theory and practice, which advocates for structured, queryable representations of data prioritizing their dynamic, event-centric dimension, and thus capable of accommodating the fluid and interconnected nature of social facts manifested in records (Dallas 2007). By deploying knowledge graphs, the Connective project ensures that the archive not only serves as a repository of digital artifacts, but also as a rich, navigable resource for exploring the depth and breadth of social media discourse.

Our dual-faceted approach to social media archiving—capturing both the experienced manifestation and the conceptual properties of data—reflects a comprehensive strategy that balances the need for authenticity with the demands of scholarly research. The project's infrastructure is designed with the "fitness for purpose" principle in mind, tailored to meet the specific needs of the social media research community (Dallas 2016). By preserving the experienced context of social media interactions alongside their conceptual underpinnings, the project addresses the critical challenge of representing digital social interactions in a manner that is both faithful to their original context and conducive to rigorous academic inquiry.

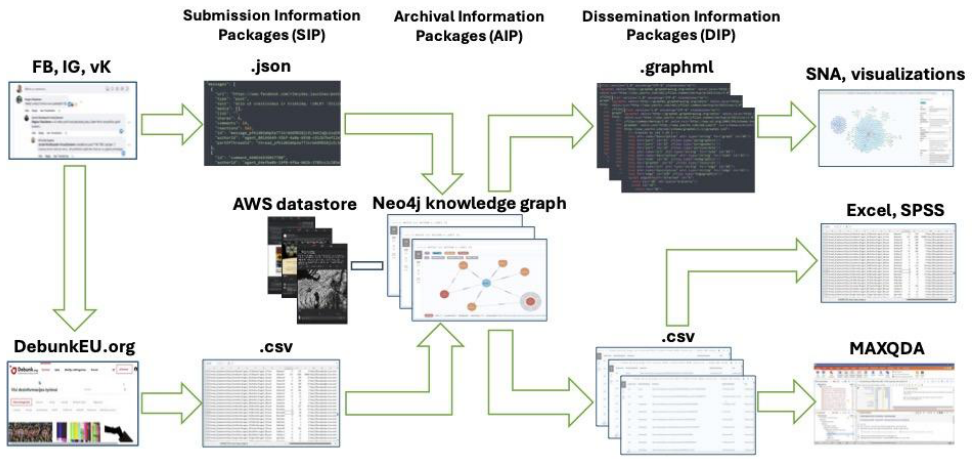
In summary, the representation of social media interactions in the Connective project's data infrastructure embodies an approach to social media archiving that upholds the authenticity of digital interactions while providing a robust framework for their analysis. Through the use of facsimile screenshots and knowledge graphs, the project captures the richness of social media discourse, ensuring its integrity, authenticity, and intelligibility for future research endeavors.

4. An Open archives social media data curation architecture

The Connective project's social media data curation architecture (Fig. 1) transcends the stages of the widely accepted Digital Curation Centre (DCC)

digital curation lifecycle and elaborations aiming to extend the model beyond the ingestion to disposal cycle (Higgins 2008; Constantopoulos et al. 2009). Drawing on the principles of the records continuum model (Upward, McKemmish, and Reed 2011), the architecture is designed to accommodate the dynamic and evolving nature of social media data, ensuring that the archive remains relevant and accessible to researchers across various stages of their inquiry.

Figure 1. Connective social media archive data architecture



The data representation and analysis workflow supported by our architecture begins with a multifaceted approach to data selection, guided by research questions that are served by both query-based and collection-based strategies of data capture. This dual approach allows for the targeted gathering of data pertinent to specific research questions within focal case studies, while also accommodating broader sweeps of content for scoping and exploratory analysis of the full archive. Inspired by OAIS, the ISO-standard reference model for an Open Archival Information System (CCSDS 2012), our data architecture. maps the processes, information structures, and systems used to appraise, capture, ingest, enrich, and provide research access to social media information in the Connective project, while also clearly separating representations of data in the archive between Submission Information Packages, Archival Information Packages, and Dissemination Information Packages. This provides for a reliable mechanism to ensure the integrity and authenticity of collected data, while also providing for the needs of a designated community of researchers

without sacrificing analytical power and the dynamic enrichment of data with additional properties and relationships as researchers "exercise the archive" (Dallas 2016).

Submission Information Package (SIP). Adopting the concept of the Submission Information Package (SIP) from the Open Archival Information System (OAIS) model, the project curates an initial collection of data that includes both 'raw' JSON streams and facsimile screenshots of social media pages. The JSON stream captures decoded textual and media content as it appears on the platform, preserving the 'raw' data in its most unfiltered form. Concurrently, the screenshots serve to document the experienced manifestation of the content, retaining the layout, design elements, and interactive features integral to interpretation. This combination of data formats provides a comprehensive snapshot of social media interactions, capturing both the underlying data structures and the user-facing presentation of content, and ensuring their integrity and authenticity.

Archival Information Package (AIP). The Archival Information Package (AIP) in the Connective social media archive represents a significant advancement in the curation process, where the raw submission data undergoes semantic decomposition and mapping into a knowledge graph. This transformation facilitates a structured representation of social media content, encompassing the structural and semantic elements of interactions and community structures. The knowledge graph not only captures the content and dynamics of social media interactions, but also incorporates provenance metadata detailing the circumstances of data capture. This aspect is particularly crucial in platforms like Facebook, where content visibility can vary based on user profiles and where data may be altered or deleted over time. By documenting the provenance of data, the AIP ensures the traceability and reliability of archived content, providing researchers with essential context for their analyses. Crucially, the representation of archived content by mapping originally captured data as knowledge graphs offers information relevant for research to be searched, categorized, and enriched algorithmically in ways that would have been impossible were the data retained only in their originally ingested format.

Dissemination Information Package (DIP). The Dissemination Information Package (DIP) is tailored to meet the diverse needs of researchers, offering both predefined and customizable access methods to the archived data. Researchers can interact with the neo4j knowledge graph through predefined or arbitrary Cypher queries facilitating the exploration of specific themes or patterns within the data. Additionally, scripts can generate filtered and ordered views of the knowledge graph, preparing data for export and further processing in analytical tools such as MaxQDA and various social network analysis software tools. This flexible dissemination

strategy ensures researchers can access and utilize archived data in ways that align with their specific research objectives and methodologies.

In essence, the Connective project's social media data curation architecture embodies an approach integrating the principles of open archives and digital curation to meet the complex demands of social media research. By navigating the challenges of data variability, provenance, and accessibility, the project establishes a robust framework for the long-term preservation and analysis of social media interactions, paving the way for insightful explorations of digital culture and communication in future research.

5. Curation/creation: Exercising the social media archive

The innovative approach adopted by the Connective project in developing its social media data curation architecture epitomizes a dynamic interplay between archival preservation and the active generation of new knowledge. Drawing from an agency-oriented digital curation approach (Dallas 2007), the project's methodology underscores the inclusion of evolving representations of social media interactions within Derived Archival Information Packages (CCSDS 2012), dynamically enriched with additional properties and relationships as researchers work with the data. This evolving nature of AIPs facilitates the incorporation of plural insights and analyses conducted by researchers, effectively transforming the archive into a living entity that grows in informational depth and breadth over time.

Using the Connective data infrastructure, we can search for complex lexical patterns in messages, and extract dictionaries and repertoires of themes. We can query the archive asking questions such as “which thematic categories of messages attracted on average the highest number of comments?”, useful to support a mixed-methods investigation, combining qualitative with quantitative analysis on the full corpus of conversations in the archive. We may also identify attribution and predication discursive constructs in social media interactions by initiating a collocation analysis based on dictionary-based lexical queries on the text of Messages, for example, to ask: “which identifications related to heroism are made for persons who have been identified as anti-Soviet partisans?” Applying communicative nexus theory (Laužikas and Dallas forthcoming), we can also use the knowledge graph to explore the circulation of agency between historical referents identified as Topics, people and organizations identified as Actors, and communicative acts identified as Messages, for example, asking questions such as: “when, and by whom, was the idea that Peter Cvirka was a Soviet collaborator first asserted on Facebook, and how did this idea circulate across the network?”

The Connective social media archive is distinguished by the combination of semantically rich representations of social media interactions as knowledge graphs with an open archives data architecture which, while ensuring integrity and authenticity of ingested data, allows their knowledge enhancement as researchers produce plural identifications and relationships through analysis and interpretation. This approach has some important advantages for a scholarly research data infrastructure.

Knowledge graphs as dynamic repositories. Central to this dynamic enrichment of the AIPs is the use of knowledge graphs, which serve as the backbone for representing the complex web of social media interactions. The introduction of a faceted Topic node taxonomy, grounded in an ontology that captures the semiotic and epistemic dimensions of online discourse (Kirtiklis et al. 2023), allows for the mapping of diverse aspects identified through axial coding, such as referents, cultural models, and affiliative relationships. This ontological approach enables researchers to embed within the knowledge graph new objects representing refined characterizations and relationships, thereby expanding the archive's conceptual structure.

Facilitating advanced social media analysis. This expanded conceptualization of the AIPs is instrumental in supporting advanced scenarios of social media analysis and interpretation. For instance, in analyzing social media discourse surrounding Lithuanian partisans, researchers in the Connective project can identify critical semiotic and discursive constructs such as attributional relationships (whereby individuals are ascribed certain qualities or roles), predicational statements (which articulate actions or states attributed to subjects), and pervasive metaphors (which structure understanding through conceptual mappings). The dynamic nature of the knowledge graph enables these constructs to be identified, categorized, and represented within the AIPs, transforming abstract analytical concepts into tangible elements of the curated research data archive.

Enriching grounded theory-building: By enabling the representation of such constructs within Derived AIPs, the Connective project's archival infrastructure not only supports the initial identification of these elements, but also their utilization in iterative cycles of analysis. Researchers can leverage the enriched Derived AIPs to question, refute, or reinforce grounded theories about the phenomena under investigation. For example, the attribution of heroism to historical figures in social media discourse can be examined in the light of prevailing cultural models and metaphors identified within the archive, providing an evidence-based understanding of collective memory construction and identity negotiation in digital memory.

This capability to dynamically enrich the AIPs with new knowledge graph objects and to utilize these enhancements in subsequent analyses

exemplifies the archive's potential role as an active participant in the research process. Rather than serving merely as a repository of static data, the archive becomes a collaborative platform where the boundaries between curation and creation blur, fostering a symbiotic relationship between archival practices and scholarly inquiry.

Fit for purpose: Serving research needs. The specification of a research-capable social media archive, as exemplified by the Connective project, underscores the importance of aligning archival practices with the specific needs of the research community. The ability of the archive to adapt to and incorporate the evolving insights of researchers is a testament to its fitness for purpose. By providing a flexible and expandable infrastructure that accommodates the analytical endeavors of scholars, the project ensures that the archive remains not only relevant, but indispensable to the exploration of complex social media phenomena.

In conclusion, the Connective project's approach to social media archiving, characterized by its dynamic and interactive AIPs, introduces a new direction for research-capable archival systems. By enabling the continuous enrichment of the archive with new knowledge and insights generated through scholarly analysis, the project demonstrates the critical role of archives in advancing our understanding of digital culture and communication. This model affirms that for a social media archive to be truly 'fit for purpose' it must be designed to serve the evolving needs of researchers, facilitating the discovery, analysis, and interpretation of social media interactions in ways that account for the dynamic nature, plurality, and context-dependency of social knowledge.

6. Reflections on web archiving, digital curation, and research infrastructures

The landscape of web archiving has traditionally focused on long-term preservation, fixity, and authenticity, often neglecting the dynamic and evolving nature of digital content. Driven by institutional memory organizations, web archiving initiatives have emphasized the creation of static repositories to safeguard digital heritage for future generations. While invaluable for preserving the digital record, this approach often overlooks the active use, enrichment, and transformation that characterize the lived experience of digital research ecosystems. Similarly, the field of digital curation has faced its own set of challenges. Initially vibrant and innovative, digital curation research risks stagnation due to its close ties to institutional and preservation-centric initiatives. The prevalent 'preservation vault' lifecycle model often fails to recognize the importance of knowledge representation and the specific needs of designated user communities, such as, notably, researchers. This narrow, custodial perspective on curation

overlooks the researchers' active role in curating social media records, infusing them with meaning through their analytical and interpretive work.

To address the multifaceted challenges posed by digital curation in the social media context, our approach draws on an alternative, pragmatic perspective on digital curation (Dallas 2016). This perspective prioritizes descriptive attention to digital curation as it occurs empirically "in the wild", involving a diverse range of curating actors, activities and objects of curation, over prescriptive rules within the custodial realm of archival professional work. It highlights the importance of embracing multiple viewpoints and relationships among records and their contextual frameworks, rather than adhering to rigid and formulaic approaches. It also recognizes that records are not fixed but evolve across time as knowledge and interpretive frameworks change. This aligns with the need to account for memory and identity as relevant paradigms for archival theory, advocated by Terry Cook's fundamental critique of the narrowness of dominant approaches to the design and professional practices of archives (Cook 2013). It also resonates with the records continuum approach (McKemmish 1997; Upward, McKemmish, and Reed 2011), which emphasizes that the responsibilities of archivists and data managers extend beyond the moment of record creation, encompassing the plural nature of interpretations and the need to consider diverse perspectives, particularly within the realm of online cultures.

This approach acknowledges that social media archives are not static entities but are shaped by ongoing interactions, evolving meanings, and the diverse voices and experiences of users. It is enabled by a comprehensive framework that acknowledges the dynamic nature of social media archives and the challenges posed by diverse perspectives and modes of knowing. Building upon critical digital curation and archival theory, our framework extends beyond the traditional understanding of a social media archive as merely capturing and managing fixed records in a preservation vault. Instead, it welcomes plural interpretations of social media records within the archive, recognizing the perspectival concerns of various stakeholders, and being open to the ethical and political dimensions inherent in decolonial and indigenous ways of knowing, as well as the complexities arising from the dynamics of online communication and the logics of social media platforms.

In the realm of digital research infrastructures, a tension persists between large-scale, often multinational or national projects that mimic traditional archival and library functions, and the more granular, practice-oriented data management activities of individual researchers and projects. The latter, crucial for shaping the epistemic potential and interpretive frameworks of research, remains relatively unexplored in discussions on research infrastructures. This gap highlights the need for a deeper understanding of how data models, taxonomies, and curation practices directly influence the

trajectory and outcomes of scholarly inquiry. The Connective project's approach to creating a social media data archive tailored for scholarly research represents a promising convergence of these domains. By prioritizing the significant properties of social media interactions that are vital to researchers, the project shifts attention from mere preservation to active engagement with the data. This perspective acknowledges the challenges of authenticity and integrity in social media data, while recognizing that the meaningfulness of information suitable for research in the human sciences hinges on what Ian Hacking (1999, 123) refers to as "interactive kinds"—those that are epistemically constructed and dynamically shaped by research activities.

Knowledge representation and semantic technologies, such as knowledge graphs, provide a powerful toolkit for rethinking social media web archiving. These technologies enable a more nuanced and flexible representation of social media data, facilitating the identification, annotation, and linking of content in ways that align with the conceptual and operational needs of social media research. This approach not only enhances the accessibility and usefulness of archives for researchers, but also fosters a curatorial effect through which scholarly engagement actively contributes to the construction and evolution of the research data landscape.

In this regard, the approach adopted by the Connective project may serve as a promising model for integrating web archiving, digital curation, and research infrastructures in a manner that is responsive to the dynamic and interactive nature of social media research. By fostering dialogue between these fields and placing the research process at the forefront as a data curation practice, the project lays the groundwork for a more engaged and reflective approach to the archiving and analysis of social media content. This paradigm shift holds the potential to enrich the field of social media research by opening up new avenues for exploring the complex interplay of digital interactions, cultural practices, and societal discourses in the online sphere.

References

Acker, Amelia, and Adam Kriesberg. 2017. “Tweets May Be Archived: Civic Engagement, Digital Preservation and Obama White House Social Media Data.” *Proceedings of the Association for Information Science and Technology* 54, 1: 1–9. https://doi.org/10.1002/pra2.2017.14505401001.

Bates, Marcia J. 2009. “An Introduction to Metatheories, Theories, and Models”. *Library and Information Science* 11, 444: 275–97.

Baym, Nancy K. 2010. *Personal Connections in the Digital Age*. Cambridge: Polity.

Ben-David, Anat. 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.

———. 2020. “Counter-Archiving Facebook.” *European Journal of Communication* 35, 3: 249–64. https://doi.org/10.1177/0267323120922069.

Bonacchi, Chiara. 2022. *Heritage and Nationalism: Understanding Populism through Big Data*. UCL Press. https://doi.org/10.2307/j.ctv1wdvx2p.

Borji, Samaneh, Amir Reza Asnafi, and Maryam Pakdaman Naeini. 2022. “A Comparative Study of Social Media Data Archiving Software.” *Preservation, Digital Technology & Culture* 51, 3: 111–19. https://doi.org/10.1515/pdtc-2022-0013.

Brügger, Niels. 2005. *Archiving Websites: General Considerations and Strategies*. Aaarhus: Center for Internetforskning.

———. 2018. *The Archived Web: Doing History in the Digital Age*. Cambridge, Mass.: MIT Press.

Brügger, Niels, and Ditte Laursen. 2019. *The Historical Web and Digital Humanities: The Case of National Web Domains*. Routledge.

Brügger, Niels, and Ralph Schroeder, eds. 2017. *The Web as History: Using Web Archives to Understand the Past and the Present*. London: UCL Press. https://doi.org/10.14324/111.9781911307563.

Bruns, Axel. 2019. “After the ‘APIcalypse’: Social Media Platforms and Their Fight against Critical Scholarly Research.” *Information, Communication & Society* 22 (11): 1544–66. https://doi.org/10.1080/1369118X.2019.1637447.

CCSDS. 2012. “Reference Model for an Open Archival Information System (OAIS).” Recommended Practice. Washington, DC: Consultative Committee for Space Data Systems (CCSDS).

Cook, Terry. 2013. “Evidence, Memory, Identity, and Community: Four Shifting Archival Paradigms.” *Archival Science* 13 (2–3): 95–120. https://doi.org/10.1007/s10502-012-

9180-7.
Dallas, Costis. 2007. “An Agency-Oriented Approach to Digital Curation Theory and Practice.” In *The International Cultural Heritage Informatics Meeting Proceedings*, edited by Jennifer Trant and David Bearman. Toronto: Archives & Museum Informatics. http://www.archimuse.com/ichim07/papers/dallas/dallas.html.
———. 2016. “Digital Curation beyond the ‘Wild Frontier’: A Pragmatic Approach.” *Archival Science* 16 (4): 421–57. https://doi.org/10.1007/s10502-015-9252-6.
———. 2018. “Heritage Encounters on Social Network Sites, and the Affiliative Power of Objects”. In *Culture and Perspective at Times of Crisis: State Structures, Private Initiative and the Public Character of Heritage*, edited by Sophia Antoniadou, Ioannis Poulios, George Vavouranakis, and Pavlina Raouzaiou, 116–31. Oxford: Oxbow Books.
Edwards, Adam, William Housley, Matthew Williams, Luke Sloan, and Malcolm Williams. 2013. “Digital Social Research, Social Media and the Sociological Imagination: Surrogacy, Augmentation and Re-Orientation.” *International Journal of Social Research Methodology* 16 (3): 245–60. https://doi.org/10.1080/13645579.2013.774185.
Engeström, Yrjö. 1999. “Activity Theory and Individual and Social Transformation.” In *Perspectives on Activity Theory*, edited by Yrjö Engeström, Reijo Miettinen, and Raija-Leena Punamäki-Gitai, 19–37. Learning in Doing. Cambridge; New York: Cambridge University Press.
Fiesler, Casey, and Nicholas Proferes. 2018. “’Participant’ Perceptions of Twitter Research Ethics.” *Social Media + Society* 4 (1): 205630511876336. https://doi.org/10.1177/2056305118763366.
franzke, aline shakti, Anja Bechmann, Michael Zimmer, Charles Ess, and Association of Internet Researchers. 2020. “Internet Research: Ethical Guidelines 3.0 Association of Internet Researchers.” Association of Internet Researchers. https://aoir.org/reports/ethics3.pdf.
Garton, Laura, Caroline Haythornthwaite, and Barry Wellman. 1997. “Studying Online Social Networks.” *Journal of Computer-Mediated Communication* 3 (1): 0–0. https://doi.org/10.1111/j.1083-6101.1997.tb00062.x.
Grace, Stephen, Gareth Knight, and Lynne Montague. 2009. “Investigating the Significant Properties of Electronic Content over Time (InSPECT) – Final Report.” London: King’s College London. https://significantproperties.kdl.kcl.ac.uk/inspect-finalreport.pdf.
Hacking, Ian. 1999. *The Social Construction of What?* Cambridge, Mass: Harvard University Press.
Helmond, Anne, and Fernando N. van der Vlist. 2019. “Social Media and Platform Historiography: Challenges and Opportunities.” *TMG–Journal for Media History* 22 (1). https://doi.org/10.18146/tmg.434.
Hemphill, Libby, Susan H. Leonard, and Margaret Hedstrom. 2018. “Developing a Social Media Archive at ICPSR.” In *Proceedings of Web Archiving and Digital Libraries (WADL’18)*. New York: ACM. http://deepblue.lib.umich.edu/handle/2027.42/143185.
Hockx-Yu, Helen. 2014. “Archiving Social Media in the Context of Non-Print Legal Deposit”. In. Lyon. https://library.ifla.org/id/eprint/999/.
Kelpšienė, Ingrida. 2021. “Participatory Heritage: A Multiple-Case Study of Lithuanian Grassroots Cultural Heritage Communities on Facebook.” Doctoral Dissertation, Vilnius, Lithuania: Vilnius University. https://doi.org/10.15388/vu.thesis.181.
Kelpšienė, Ingrida, Donata Armakauskaitė, Viktor Denisenko, Kęstas Kirtiklis, Rimvydas Laužikas, Renata Stonytė, Lina Murinienė, and Costis Dallas. 2023. “Difficult Heritage on Social Network Sites: An Integrative Review.” *New Media & Society* 25 (11): 3137–

64. https://doi.org/10.1177/14614448221122186.
Kietzmann, Jan H., Kristopher Hermkens, Ian P. McCarthy, and Bruno S. Silvestre. 2011. "Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media." *Business Horizons* 54 (3): 241–51.
Kirtiklis, Kęstas, Rimvydas Laužikas, Ingrida Kelpšienė, and Costis Dallas. 2023. "An Ontology of Semiotic Activity and Epistemic Figuration of Heritage, Memory and Identity Practices on Social Network Sites." *SAGE Open* 13 (3): 1–25. https://doi.org/10.1177/21582440231187367.
Littman, Justin, Daniel Chudnov, Daniel Kerchner, Christie Peterson, Yecheng Tan, Rachel Trent, Rajat Vij, and Laura Wrubel. 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.
Lomborg, Stine. 2012. "Researching Communicative Practice: Web Archiving in Qualitative Social Media Research." *Journal of Technology in Human Services* 30 (3–4): 219–31. https://doi.org/10.1080/15228835.2012.744719.
Lomborg, Stine, and Anja Bechmann. 2014. "Using APIs for Data Collection on Social Media." *The Information Society* 30 (4): 256–65. https://doi.org/10.1080/01972243.2014.915276.
Lotman, Juri. 2005. "On the Semiosphere." Translated by Wllma Clark. *Σημειωτική-Sign Systems Studies*, 1: 205–29.
Manca, Stefania. 2020. "Bridging Cultural Studies and Learning Science: An Investigation of Social Media Use for Holocaust Memory and Education in the Digital Age." *Review of Education, Pedagogy, and Cultural Studies* 43 (3). https://www.tandfonline.com/doi/abs/10.1080/10714413.2020.1862582.
Marres, Noortje, and Esther Weltevrede. 2013. "Scraping the Social?: Issues in Live Social Research." *Journal of Cultural Economy* 6 (3): 313–35. https://doi.org/10.1080/17530350.2013.772070.
McKemmish, Sue. 1997. "Yesterday, Today and Tomorrow: A Continuum of Responsibility." In *Proceedings of the Records Management Association of Australia 14th National Convention, 15–17 Sept. 1997*. Perth, Western Australia: RMAA. http://www.infotech.monash.edu.au/research/groups/rcrg/publications/recordscontinuum-smckp2.html.
Michel, Alejandra, Jessica Pranger, Friedel Geeraert, Sven Lieber, Peter Mechant, Eveline Vlassenroot, Sally Chambers, Julie Birkholz, and Fien Messens. 2021. "WP1 Report: An International Review of Social Media Archiving Initiatives." Report. https://orfeo.belnet.be/handle/internal/7741.
Milligan, Ian, Nick Ruest, and Jimmy Lin. 2016. "Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses." In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 107–10. Newark New Jersey USA: ACM. https://doi.org/10.1145/2910896.2910913.
Ogden, Jessica, and Emily Maemura. 2021. "'Go Fish': Conceptualising the Challenges of Engaging National Web Archives for Digital Research." *International Journal of Digital Humanities* 2 (1–3): 43–63. https://doi.org/10.1007/s42803-021-00032-5.
Papacharissi, Zizi, ed. 2011. *A Networked Self: Identity, Community and Culture on Social Network Sites*. New York: Routledge.
Pehlivan, Zeynep, Jérôme Thièvre, and Thomas Drugeon. 2021. "Archiving Social Media: The Case of Twitter." In *The Past Web: Exploring Web Archives*, edited by Daniel Gomes, Elena Demidova, Jane Winters, and Thomas Risse, 43–56. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-63291-5_5.
Perriam, Jessamy, Andreas Birkbak, and Andy Freeman. 2020. "Digital Methods in a Post-API Environment." *International Journal of Social Research Methodology* 23 (3): 277–

90. https://doi.org/10.1080/13645579.2019.1682840.
Robinson, Ian, Jim Webber, and Emil Eifrem. 2015. *Graph Databases: New Opportunities for Connected Data*. O'Reilly Media.
Rogers, Richard. 2013. *Digital Methods*. Cambridge, Massachusetts: The MIT Press.
Rutten, Ellen. 2013. "Why Digital Memory Studies Should Not Overlook Eastern Europe's Memory Wars." In *Memory and Theory in Eastern Europe*, edited by Uilleam Blacker and Alexander Etkind, 219–31. New York: Palgrave Macmillan. http://link.springer.com/chapter/10.1057/9781137322067_11.
Schäfer, Mirko Tobias, and Karin van Es, eds. 2017. *The Datafied Society. Studying Culture through Data*. Amsterdam: Amsterdam University Press. http://en.aup.nl/books/9789462981362-the-datafied-society.html.
Shibuya, Yuya, Andrea Hamm, and Teresa Cerratto Pargman. 2022. "Mapping HCI Research Methods for Studying Social Media Interaction: A Systematic Literature Review." *Computers in Human Behavior* 129: 107131.
Snelson, Chareen L. 2016. "Qualitative and Mixed Methods Social Media Research: A Review of the Literature." *International Journal of Qualitative Methods* 15 (1): 1609406915624574. https://doi.org/10.1177/1609406915624574.
Social Media Lab, Toronto Metropolitan University. 2024. "Social Media Research Toolkit." *Social Media Lab* (blog). January 2024. https://socialmedialab.ca/apps/social-media-research-toolkit-2/.
Stieglitz, Stefan, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. 2018. "Social Media Analytics – Challenges in Topic Discovery, Data Collection, and Data Preparation". *International Journal of Information Management* 39 (Complete): 156–68. https://doi.org/10.1016/j.ijinfomgt.2017.12.002.
Stoycheff, Elizabeth, Juan Liu, Kunto A. Wibowo, and Dominic P. Nanni. 2017. "What Have We Learned about Social Media by Studying Facebook? A Decade in Review." *New Media & Society* 19 (6): 968–80. https://doi.org/10.1177/1461444817695745.
Thomson, Sara Day. 2017. "Preserving Social Media: Applying Principles of Digital Preservation to Social Media Archiving." In *Researchers, Practitioners and Their Use of the Archived Web*, 1–13. London: School of Advanced Study, University of London. https://doi.org/10.14296/resaw.0007.
Thomson, Sara Day, and William Kilbride. 2015. "Preserving Social Media: The Problem of Access." *New Review of Information Networking* 20 (1–2): 261–75. https://doi.org/10.1080/13614576.2015.1114842.
Tufekci, Zeynep. 2017. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. New Haven; London: Yale University Press.
Upward, Frank, Sue McKemmish, and Barbara Reed. 2011. "Archivists and Changing Social and Information Spaces: A Continuum Approach to Recordkeeping and Archiving in Online Cultures." *Archivaria* 72 (January): 197–237.
Vlassenroot, Eveline, Sally Chambers, Emmanuel Di Pretoro, Friedel Geeraert, Gerald Haesendonck, Alejandra Michel, and Peter Mechant. 2019. "Web Archives as a Data Resource for Digital Scholars." *International Journal of Digital Humanities* 1: 85–111.
Vlassenroot, Eveline, Sally Chambers, Sven Lieber, Alejandra Michel, Friedel Geeraert, Jessica Pranger, Julie Birkholz, and Peter Mechant. 2021. "Web-Archiving and Social Media: An Exploratory Analysis." *International Journal of Digital Humanities* 2 (1): 107–28. https://doi.org/10.1007/s42803-021-00036-1.
Voss, Alex, Ilia Lvov, and Sara Day Thomson. 2017. "Data Storage, Curation and Preservation." In *The SAGE Handbook of Social Media Research Methods*, edited by Luke Sloan and Anabel Quan-Haase, 161–76. SAGE Publications Ltd London. https://www.torrossa.com/gs/resourceProxy?an=5018794&publisher=FZ7200#page=190.

Weller, Katrin. 2014. “What Do We Get from Twitter—and What Not? A Close Look at Twitter Research in the Social Sciences.” *Knowledge Organization* 41 (3): 238–48. https://doi.org/10.5771/0943-7444-2014-3-238.

Wilson, Steven Lloyd. 2022. *Social Media as Social Science Data*. 1st ed. Cambridge University Press. https://doi.org/10.1017/9781108677561.

Winters, Jane. 2017. “Coda: Web Archives for Humanities Research–Some Reflections.” In *The Web as History: Using Web Archives to Understand the Past and Present*, edited by Niels Brügger and Ralph Schroeder, 238–48. London: UCL Press.

Zimmer, Michael, and Katharina Kinder-Kurlanda. 2017. *Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts*. New York, NY: Peter Lang.

Conclusion

A Highly transformative age for web archives

Nicola Bingham, Valérie Schafer, Jane Winters, Anat Ben-David

Abstract: This chapter explores the evolving landscape of web archiving. It considers how web archives document challenging times, may help to analyse them, and respond to events, disruptions, social demands, and crises. It examines emergency response practices and research trends. The chapter also addresses current and forthcoming challenges such as adapting to platformization, AI, the closure of APIs, and evolving legal frameworks. It highlights how web archives are dynamic entities intertwined with contemporary socio-technical contexts, continually adapting to navigate the complexities of a highly transformative age.

Keywords: web studies, web archiving, collections, crisis, emergency responses.

Over nearly three decades of web archiving, we have witnessed profound transformations in the landscape: the emergence of different players, practices, and processes; the evolution of aims and goals; the recognition of new challenges; and the blossoming of collaborations between archival institutions and researchers. Web archive studies has begun to take shape as a discrete field of research, related to but distinct from digital humanities (Brügger 2021), data science and platform studies, and is itself beginning to be interrogated critically. Ben-David (2021, 181–82) notes that "The field of web archiving and web archive research is maturing" and consequently, there is now "room for thinking critically about web archives and for rethinking some of their premises".

Web archives have long been viewed as a rich and "important primary source for humanities researchers" among others (Winters 2017, 245), amenable to qualitative and quantitative research alike. There is scope both for a detailed study of the French presence in London (Huc-Hepher 2021) and for a survey of the entire Danish web domain (Nielsen 2021). The organization of many web archives is both shaped by and shapes these micro- and macro-level approaches, with carefully curated special collections complementing the vast, heterogeneous domain crawls undertaken by many national libraries. The lens through which we view web archives is, however, changing. They remain invaluable repositories of information but are increasingly being considered as important cultural heritage artifacts in their own right. The value of even online memes as cultural heritage is evident from a project like the Meme Wall, which was

Nicola Bingham, British Library, United Kingdom, Nicola.Bingham@bl.uk, 0000-0002-5510-9869
Valérie Schafer, University of Luxembourg, Luxembourg, valerie.schafer@uni.lu, 0000-0002-8204-1265
Jane Winters, University College London, United Kingdom, jane.winters@sas.ac.uk, 0000-0001-5502-5887
Anat Ben-David, University of Israel, Israel, anatbd@gmail.com, 0000-0003-4510-5634

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Nicola Bingham, Valérie Schafer, Jane Winters, Anat Ben-David, *Conclusion: A Highly transformative age for web archives* , © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.29, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 343-362, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

created by the Saving Ukrainian Cultural Heritage Online (SUCHO)[1] initiative. The securing of information in the face of conflict was important, but so too was the preservation of digital culture.

Web archiving has also had to respond to and try to keep pace with increasingly rapid change in the broader digital landscape. New web and social media platforms have required the development and combination of different tools and approaches. Social media, in particular, is something of a "moving target", requiring agility and innovation from those who would archive and study it. More sophisticated approaches to documenting and sharing data have also been influential for research and practice in web archive studies, although they are not always easily accommodated. The growing emphasis on data that is Findable, Accessible, Interoperable and Reusable (FAIR) and the influence of open science initiatives have changed the field significantly, but the relative lack of change in the interlocking legal frameworks that govern the harvesting and archiving of the web creates friction.

In this final chapter, and as we are coming to the end of this collective reflection, we would argue that web archives and web archive studies are reaching an inflection point. After a "long process resulting in the consolidation of standards, best practices, shared methods, tools, and knowledge" (Ben David 2021, 182), there is an opportunity to reevaluate web archiving research and practice and to reconsider the relationship between web archives and their contemporary socio-technical contexts. Web archives are not merely static repositories; they are dynamic entities closely entangled with contemporary challenges. These include ethical approaches to the archiving, preservation, and reuse of personal and public data; the environmental impact of digital preservation in the face of a climate crisis; and the requirement to respond swiftly to unforeseen events and crises.

All of these transformations unfold simultaneously within web archiving institutions and in the broader context of changing digital and scientific practice. They affect the objects and subjects of study; methods of data collection and preservation; and the demands and expectations placed on web archives by society. In this chapter, we explore the extent to which web archives are both active participants in and influenced by this highly transformative age, and how they are responding to it. It begins by considering web archives as 'archives of crisis', which play a vital role in recording the traces of natural and man-made crises as they play out in online spaces. It then discusses how web archives are (and are not) in tune with these challenging and febrile times, while also exploring the processes of constant renewal, adaptation, and ultimately transformation that have

[1] https://www.sucho.org. For the Meme Wall, see https://memes.sucho.org

allowed web archives to weather the digital and social storms of the early 21st century. Finally, it identifies current and future challenges that will necessitate continuing adaptation and innovation in web archiving and web archive studies.

1. Web archives responding to events, disruptions, social demands and crises

Web archives serve as snapshots of our online world and attempt to mirror the challenges and crises that unfold. However, they are always filtered through the lens of human decisions, technological limitations, and resource constraints, embedding an inherent subjectivity. In acknowledging disruptions and crises, web archives exhibit a long-standing vigilance and responsiveness, reacting by creating special collections that highlight these pivotal moments.

1.1 Societal expectations and institutional roles

Societal expectations place a significant burden on heritage institutions in addressing crises and collecting information about them. Libraries and archives are seen as custodians of history, responsible for preserving the collective memory of society. Therefore, their role in archiving digital content during crises is crucial in ensuring that these moments are not lost or distorted with time. The public expects these institutions to capture a diverse range of perspectives, ensuring that the archive is representative and reflective of the entirety of an event. This situation was, for instance, heightened during the COVID-19 crisis, when institutions preserved a born-digital record of the pandemic (see for instance "Mapping the archival horizon: A Comprehensive survey of COVID-19 web collections in European GLAM institutions" by Nicola Bingham).

Web archives have always shown sensitivity to disruption and crisis, and web archiving institutions have demonstrated an awareness, and responsiveness to traumatic events that predates the archived web. Recently, however, there has been a noticeable professionalization of emergency response practices within the web archiving community. While the British Library, through the UK Web Archive (UKWA), had already reflected the 2005 terrorist attacks in London through a small, curated collection, the web archiving responses to the Paris terrorist attacks in 2015 were quite different. Ten years after events in London, the technical means, the digital landscape, and the skills and resources available to web archivists had changed, leading the French institutions that preserve the web—BnF (Bibliothèque nationale de France) and Ina (Institut National de l'Audiovisuel)—to collect a vast amount of websites and social media content (i.e., 20 million tweets preserved by Ina on the 13 November terrorist attacks, see Schafer et al. 2019).

This shift towards "living archives" (Rollason-Cass and Reed 2015) is characterized by increased experience, better guidelines, international collaboration, and heightened reactivity. The COVID-19 crisis expedited this professionalization, as evidenced in the next section, emphasizing the need for swift and comprehensive archiving of digital content during unforeseen events. Cultural heritage institutions have also fostered widening participation in relation to web archiving, either through special collections curated by researchers or research teams (as seen in UKWA or the BnF, for example), through calls for participation to the general public, or by involving colleagues during the COVID crisis. For instance, the BnF enriched its COVID collection with more local perspectives thanks to its local correspondents or integrated BnF staff, some of whom were isolated during the lockdown and unable to perform their usual tasks (Gebeil et al. 2020).

In the realm of professionals, international participation and increasingly refined processes within the global context, particularly thanks to the IIPC (International Internet Preservation Consortium), also demonstrate international solidarity coupled with a community of practices.

1.2 Professionalization of emergency response practices

The period encompassing the pandemic from 2020 to 2022 expedited the professionalization trajectory of emergency response practices within the sphere of web archiving. This swift evolution was evidenced by enhanced expertise, heightened responsiveness, the formulation of guiding frameworks, and concerted collaboration among cultural heritage institutions.

The importance placed on preserving online materials related to the pandemic may be illustrated by the increase in data allocation for the IIPC COVID-19 collection[2] from the Internet Archive/Archive-It. An extra two terabytes of data were allocated, free of charge, bringing the total data budget to five terabytes specifically for archiving COVID-19-related web content. This gesture signifies a proactive response to the exceptional volume and critical nature of online information surrounding the global health crisis and underscores a collective effort to ensure comprehensive and robust preservation of international web content for future historical, research, and public informational purposes (Geeraert and Bingham 2020).

Libraries and archives demonstrated multifaceted approaches to fostering collaboration in COVID-19 web archiving activities. Notably, cultural institutions actively forged collaborative alliances, recognizing the inherent complexity of documenting an issue of such magnitude and acknowledging

[2] See the collection at https://archive-it.org/collections/13529

the necessity for collective effort in this endeavor. The Royal Danish Library, for example, was part of a general project documenting coronavirus lockdowns in Denmark in 2020. This effort was a cooperation between several cultural institutions, including the National Archives (Rigsarkivet), the National Museum (Nationalmuseet), the Workers Museum (Arbejdermuseet), and local archives, which resulted in a more comprehensive collection than could be achieved by any one institution on its own (Schostag 2020).

COVID-19 archiving projects accelerated a trend for collecting a diverse range of voices in web archives:

> It seems as though a threatening present and the urge to collect as many voices as possible on the COVID-19 crisis in web archives encourage us to question history as it has been written and discussed in the modern era (Priem and Grosvenor 2022).

The urgency of capturing numerous perspectives amid the COVID-19 crisis prompted a re-examination of historical narratives. The authors suggest that both the digital age and the pandemic fostered an augmented awareness of establishing new dimensions for engaging with and reflecting upon the past.

Chiara Zuanni (2022) makes the point that the pandemic led to a surge in born-digital material collected by museums and memory organizations, posing challenges in collection, preservation, and display. The volume and hybrid nature of these collections (mixing physical and digital objects) necessitated new approaches and technical capabilities, pushing institutions to develop innovative methods.

As mentioned above, institutions also adopted participatory approaches, aiming for more comprehensive and diverse collections, while at the same time trends towards grassroots and bottom-up initiatives are also becoming apparent in the field of web archiving.

1.3 Participation and grassroots efforts

During the Arab Spring, collections were curated by institutions such as the National Library of Tunisia (BnT), the Internet Archive and the Library of Congress. In the case of the BnT, the institution underwent a transformative shift in its collecting approach to web archiving, as detailed by Raja Ben Slama (2023 and her chapter "Web archiving in Tunisia post-2011: The National Library of Tunisia's experience"). The BnT recognized the need for a collective effort involving public institutions, associations, and volunteer researchers to collect and archive scattered digital documents from the web and citizens' mobile phones. This collaboration aimed to capture firsthand accounts and materials from those who participated in the uprisings. It highlighted the necessity of adapting to the changing landscape by creating a web archiving unit outside the traditional legal deposit service

to address the lack of provision for archiving documents from social networks and various creative forms that emerged post-revolution. These challenges were outside the framework of existing legal regulations and required innovative solutions. This suggests an ongoing process of adaptation and growth in response to the dynamic nature of digital content and the challenges of preserving and organizing it effectively, exacerbated by the temporal urgency of responding to crisis events.

Preserving online content during crisis events like terrorist attacks and movements for social justice has spurred advancements in both web archiving technologies and policies guiding collection development. The Documenting the Now project[3] played a significant role in archiving digital content related to the shooting of Michael Brown in Ferguson, Missouri, in 2014. This initiative aimed to capture and preserve the digital traces of social media conversations, images, videos, and online content that emerged in the aftermath of this pivotal event. The project recognized the importance of archiving these digital conversations in real-time. It developed tools and methodologies to capture and preserve Twitter feeds, Facebook posts, Instagram images, and other social media content.

The project also addressed ethical considerations regarding the archiving of sensitive and potentially traumatic material. It engaged with issues of consent, privacy, and the responsible preservation of digital records in a sensitive societal context. In its second phase, Documenting the Now underlined the need for local participation and "digital community-based archives from the perspectives of local activists and in equitable partnership with them" [4].

This trend of non-institutional actors playing a pivotal role in archiving critical events appears to be strengthening. Taking on the challenge of community engagement in web archiving, SUCHO exemplified success by mobilizing volunteers to safeguard Ukraine's online cultural heritage. As the war unfolded, the response was swift, and the project launched on March 1, 2022, successfully bringing together over 1,500 volunteers from more than 38 countries. Coordinating such an initiative required establishing procedures to guide volunteer efforts effectively. According to the 2022 report[5], a metadata team created, for example, "metadata guidelines and video tutorials for direct upload of individual items with metadata to Internet Archive [...]", while also launching a test aimed at volunteers to better adapt these guidelines. It differs somewhat from the guidelines and target audience of the Archive Team, which also issued calls for volunteers and saved numerous online platforms in jeopardy, such as

[3] http://www.docnow.io/
[4] https://archive.mith.umd.edu/mith-2020/documenting-the-now-phase-2/
[5] https://www.sucho.org/assets/Mar-Dec-2022_End_of_Year_Updates.pdf

Geocities or Mobileme, but with a more technical approach[6]. Non-institutional, grassroots efforts in web archiving play a crucial role in capturing diverse perspectives, filling gaps in institutional collections, and ensuring the preservation of digital heritage in a more inclusive and comprehensive manner. As noted by Cui Cui et al. (2023), such efforts often target content that might not be on the radar of larger institutions. They focus on niche topics, local events, or marginalized communities that might not receive adequate attention from traditional archives. This engagement fosters a sense of ownership and empowerment within the community, allowing them to contribute to preserving their history and narratives.

However, challenges persist and not all crises and conflicts receive equal attention or coverage within web archives due to limitations in resources and differing priorities among institutions. Bridging these gaps is crucial to ensure more comprehensive web studies and inclusive representation of our digital history.

2. Web studies in tune with the challenging times

While our initial focus has centered on the dedicated efforts and practices of web archiving, this digital heritage also serves as a valuable resource for researchers. A comprehensive approach to crises and to our highly transformative age necessitates an intrinsic connection to related research endeavors. This interest is exemplified in various chapters of this book, showcasing a sensitivity to controversies, crises, and related memories. The chapters delve into diverse topics, such as "The Words of online hospitality" by Dana Diminescu and Quentin Lobbé; "Mapping the archival horizon: A Comprehensive survey of COVID-19 web collections in European GLAM institutions" by Nicola Bingham; and "Archiving public health discourse in the UK Web Archive" by Alice Austin. Researchers have embraced the challenge of using web archives to enhance the understanding of the first decades of the 21st century.

2.1 Web studies to analyze challenging times

Research programs have emerged as proactive initiatives addressing critical issues. Noteworthy examples include ASAP[7] (Archives Sauvegarde Attentats Paris, funded by the French CNRS), launched in the aftermath of terrorist attacks in 2016, WARCnet's dedicated focus on COVID studies within working group 2[8], and the recent project by Anat Ben-David on the

[6] https://wiki.archiveteam.org/index.php/ArchiveTeam_Warrior

[7] https://asap.hypotheses.org/author/web90

[8] https://cc.au.dk/en/warcnet

history of climate news images using web archives. These research programs, among others, reflect a sensitivity to crises and social transformations. Web archiving collections support these current trends, whether they are constituted by researchers or by web archivists, as demonstrated by those accessible in Archive-It that, for example, capture social movements. From #Metoo and Black Lives Matter to the social upheavals following Fukushima in 2011 and the 'Arab Spring', these collections rely on a collaborative and crowdsourcing model that was established in 2007 after the Virginia Tech campus shooting.

Recognizing the complex nature of crises, there is a growing need for interdisciplinary approaches. While technical and cognitive skills are crucial, recent global challenges demand a broader perspective. Geopolitical, health-related, and social issues associated with recent crises transcend digital expertise, necessitating collaboration among diverse professionals. Initiatives like WARCnet exemplify this interdisciplinary as well as interprofessional model, gathering together web archivists and researchers. It also calls more and more for international research teams. Despite the efforts of a reactive and committed community, challenges persist. The limitations of restricted time and funding pose hurdles to comprehensive archiving and research. Balancing the urgency of capturing the present with sustainable, long-term strategies remains a critical consideration.

2.2 Memories as a key research theme

Web archives are also at the core of some memory studies (see, for example, Clavert 2018 and Gebeil 2016; 2021). This field is increasingly growing in web studies, as evidenced by various chapters in this book, such as "Websites as historical sources? The benefits and limitations of using the websites of former repatriates for the history of schooling in colonial Algeria?' by Christine Mussard, and "A Social media archive for digital memory research" by Costis Dallas and Ingrida Kelpšienė. Studies of socio-digital networks also contribute significantly to this field. Moreover, they intentionally act as memory producers, as noted by Jacobsen and Beer (2022). In the study of social networks, it is worth highlighting the significance of distant reading and the technical and scientific challenges involved, as identified early on by Frédéric Clavert and actively addressed in the CONNECTIVE social media digital archive project presented by Dallas and Kelpšienė in this book. This project accounts for user interactions through posts, comments, and reactions on social media, representing these interactions in a knowledge graph of deposits, agents, threads, and messages representing these social media interactions.

Mentioning knowledge graphs and distant reading, advancements in tools and processes, including but not limited to distant reading and metadata analysis, have to be underlined.

A WARCnet report on *Skills, tools, and knowledge ecologies in web archive research* (Healy et al. 2022) highlighted the sheer range of skills and professional knowledge required to work with web archives, from software and tools, through digital curation processes and workflows to data analysis and web design methods. Researchers and practitioners are required both to acquire a broad base of technical and archival skills and to develop their knowledge and expertise in order to keep pace with the rapidly changing web, digital preservation, and cultural heritage landscapes. Web archiving in national libraries and other memory institutions is generally undertaken by small teams who are facing increasing volumes of work against a background of systemic underinvestment in the cultural heritage sector. In this context, it is difficult to find either the time or the resourcing for effective and sustained programs of training and professional development. With a few notable exceptions, the digital skills training offered to humanities researchers does not encompass born-digital archives. Consequently, they are frequently left to develop their skills independently, learning through trial and error.

Like many of the other challenges identified in this chapter, collaboration and partnership offer an effective means of addressing the skills deficit (see "A network to develop the use of web archives: Three outcomes of the ResPaDon project" by Emmanuelle Bermès et al.). Both WARCnet and RESAW have shown sustained commitment to skills training for web archiving and web archive studies. The IIPC is similarly committed to upskilling web archivists and sharing knowledge between established and new entrants in web archiving. The development of stronger connections with professional bodies in the fields of Digital Humanities, Research Software Engineering and Cultural Heritage, for example the Society of Research Software Engineering, would usefully strengthen these existing networks.

Web researchers and archivists face significant hurdles when creating live web corpora, notably due to the complexity of web scraping tools. Michael Black (2016) emphasizes the lack of a universal solution due to diverse content hosting platforms and the rapid evolution of web language standards. This diversity necessitates tailored tools, posing a challenge for archivists seeking options aligned with their needs amidst commercially oriented or larger research-based offerings. Despite the availability of resources like Voyant and robust programming libraries for text mining novices, effectively using these tools requires additional skills and training

for web researchers and archivists (see for instance the studies "Web archives and hyperlink analyses: The case of videnskab.dk 2009–2022" by Niels Brügger and Katharina Sølling Dahlman, and "Multi-level structure of the First Tuesday communities after the 2000 dot-com crash: A social network analysis of economic actors based on web archives" by Quentin Lobbé).

The development of appropriate tools and infrastructure are essential if researchers and practitioners are to have access to environments in which they can hone and refresh their skills. Initiatives like ARCH, a platform developed by the Archives Unleashed team[9] and then adapted by the Internet Archive "for building research collections, analyzing them computationally, and generating datasets from terabytes and even petabytes of data[10]" have enormous potential to open up web archives for researchers who have limited or no access to technical support in their own institutions. Crucially, training programs, tools, and infrastructure will be required to keep ahead of developments on the live web and help answer issues underlined in chapters of this book, like "Making social media archives: Limitations and archiving practices in the development of representative social media collections" by Beatrice Cannelli and "Challenges in archiving the personalized web" by Erwan Le Merrer, Camilla Penzo, Gilles Tredan, and Lucas Verney. Of course, these challenges go beyond technical issues to encompass ethical and legal considerations.

2.4 Ethical considerations in challenging times

Researchers and web archivists navigate both legal and ethical dimensions in their work. The impact of the GDPR in Europe looms large, particularly concerning the proliferation of personal data in web archives. Research guidelines and policies, exemplified by the shift towards FAIR data, add further layers of complexity. Compliance is not always straightforward, given the limited shareability of web archives content, often constrained by legal deposit frameworks in various countries. Issues related to author rights further complicate matters. However, there is a notable shift towards the shareability of seed lists and metadata, as seen in a project like AWAC2 (Aasman et al. 2021), based on the COVID IIPC collections in Archive-It, which facilitated a comprehensive study of COVID data through distant reading. The use of permalinks also supports more widespread citation. Yet, many challenges persist, especially with vast collections of web archives whose accessibility is limited to library reading rooms. Internet Archive, Arquivo.pt, and projects like SUCHO provide

[9] https://archivesunleashed.org/arch/
[10] https://webservices.archive.org/pages/arch

online access, but the question of reuse remains quite unresolved. Ethical dimensions, explored in the context of Geocities by Ian Milligan, are increasingly pertinent too. He emphasized the ethical dilemma of studying personal pages, asserting that:

> Leaving people out isn't ethical either. Moving to a full opt-in process would likely lead to the historical record being dominated by corporations, celebrities and other powerful people, tech males, and those [who] wanted their public face and history to be seen a particular way (Milligan 2018).

Meghan Dougherty's analysis already highlighted the evolving ethical landscape, shifting from concerns about copyright permissions to a focus on privacy:

> The debate is not simply a matter of whether or not it is ethical to preserve what some users consider to be ephemeral artifacts in permanent and accessible storage. The debate is far more complicated, involving various information behaviors, conflicting expectations, and different interpretations of how our information online represents our most intimate selves (Dougherty 2013).

Ethical debates extend beyond privacy, encompassing issues of inclusiveness, power relations, and potential invisibilization. In a white paper, Documenting the Now highlights tensions such as user awareness, the potential fraudulent use of social media content, and the

> heightened potential for harm to members of marginalized communities using the web and social media, especially when those individuals participate in activities such as protests and other forms of civil disobedience that are traditionally heavily monitored by law enforcement (Jules et al. 2018, 9).

The National Forum on Ethics and Archiving the Web, organized by Rhizome in 2018, further underscored the multifaceted nature of ethical challenges, with panel discussions, like "Archiving Trauma" and "Documenting Hate". Pamela M. Graham (2017) emphasizes the intertwining topics of ethics and diversity. Transparency in the archiving process is considered crucial for creating collections ethically. This article also challenges the black box of web search engine algorithms, referring to Safiya Noble's *Algorithms of oppression*, and pointing out that in web archives, biases should be mitigated rather than perpetuated: by developing "effective search functions, we have the opportunity to offer a very different use experience than what the live web affords".

3. Trends and future challenges

While some challenges have already been underlined, for example the need for constant renewal and adaptation of skills, the requirement to face the rapid changes implemented by platforms and social media networks, the

issue of inclusiveness and public participation, and the need for co-shaping and co-sharing web archives, there are still others that promise to be highly transformative for web archives and practices. Artificial intelligence, misinformation, platformization, and asymmetries are key topics to be addressed to better adapt or consider web archiving in the near future and may become strong drivers of transformation.

3.1 Anticipating the future with artificial intelligence (AI) and machine learning (ML)

AI and ML are reshaping the landscape of web archiving, expanding the types of data captured, diversifying the methods of analysis, and highlighting the challenges of ethical considerations, compliance, and the preservation of diverse perspectives. AI has undoubtedly had a positive impact on web archiving in several areas, for example by enabling the capture and organization of events that might typically fall outside the purview of institutional archives' defined collection policies, or within conventional archival collection policies. This expanded scope enhances the inclusivity and depth of archived content (Sönmez et al. 2016).

In their chapter, Emmanuelle Bermès et al. discuss the ResPaDon project which aims to provide the research community and archivists with methods and tools for the building, analysis, and dissemination of web corpora. To this end, Sciences Po médialab and the BnF organized, ran, and evaluated an experiment based on the use of the Hyphe web crawler on web archives.

The expansion of web archivist and researcher-friendly tools integrating AI and ML are also to be noted, while there arise significant ethical and legal considerations that archivists and web researchers must confront. Black (2016) points out, for example, that while web scraping might not inherently violate intellectual property laws, recent US and EU court cases have scrutinized whether scraping affects data value or causes economic harm to data hosts.

Guidelines are beginning to be developed, for example the OCLC publication, "Guidelines for Libraries' Responsible Use of AI: Responsible Operations", a guide developed in collaboration with professionals from various sectors, presents a framework to address technical, organizational, and social challenges related to the operationalization of data science, ML, and AI in libraries. This agenda highlights seven areas of investigation, providing recommendations to guide discussions and actions toward responsible engagement with these technologies (Padilla 2019).

Lynch (2017) raises concerns about stewardship for algorithms, highlighting that current archival techniques and training might not be sufficient to preserve the perspective of the "age of algorithms" for future understanding. There is a need to adapt archivists' education to capture the impact and implications of algorithms on data capture and processing.

Jaillant and Caputo (2022) highlight significant challenges faced by web archives in utilizing AI. They emphasize the risks associated with algorithmic errors, citing an instance where open-source software flagged innocuous terms incorrectly, leading to false positives. They also address the ethical and social implications of AI-driven decision-making due to the opacity of AI processes and the potential biases within training data. To mitigate these challenges, they advocate for "Explainable AI", which facilitates human comprehension of machine-generated outcomes and stresses the need for interdisciplinary collaborations among archivists, Digital Humanists, and Computer Scientists to navigate these ethical complexities. Recognizing the interdisciplinary nature of these challenges, various networks and initiatives have emerged, such as AURA and AEOLIAN, fostering exchanges among scholars, archivists, librarians, and museum professionals. While AI has gained prominence in various sectors, its application in libraries and archival institutions remains at an experimental stage, necessitating more robust and compelling case studies to drive advancements in this domain.

3.2 Misinformation in a post-truth era

A second challenge, which is already very much with us—"post-truth" was the Oxford Dictionaries international word of the year in 2016—is that of dealing with mis- and disinformation as it enters web and social media archives. Archives have always included misinformation, for example medieval charters that are only identified as forgeries centuries after their creation, but the scale of the challenge when dealing with the archived web is unprecedented. The Archive of Tomorrow project, which ran for 14 months from February 2022, was set up to identify and preserve both online information and misinformation related to public health in the UK, and in particular to the COVID-19 pandemic of 2020 onwards. The project's final report notes that "The need to identify and archive both accurate information as well as the inaccurate is now a pressing societal need" (Archive of Tomorrow 2023), but this is not an easy thing to do. The project identified several practical challenges, from the "question of how to name and describe a collection which was explicitly open to capturing misinformation as well as information" to "Who is responsible for identifying and labeling misinformation in research collections?". The main challenge, however, remains one of how to present and contextualize misinformation such that it can be distinguished from other archived data. This is feasible, although labor intensive, for smaller special collections, but becomes difficult, if not impossible, at the scale of a national web domain. How can what Acker and Chaiet (2020) describe as "The weaponization of web archives", contributing to an online "misinfodemic", be combated?

Metadata and documentation are important tools for web archives, allowing users to investigate provenance and make informed decisions about the accuracy, currency and even reliability of the information they are looking at. Dense metadata and descriptive text run the risk of being ignored or misunderstood, especially when the archived web is itself such a complex source, so the accessible presentation of key contextual information will be crucial in helping researchers and readers to navigate around (mis)information effectively. Expert human curation will remain important, but the use of artificial intelligence to flag problematic content and assist with the creation of metadata is likely to make the task of contextualizing material in large born-digital archives easier. This will not, however, address every kind of misuse, like the "screensampling" identified by Acker and Chaiet (2020), which involves posting screenshots of archived URLs to remove the ability to click on or track these static images of archived online sources.

3.3 Closure of APIs and platformization

Within the framework of IIPC WAC24[11], Frédéric Clavert invited scholars to discuss "Archiving social media in an age of APIcalypse". This discussion arose after two major platforms, Twitter (now X) and Reddit, placed access to their APIs behind a paywall in the early part of 2023. As noted by Clavert, Application Programming Interfaces (APIs) play a crucial role in accessing data and harvesting for archived collections and various research projects. This closure echoes past incidents, such as LinkedIn restricting data access in 2015 and Facebook reducing API functionalities post the Cambridge Analytica scandal, and it is referred to as an "APIcalypse" by Axel Bruns (2019). The closure of APIs has multifaceted repercussions. Notably, it adversely affects collections of the kind that have previously been allowed, for example Clavert's analysis of World War I memories (2018) and Nick Ruest's immediate collection during the Charlie Hebdo attacks[12]. Institutions also leverage APIs, as demonstrated by the INA's creation of extensive collections, some deeply tied to highly transformative events like the Charlie Hebdo attacks and social movements like the protests of the Gilets Jaunes.

Terms of use or access can regularly evolve, necessitating a reconfiguration of collection methods, while consideration must also be given to frequent updates and sometimes unsatisfactory crawls. As noted by Ben Els (National Library of Luxembourg) Facebook actively blocks crawler robots, necessitating the collection of more data for reliable results,

[11] https://netpreserve.org/ga2024/
[12] https://ruebot.net/tags/charliehebdo/

while the cost of capturing Facebook may exceed that of archiving a regular website, with issues such as non-functional videos (Els and Schafer 2020). "The Telegram Archive of the War in Ukraine" serves as a testament to these challenges[13], emphasizing the need for human curation and the constant issues at stake (Holownia and Socha 2022).

The ongoing platformization and the challenges and limits posed by giant web companies to web archiving, are concerning. Loss of functionality and a strong dependence on proprietary platforms are crucial considerations, especially as their influence and usage continue to strengthen, while also touching upon asymmetries and gaps in web archiving.

3.4 Asymmetries and gaps

Although web archives have been characterized by ingenuity, innovation, and responsiveness to both technological and societal change, there remain significant asymmetries in the collectivity of the archived web—an overrepresentation of some voices and experiences and an underrepresentation of others.

Web archiving continues to be an activity primarily in and of the Global North, and this necessarily distorts the digital historical record. The work of the IIPC in organizing global collections, for example the Novel Coronavirus (COVID-19) special collection[14], goes some way to addressing this imbalance. At the time of writing, the country most represented in the COVID-19 collection is the US (1,988 websites), but it is followed by Brazil, Argentina, Peru, Uruguay, Chile, and Bolivia. There are, however, only 85 Chinese websites included in the collection, the first African country to be mentioned is Angola (59 websites) and there are only 15 archived resources for India. The Internet Archive's Whole Earth Web Archive (WEWA), launched in October 2019, was designed as "a proof-of-concept to explore ways to improve access to the archived websites of underrepresented nations around the world". The Internet Archive is committed to "undertaking active outreach to national and heritage institutions in these nations, and to related international organizations, to ensure this work is guided by broader community input" (Bailey 2019), but the WEWA remains a Global North initiative on behalf of nations without their own web archiving infrastructure.

Lor and Britz (2004) have highlighted the complex morality of what they describe as "South-North web archiving", and the need for clarity both about the motivations of archiving institutions and about the balance between the right of access to information and the right to own and control

[13] https://storymaps.arcgis.com/stories/0af72de4b008461bb441fc62fffb9f8d
[14] https://archive-it.org/collections/13529

it. There is much that could be learned from the CARE Principles for Indigenous Data Governance—Collective Benefit, Authority to control, Responsibility, and Ethics—which are not yet part of the mainstream of web archiving research and practice. In the field of Digital Humanities, Quintanilla and Horcasitas (2013) have called for "transnational solidarity", based on "relationships and networks of care that exceed the logic of national boundaries" and which can "lay the groundwork for decolonial and sustainable futures". These are calls to action that web archiving and web archive studies can take up on their own terms.

A particular challenge is the large-scale, top-down, and primarily automated nature of national domain crawls. They are designed to be as comprehensive as possible, but they can be neither complete nor consultative; nor do they systematically include most social media platforms. Consequently, it is at the level of special collections, where careful curation is possible, that gaps and asymmetries in web archives can more easily be addressed. Schafer and Winters (2021) note that "With regards to inclusiveness, there have been some notable efforts to diversify special collections, so that web archives become visibly more inclusive". Ensuring that this is true for the vast web archives of national domains or major global events will, however, remain difficult.

Finally, we cannot ignore a current and future challenge related to sustainability, environmental impacts, resource allocation, and long-term preservation strategies. The environmental cost of AI is increasingly being discussed and there are calls to embrace "digital frugality", and web archives must be involved in these conversations. The experience of collecting born-digital records during the pandemic acted as a catalyst for recognizing the need for more refined and sustainable digital preservation practices. This necessity stimulated demand for the development of appropriate workflows and strategies tailored specifically for born-digital objects. For instance, Blair et al. (2021) covered how to pitch web archiving to funding bodies, how to make appraisal decisions when gathering URL seeds, how to manage crawling within a limited data budget, and tools and techniques for managing this work between several people working remotely. As highlighted by Pendergrass et al. (2019):

> [...] it is time for all cultural heritage professionals who work with digital content to engage with this urgent issue and to critically evaluate current practices in appraisal, permanence, and availability of digital content to create environmentally sustainable digital preservation.

To conclude, as demonstrated during the whole book and in this final chapter, web archives and the efforts of web archivists are vital in documenting and preserving digital material during times of crisis and

disruption. However, inherent biases, resource constraints, and challenges create imperfections and asymmetries in the representation of these events. While increased collaboration within the field is a positive step forward, addressing issues of participation, openness, and resource allocation are crucial in creating more balanced and inclusive web archive collections. Continuities in debates persist alongside obstacles and paradoxes, notably in the realm of author rights, re-use, and shareability. Despite ongoing efforts and notable progress, these issues linger, challenging access and dissemination.

Amidst the call for some change, there is an equally pressing need for stability. Balancing the rapid pace of digital and social changes with the necessity for established practices and research frameworks requires a delicate equilibrium. Shared frames and moments of respite are essential, while the risk of presentism and constant emergency in the face of the numerous crises that arise is to be avoided. The urgency of capturing and documenting the present must be counterbalanced with a nuanced understanding of heritagization and historical context to avoid distortions and oversights in the archived narrative. The allocation of means and resources also emerges as a critical consideration in sustaining effective web archiving initiatives. Striking a balance between social expectations, the demands for innovation and the realities and practicalities of crawling and resource management remains an ongoing challenge.

References

Aasman, Susan, Niels Brügger, Frédéric Clavert, Karin de Wild, Sophie Gebeil, and Valérie Schafer. 2021. "Analysing Web Archives of the Covid-19 Crisis through the IIPC collaborative collection: early findings and further research questions." *International Internet Preservation Consortium Blog*, November 2, 2021. 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/

Acker, Amelia, and Mitch Chaiet. 2020. "The weaponization of web archives: data craft and Covid-19 publics." *Harvard Kennedy School Misinformation Review* 1. https://doi.org/10.37016/mr-2020-41

Archive of Tomorrow. 2023. *Archive of Tomorrow: Capturing Public Health Discourse in the UK Web Archive*. Edinburgh: National Library of Scotland.

Bailey, Jefferson. 2019. "The Whole Earth Web Archive." *Internet Archive Blogs*, October 30, 2019. https://blog.archive.org/2019/10/30/the-whole-earth-web-archive/.

Ben-David, Anat. 2021. "Critical web archive research." In *The Past Web: Exploring Web Archives*, edited by Daniel Gomes, Elena Demidova, Jane Winters, and Thomas Risse, 181– 88. Cham: Springer Nature.

Ben Slama, Raja. 2023. "Web archiving in Tunisia after 2011. Experience of the National Library of Tunisia." Abstract for the RESAW 2023 Conference. https://resaw2023.sciencesconf.org/resource/page/id/14

Black, Michael L. 2016. "The World Wide Web as complex data set: Expanding the Digital Humanities into the twentieth century and beyond through Internet research." *International Journal of Humanities and Arts Computing* 10, 1: 95–109. 10.3366/ijhac.2016.0162

Blair, Lindsey, Claire Drone-Silvers, Denise Rayman, and Rhys Weber. 2021. "Building a COVID-19 Web Archive with Grant Funding." *Midwest Archives Conference Annual Meeting Presentations*. Iowa State University Digital Press. https://www.iastatedigitalpress.com/macmeetings/article/id/12582/

Brügger, Niels. 2021. "Digital humanities and web archives: possible new paths for combining datasets." *International Journal of Digital Humanities* 2, 145–68.

Bruns, Axel. 2019. "After the 'APIcalypse': social media platforms and their fight against critical scholarly research." *Information, Communication & Society* 22, 11: 1544–66. 10.1080/1369118X.2019.1637447

Clavert, Frédéric. 2018. "Temporalités du Centenaire de la Grande Guerre sur Twitter." In

Temps et temporalités du web, edited by Valérie Schafer, 113–34. Nanterre: Presses universitaires de Paris Nanterre.

Cui, Cui, Stephen Pienfield, Andrew Cox, Frank Hopfgartner. 2023. "Participatory Web Archiving: The path towards more inclusive web archives?" Abstract for the RESAW Conference 2023. https://resaw2023.sciencesconf.org/433545

Dougherty, Meghan. 2013. "Property or Privacy? Reconfiguring Ethical Concerns Around Web Archival Research Methods." AOIR Selected Paper, Denver, USA. https://spir.aoir.org/ojs/index.php/spir/article/view/8804/pdf

Els, Ben, and Valérie Schafer. 2020. "Exploring special web archive collections related to COVID-19: The case of the BnL." *WARCnet Papers*. Aarhus: WARCnet.

Gebeil, Sophie. 2021. *Website Story. Histoire, mémoires et archives du web*. Bry-sur-Marne: INA.

Gebeil, Sophie. 2014. "Le web, nouvel espace de mobilisation des mémoires marginales. Les mémoires de l'immigration maghrébine sur l'internet français (2000–2013)." *Cahiers Mémoires et Politique* 2. https://doi.org/10.25518/2295-0311.115

Geeraert, Fridel, and Nicola Bingham. 2020. "Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection, An interview with Nicola Bingham (British Library) conducted by Friedel Geeraert (KBR)." *WARCnet Papers*. Aarhus: WARCnet.

https://cc.au.dk/fileadmin/user_upload/WARCnet/Geeraert_et_al_COVID-19_IIPC__1_.pdf

Graham, Pamela. 2017. "Guest Editorial: Reflections on the Ethics of Web Archiving." *Journal of Archival Organization* 14, 3-4: 103–10. 10.1080/15332748.2018.1517589)

Healy, Sharon, Helena Byrne, Katharina Schmid, Nicola Bingham, Olga Holownia, Michael Kurzmeier, and Robert Jansma. 2022. "Skills, tools, and knowledge ecologies in web archive research." *WARCnet special report*. Aarhus: WARCnet. https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Skills_Tools_and_Knowledge_Ecologies.pdf

Holownia, Olga, and Sacha Kelsey. 2022. "Web Archving the War in Ukraine." *International Internet Preservation Consortium Blog*, July 20, 2022. https://netpreserveblog.wordpress.com/2022/07/20/web-archiving-the-war-in-ukraine/

Huc-Hepher, Saskia. 2021. "Queering the web archive: a xenofeminist approach to gender, function, language and culture in the London French special collection." *Humanities and Social Sciences Communications* 8, 1–15.

https://doi.org/10.1057/s41599-021-00967-8

Jacobsen, Ben, and David Beer. 2021. *Social Media and the Automatic Production of Memory: Classification, Ranking, and Sorting of the Past*. Bristol: Bristol University Press.

Jaillant, Lise, Annalina Caputo. 2022. "Unlocking digital archives: cross-disciplinary perspectives on AI and born-digital data." *AI & Society* 37, 823–35. https://doi.org/10.1007/s00146-021-01367-x

Jules, Bergis, Ed Summers, and Vernon Mitchell. 2018. "Documenting The Now White Paper. Ethical Considerations for Archiving Social Media Content Generated by Contemporary. Social Movements: Challenges, Opportunities, and Recommendations." https://www.docnow.io/docs/docnow-whitepaper-2018.pdf

Lor, Peter, and Johannes Britz. 2004. "A moral perspective on South-North web archiving." *Journal of Information Science* 30, 6: 540–49. https://doi.org/10.1177/0165551504047925

Lynch, Clifford. 2017. "Stewardship in the 'Age of Algorithms'." *First Monday* 22, 12. https://doi.org/10.5210/fm.v22i12.8097

Milligan, Ian. 2018. "The Ethics of Studying Geocities." Ethics and Archiving the Web

Conference. New York: New Museum.
https://ianmilli.wordpress.com/2018/03/27/ethics-and-the-archived-web-presentation-the-ethics-of-studying-geocities/

Nielsen, Jane. 2021. "Quantitative approaches to the Danish Web Archive." In *The Past Web: Exploring Web Archives*, edited by Daniel Gomes, Elena Demidova, Jane Winters, and Thomas Risse, 165–79. Cham: Springer Nature.

Quintanilla, Olivia, and Jeanelle Horcasitas. 2023. "A call to research action: transnational solidarity for digital humanists." In: *Debates in the Digital Humanities* 2023, edited by Matthew Gold, and Lauren Klein. Minneapolis: University of Minnesota Press.

Padilla, Thomas. 2019. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. Dublin, OH: OCLC Research. https://doi.org/10.25333/xk7z-9g97.

Pendergrass, Keith, Walker Sampson, Tim Walsh, and Laura Alagna. 2019. "Toward Environmentally Sustainable Digital Preservation." *The American Archivist* 82, 1: 165–206.

Priem, Karin, and Ian Grosvenor. 2022. "Future Pasts: Web Archives and Public History as Challenges for Historians of Education in Times of COVID-19." In *Exhibiting the Past. Public Histories of Education,* edited by Frederik Herman, Sjaak Braster, and Maria del Mar del Pozo Andrés, 177–96. Berlin, Boston: De Gruyter Oldenbourg. https://doi.org/10.1515/9783110719871-009

Rollason-Cass, Sylvie, and Scott Reed, "Living Movements, Living Archives: Selecting and Archiving Web Content During Times of Social Unrest." *New Review of Information Networking* 2, 1–2: 241–47.

Ruest, Nick, Jimmy Lin, Ian Milligan, and Sam Fritz. 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* (JCDL '20), 157–166. New York: Association for Computing Machinery. https://doi.org/10.1145/3383583.3398513

Schafer, Valérie, Gérome Truc, Romain Badouard, Lucien Castex, and Francesca Musiani. 2019. "Paris and Nice terrorist attacks: Exploring Twitter and web archives." *Media, War & Conflict* 12, 2: 153–70. https://doi.org/10.1177/1750635219839382

Schafer, Valérie, and Ben Els. 2020. "Exploring special web archive collections related to COVID-19: The case of the BnL An interview with Ben Els (BnL) conducted by Valérie Schafer (C^2DH, University of Luxembourg)." *WARCnet Papers*. Aarhus: WARCnet https://cc.au.dk/fileadmin/user_upload/WARCnet/Schafer_et_al_COVID-19_BnL.pdf

Schafer, Valérie, and Jane Winters. 2021. "The values of web archives." *International Journal of Digital Humanities* 2, 129–44.

Schostag, Sabine. 2020. "The Danish Coronavirus web collection – Coronavirus on the curators' minds." *International Internet Preservation Consortium Blog,* July 29, 2020.
https://netpreserveblog.wordpress.com/2020/07/29/the-danish-coronavirus-web-collection/

Sönmez, Çağıl, Arzucan, Özgür, and Yörük Erdem. 2016. "Towards building a political protest database to explain changes in the welfare state." *10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.* https://doi.org/10.18653/v1/W16-2113

Winters, Jane. 2017. "Coda: Web archives for humanities research – some reflections." In *The Web as History: Using Web Archives to Understand the Past and Present*, edited by Niels Brügger, and Raph Schroeder, 238–48. London: UCL Press.

Zuanni, Chiara. 2022. "Contemporary Collecting in a Pandemic: Challenges and Solutions for Documenting the COVID-19 Pandemic in Memory Organizations." *Heritage* 5, 4: 3616–27. 10.3390/heritage5040188

Given recent global crises, the imperative to preserve and analyze online content has never been more vital to enhancing our comprehension of contemporary changes. This book, the outcome of the 5th international RESAW conference that convened experts from fifty disciplines across seventeen countries in Marseille in June 2023, tackles the multifaceted challenges of web archiving. It underscores the dual roles of web archiving, as cultural heritage and as essential source material for researchers delving into contemporary events and the evolution of digital culture. Through twenty chapters, it explores the development of web archiving and examines how technical, cultural, geopolitical, societal, and environmental shifts impact its conception, study, and dissemination.

Sophie Gebeil is a lecturer in contemporary history at Aix-Marseille University (Telemme laboratory). Her research focuses on the study of memorial practices, using web archives as a historical source. She is a member of the Institut Universitaire de France, from 2023 to 2028.

Jean-Christophe Peyssard is head of the Multimedia Library at the Maison méditerranéenne des sciences de l'Homme (MMSH). As a research engineer specialized in MENA Studies at the CNRS, his fields of expertise cover scholarly communication, scientific publishing, library science, digital humanities, and web archives.

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 979-12-215-0412-5 (Print)
ISBN 979-12-215-0413-2 (PDF)
ISBN 979-12-215-0414-9 (XML)
DOI 10.36253/979-12-215-0413-2
www.fupress.com