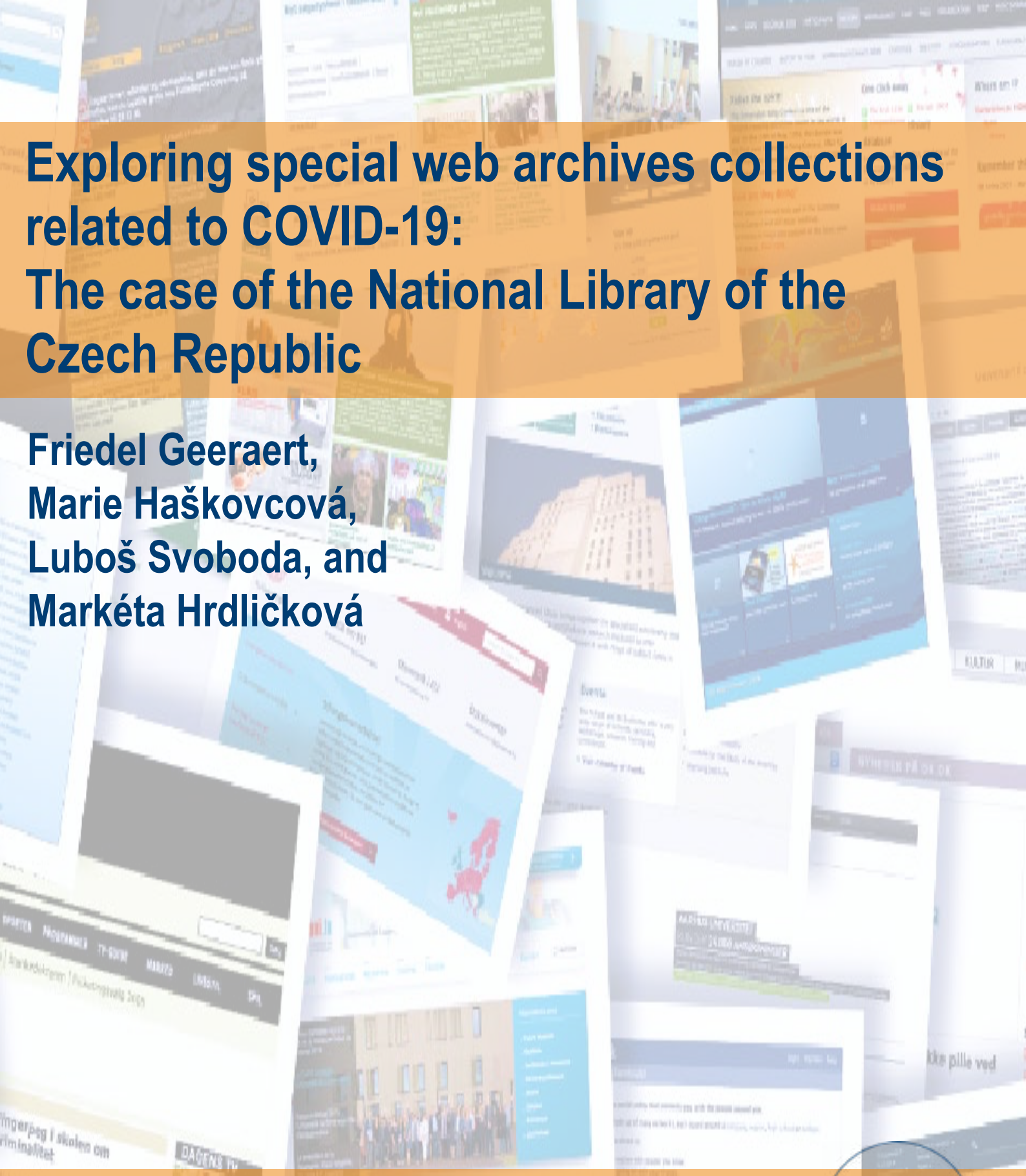
Exploring special web archives collections related to COVID-19: The case of the National Library of the Czech Republic

Friedel Geeraert,
Marie Haškovcová,
Luboš Svoboda, and
Markéta Hrdličková



WARCNET PAPERS

WARCnet
web archive studies

Exploring special web archives collections related to COVID-19: The case of the National Library of the Czech Republic

An interview with Marie Haškovcová, Luboš Svoboda and Markéta Hrdličková (National Library of the Czech Republic) conducted by Friedel Geeraert (KBR)

friedel.geeraert@kbr.be

WARCnet Papers
Aarhus, Denmark 2023

WARCnet Papers ISSN 2597-0615.

Friedel Geeraert, Marie Haškovcová, Luboš Svoboda and Markéta Hrdličková: *Exploring special web archives collections related to COVID-19: The case of the National Library of the Czech Republic*

Published by the research network WARCnet, Aarhus, 2023.

Cover design: Julie Brøndum
ISBN: 978-87-94108-15-7

WARCnet
Department of Media and Journalism Studies
School of Communication and Culture
Aarhus University
Helsingforsgade 14
8200 Aarhus N
Denmark
warcnet.eu

The WARCnet network is funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

DANMARKS FRIE FORSKNINGSFOND
INDEPENDENT RESEARCH FUND DENMARK

WARCnet Papers

Niels Brügger: *Welcome to WARCnet* (May 2020)

Ian Milligan: *You shouldn't Need to be a Web Historian to Use Web Archives* (Aug 2020)

Valérie Schafer and Ben Els: *Exploring special web archive collections related to COVID-19: The case of the BnL* (Sep 2020)

Niels Brügger, Anders Klindt Myrvoll, Sabine Schostag, and Stephen Hunt: *Exploring special web archive collections related to COVID-19: The case Netarkivet* (Oct 2020)

Friedel Geeraert and Nicola Bingham: *Exploring special web archives collections related to COVID-19: The case of the UK Web Archive* (Nov 2020)

Friedel Geeraert and Barbara Signori: *Exploring special web archives collections related to COVID-19: The case of the Swiss National Library* (Nov 2020)

Matthew S. Weber: *Web Archives: A Critical Method for the Future of Digital Research* (Nov 2020)

Niels Brügger: *The WARCnet network: The first year* (Jan 2021)

Susan Aasman, Nicola Bingham, Niels Brügger, Karin de Wild, Sophie Gebeil and Valérie Schafer: *Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections* (Dec 2021)

Olga Holownia, Friedel Geeraert, Abbie Grotke, Jennifer Harbster and Gulnar Nagashybayeva: *Exploring special web archives collections related to COVID-19: The case of the Library of Congress* (Feb 2022)

Niels Brügger: *The WARCnet network: The second year* (Dec 2022)

Michael Kurzmeier: *Using a national web archive for the study of web defacements? A case-study approach* (Aug 2023)

Helle Strandgaard Jensen: *Any Teletubbies Caught in the Web?* (Aug 2023)

Valérie Schafer, Jérôme Thièvre and Boris Blanckemane: *Exploring special web archives collections related to COVID-19: The case of INA* (Aug 2020)

Niels Brügger, Valérie Schafer, Jane Winters (Eds.): *Perspectives on web archive studies: Taking stock, new ideas, next steps* (Sep 2020)

Friedel Geeraert and Márton Németh: *Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary* (Oct 2020)

Friedel Geeraert and Nicola Bingham: *Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection* (Nov 2020)

Caroline Nyvang and Kristjana Mjöll Jónsdóttir Hjörvar: *Exploring special web archives collections related to COVID-19: The Case of the Icelandic web archive* (Nov 2020)

Sophie Gebeil, Valérie Schafer, David Benoist, Alexandre Faye and Pascal Tanesie: *Exploring special web archive collections related to COVID-19: The case of the French National Library (BnF)* (Dec 2020)

Karin de Wild, Ismini Kyritsis, Kees Teszelszky and Peter de Bode: *Exploring special web archive collections related to COVID-19: The Dutch Web archive (KB)* (Nov 2021)

Michael Kurzmeier, Joanna Finegan and Maria Ryan: *Exploring special web archives collections related to COVID-19: The National Library of Ireland* (Feb 2022)

Emily Maemura: *Towards an Infrastructural Description of Archived Web Data* (May 2022)

Olga Holownia, Friedel Geeraert and Paul Koerbin: *Exploring special web archives collections related to COVID-19: The case of the National Library of Australia* (Dec 2022)

Helena Byrne, Beatrice Cannelli, Carmen Noguera, Michael Kurzmeier, Karin de Wild: *Looking ahead: after web (archives)?* (Aug 2023)

Friedel Geeraert, Marie Haškovcová, Luboš Svoboda and Markéta Hrdličková: *Exploring special web archives collections related to COVID-19: The case of the National Library of the Czech Republic* (Aug 2023)

WARCnet Special Reports

Sharon Healy, Helena Byrne, Katharina Schmid, Nicola Bingham, Olga Holownia, Michael Kurzmeier, Robert Jansma: *Skills, Tools, and Knowledge Ecologies in Web Archive Research* (August 2022)

All WARCnet Papers and WARCnet Special Reports can be downloaded for free from the project website warcnet.eu.

Exploring special web archives collections related to COVID-19: The case of the National Library of the Czech Republic

An interview with Marie Haškovcová, Luboš Svoboda and Markéta Hrdličková (National Library of the Czech Republic) conducted by Friedel Geeraert (KBR)

Abstract: This WARCnet paper is part of a series of interviews with web archivists who have been involved in special collections related to COVID-19. The aim of the series is to provide a general overview of COVID-19 web archives.

Keywords: web archives, social networks, COVID-19, special collections, National Library of the Czech Republic

This WARCnet paper is part of a series of interviews with web archivists who have been involved in special collections related to COVID-19. The interview was conducted in writing on 26 June 2023 with Marie Haškovcová (Head of the Web Archiving Department), Luboš Svoboda (Curator and Social Media researcher) and Markéta Hrdličková (Curator of the websites) at the National Library of the Czech Republic.

The idea to create a Czech web archive was conceived in the late 1990s. In the year 2000, Webarchiv was founded as a project of the National Library of the Czech Republic, the Moravian Library in Brno, and the Masaryk University. Webarchiv is now a stable part of the National Library of the Czech Republic. The first websites were archived in 2001, and regular harvesting began in 2005. The Czech Webarchiv currently comprises approx. 560 TB of data.

The Library's collection policy includes several acquisition strategies. Thanks to a cooperation with the Czech domain provider CZ.NIC we can realise comprehensive harvests; second-order domains of *.cz are harvested once or twice a year (totalling about 1.4 million URLs). This collection captures an image of the Czech Internet at a given time.

Selected valuable resources are harvested more frequently and in-depth (selective harvests). The collection comprises resources with cultural, historical and research value across all social themes. The main aim of this collection is to create a representative sample

of Czech cultural heritage that is published on the web. The collection is created in accordance with the National Library´s acquisition strategy fund and uses the Conspectus method which is a classification system for sorting documents or knowledge. The frequency of harvesting depends on the nature of the resource, how often it is updated, how extensive it is, etc.

Resources focused on specific events or topics are collected through topical harvests. These harvests focus on the impact of a specific event on the web. Archiving of these resources is carried out once, or repeatedly, depending on the purpose and duration of the event. Thematic harvests are carried out because of the need for a deeper capture of the imprint of the given topic on the web.

Webarchiv also accepts suggestions for resources to be included in selective harvests via a form on the website. Publishers, webmasters or users can alert the web archive to interesting pages using the form on the website of Webarchiv.

Czech legislation allows the National Library to make copies of works for archiving and conservation purposes, but does not allow them to be made available elsewhere other than in the National Library building. The national legal deposit legislation does not include web content yet. Proposals to change the legal regulations are being discussed within the legislative process. Web archiving practices, therefore, follow the legislation which allows the reproduction of the work (web pages) solely for archiving purposes. Online access outside the library is only available to resources that are licensed under a Creative Commons license or for which an agreement with the publisher was reached. The entire web archive is available to the public on several terminals in the Library. In this way resources from comprehensive and thematic harvests, in particular, are accessible, but also resources selected as part of selective harvests that have not been covered by a licence. Records of web archive resources within the selective harvests are available in the catalogue of the National Library.

The reasons for the special collection

Why did you create a special COVID-19 collection?

National Library of the Czech Republic: COVID-19 affected the entirety of society intensively and in an unprecedented manner, which is why the Czech Webarchiv tried to capture the events on the web related to the spread and impact of the virus in a special thematic collection, which can serve as a unique dataset for research in the future.

The scope of the COVID-19 collection

What exactly did you collect? Websites, social media? Which specific platforms, hashtags, profiles or languages?

National Library of the Czech Republic: We tried to have a broad coverage of the topic - on the web as well as on social media. The collection consists of government websites,

professional scientific and popularising resources, media reviews, pages of volunteer activities and civic initiatives. It includes resources examining economic, legal and social impacts, statistics, mathematical models, or also artistic reflections or disinformation that have appeared in connection with COVID-19.
We collected Facebook (specific search queries, profiles, groups), Twitter (hashtags, specific search queries, profiles), Instagram (specific profiles), YouTube (specific videos), TikTok (although it was unsuccessful due to TikTok's shadow ban). At the same time, we harvest about 800 journalistic Twitter accounts on a regular basis.
So, we knew that the COVID-19 theme wouldn't escape us, even if it didn't make it directly into the collection. We think we have yet to work out technically how to link the collections back somehow in an automated way and how to mine or cluster a topic from already collected regular collections.

Could you provide more information with regard to the amount of data collected and the nature of the collected data?

National Library of the Czech Republic: So far, we have collected about 20 TB of data and the harvests are still running. Due to the exceptional situation that developed rapidly, we started harvesting resources daily on March 12, 2020, with the intention of capturing all changes that occurred on selected sources. It was the first collection we started harvesting on a daily basis, and also the first time that automated procedures were used meaning that the IT operator did not start the harvest manually. We did not assign descriptive metadata to individual sources, but rather to the entire collection.

How did you go about archiving on a national level about an event that is fundamentally global?

National Library of the Czech Republic: Webarchiv, in accordance with the legislation, focuses on archiving Czech resources. This covers content created by authors originally from the Czech Republic, written in the Czech language, dealing with Czech content or published in the Czech territory. Curators who created the collection focused on resources that met at least one of these criteria. Certain advantages can be seen in the linguistic specificity of Czech, but it is only one of the criteria mentioned above. Many Bohemian documents may not be published in the Czech language.

A number of Czech sources naturally refer to foreign content. We focus on a number of other topics that are global, such as climate change or artificial intelligence, GitHub, and Web3. We believe that national content cannot be strictly defined for global topics.
In the case of social media harvesting, we are using advanced filtering by language, especially on Twitter, which is the easier case. On the other networks, we had to rely on #kovid19 hashtag, spelled with K.

The frame of this special collection

When did you start collecting and when did/do you plan to stop? What was the capture frequency?

National Library of the Czech Republic: We started collecting resources reflecting the topic of the COVID-19 pandemic in February 2020. The first harvest took place on February 24 2020. We started harvesting resources daily on March 12, 2020. In accordance with the development of the situation, we operatively changed the frequency of harvesting (from multiple times a day to a monthly frequency).

When to stop the collection is an interesting question, because only now is the web archive beginning to include content reflecting what the COVID-19 period and the resulting isolation, actually caused. These evaluations are only now beginning to appear. Diseases such as post-COVID-19 syndrome, or the connection with the economic crisis, the psychological consequences, etc. are an important aspect of the COVID-19 collection. It's hard to determine the end of COVID-19. We think the pragmatic part of it - the politics of disk space - will ultimately decide it for us.

How did you carry out quality control on the collection (if applicable)?

National Library of the Czech Republic: The quality assurance of the sources was carried out by the curators (a visual comparison between the live web and the archived version of the websites) and was approached by random selection within limited extent.

Did you encounter any issues, challenges, or limits related to the collecting activity?

National Library of the Czech Republic: We solved how to set the collection policy by testing harvesting on a daily basis. In the early days of the collection there were technical problems related to the interruption of the harvest and we solved the limitation related to the space on the storage.

Harvesting social media, brought a lot of technical challenges with it because it is a new activity for us and we don't have a solid infrastructure yet for our daily practices. The curator is also the operator, administrator, quality checker and developer. Social media platforms also make it difficult to crawl content by blocking accounts, using pop-ups and captcha mechanisms, and by changing their environments frequently. Most striking is TikTok, which simply gives an undocumented error without any explanation. We also have problems with YouTube (Google account), which finds our Browsertrix software browser untrustworthy so we cannot get our crawler to log in and successfully harvest the content e.g. whole YouTube profiles.

Accessibility and searchability

How can users access and search in this collection?

National Library of the Czech Republic: Due to legislative restrictions, the entire collection is only available in the building of the National Library. Only licensed resources are freely available, meaning resources for which we came to an agreement (contract) with the publisher or resources with a Creative Commons licence. The COVID-19 collection is available on the Webarchiv website.

Have researchers already expressed interest in using the COVID-19 collection?

National Library of the Czech Republic: For now, we have focused on creating a relevant information resource for future research. Some researchers find our data interesting, but no real research has yet been done on our data. We believe that the collection we are building will be a valuable dataset for future research.

Last year, we prepared a survey aimed at the academic sphere's awareness of web archiving. The name of the survey was "Do you use Webarchiv? Survey of the needs of users of the Czech web archive". For example, we tried to find out what tools and formats researchers use, what type of data and how they would like to use it, or whether they use archived versions of the website in citations of online sources. Respondents usually did not have a concrete idea of what the dataset for their further research should look like, or what data format they would like to work with, or what metadata they would need. Therefore, it is important to raise awareness about web archive data, develop user-friendly interfaces and cooperate with researchers. There are still many questions about how to use archived data.

How do you communicate about this special collection?

National Library of the Czech Republic: We promoted the collection focused on COVID-19 through social networks, articles and at conferences.

Did you have any partnerships with local stakeholders (Archive-It, IIPC, etc.) during the collection process?

National Library of the Czech Republic: In order to create a valuable collection of resources we also addressed the public, experts and librarians with a request to send suggestions for resources to archive via a special form. We promoted this through posts on social networks and mailing lists. This way we obtained about 30 suggestions for web resources to archive.

The Czech web archive also contributed a selection of Czech sources to the international Novel Coronavirus (COVID-19) outbreak collection initiated by the IIPC.

REFERENCES

Haškovcová, M., Svoboda, L. & Hrdličková, M. (2022) Používáte Webarchiv? Průzkum potřeb uživatelů českého webového archive. ProInflow, 14(1–2). https://doi.org/10.5817/ProIn2022-2-2.

National Library of the Czech Republic. *Webarchiv (the Czech Web Archive).* Retrieved from https://www.webarchiv.cz/.

National Library of the Czech Republic. *COVID-19*. Retrieved from https://www.webarchiv.cz/cs/tematicke-kolekce/covid-19.

National Library of the Czech Republic. *COVID-19. Nechte se Webarchivovat!* Retrieved from https://www.webarchiv.cz/cs/pridat-web.

We would like to thank Nicola Bingham (British Library) for her help in proofreading this interview.

WARCnet Papers is a series of papers related to the activities of the WARCnet network. WARCnet Papers publishes keynotes, interviews, round table discussions, presentations, extended minutes, reports, white papers, status reports, and similar. To ensure the relevance of the publications, WARCnet Papers strives to publish with a rapid turnover. The WARCnet Papers series is edited by Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster and Michael Kurzmeier. In cases where a WARCnet Paper has gone through a process of single blind review, this is mentioned in the individual publication.

The aim of the WARCnet network is to promote high-quality national and transnational research that will help us to understand the history of (trans)national web domains and of transnational events on the web, drawing on the increasingly important digital cultural heritage held in national web archives. The network activities run in 2020-22, hosted by the School of Communication and Culture at Aarhus University, and are funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

WARCNET PAPERS

INDEPENDENT RESEARCH FUND DENMARK

warcnet.eu warcnet@cc.au.dk twitter: @WARC_net facebook: WARCnet
youtube: WARCnet Web Archive Studies slideshare: WARCnetWebArchiveStu