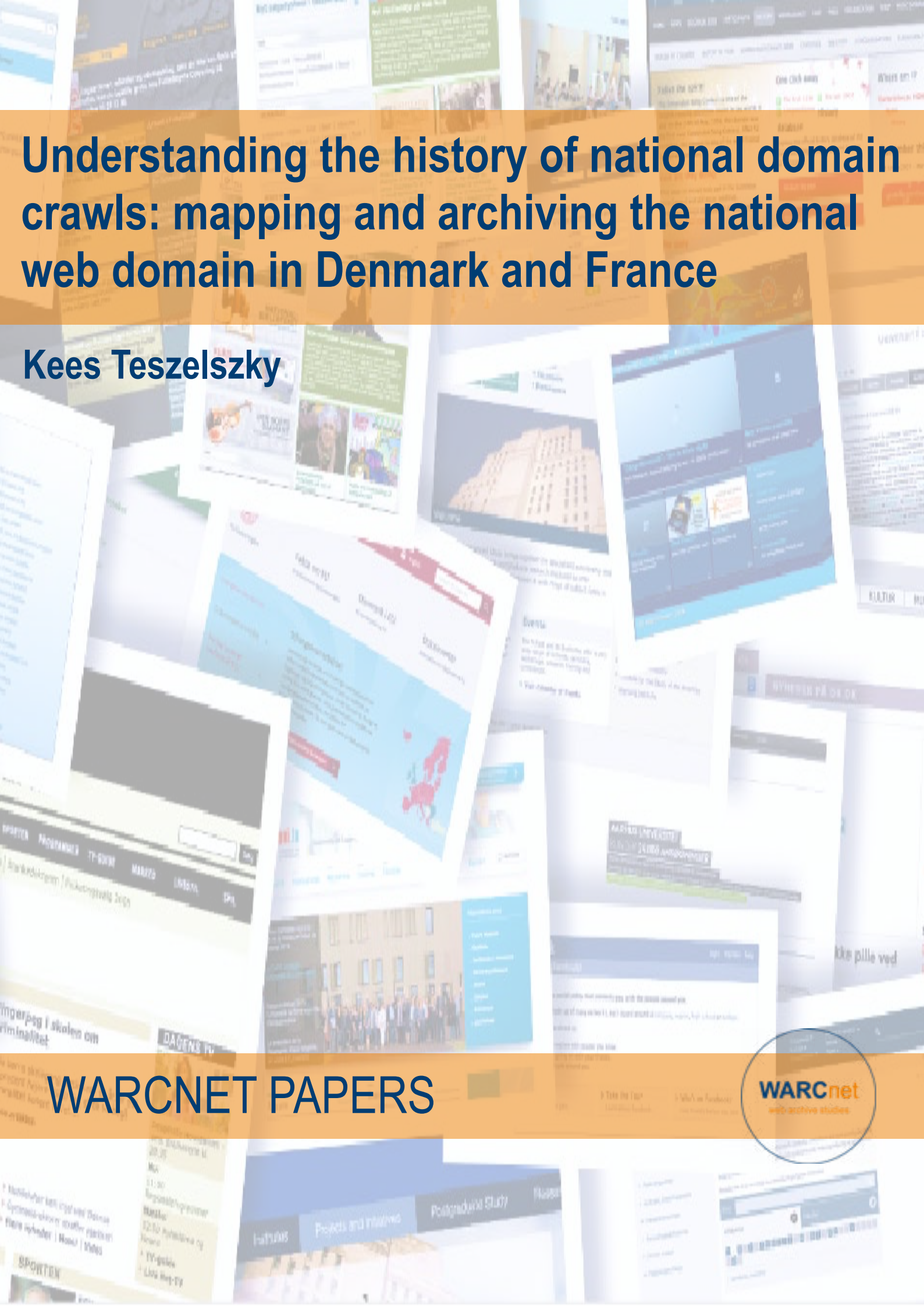
Understanding the history of national domain crawls: mapping and archiving the national web domain in Denmark and France

Kees Teszelszky



WARCNET PAPERS

WARCnet
web archive studies

Understanding the history of national domain crawls: mapping and archiving the national web domain in Denmark and France

Kees Teszelszky
Kees.Teszelszky@KB.nl

WARCnet Papers
Aarhus, Denmark 2023

WARCnet Papers ISSN 2597-0615.

Kees Teszelszky: *Understanding the history of national domain crawls: mapping and archiving the national web domain in Denmark and France*

Published by the research network WARCnet, Aarhus, 2023.

Cover design: Julie Brøndum
ISBN: 978-87-94108-20-1

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).

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)

Niels Brügger: *The WARCnet network: The third year* (Aug 2023)

Kees Teszelszky: *Understanding the history of national domain crawls: mapping and archiving the national web domain in Denmark and France* (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)

Niels Brügger: *Why onsite meetings are important: Reporting from five Short-Term Network Stays* (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* (Aug 2022)

Sharon Healy, Helena Byrne, Katharina Schmid, Juan-José Boté-Vericad, Lanna Floody: *Towards a Glossary for Web Archive Research: Version 1.0* (Aug 2023)

Sharon Healy, Helena Byrne: *Scholarly Use of Web Archives Across Ireland: The Past, Present & Future(s)* (Aug 2023)

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

Understanding the history of national domain crawls: mapping and archiving the national web domain in Denmark and France

Kees Teszelszky

Interviews with Anders Klindt Myrvoll (Det Kgl. Bibliotek/Royal Danish Library) and Vladimir Tybin (Bibliothèque nationale de France) conducted by Kees Teszelszky, KB, National Library of The Netherlands

Abstract: *If a researcher wants to study national web domains in web archives from a comparative perspective, they will head for national libraries and other national heritage institutions which map and crawl the national web domain of a country. Yet, it is very hard to compare different archived national web domains, as the various archive strategies and the legal, technical and cultural backgrounds of the heritage institutions differ from each other. This report explored the differences between the national web archives of France and Denmark and the domains crawls of France and Denmark by comparing the legal deposit laws of the different countries, how the law was put into practice and shaped the library policies, practices and results of BnF and KB Denmark. Following the work of Emily Maemura (2022) it focuses on contextualising the differences between these national collections and the sociotechnical systems that generate data from the national web domain and preserve this data.*

Keywords: web archives, legal deposit law, libraries

INTRODUCTION

If we want to write the history of the worldwide web in the future, we need to do high-quality national and transnational research now that will help us write the history of (trans)national web domains and of transnational events on the web. This research will draw on the increasingly important born digital cultural heritage preserved in national web archives, like national libraries, national archives and other large institutions which are set up by states to safeguard the heritage of the nation state. The data from a domain crawl, in which as much content of the national web

domain as possible is collected for future use by a designated national heritage institution, is the main source for doing such research.

But how can the digital culture of a time-bound concept like the nation-state be preserved for eternity? National libraries and other national heritage institutions that preserve digital born heritage for the future often have the (legal) task of exploring the boundaries of the national web domain and taking at least an annual snapshot of the national web domain by doing or commissioning a domain crawl. Nevertheless, a 'national web domain' is not the same as a 'domain crawl of the national domain' and certainly not the same as the 'archived version of the national web domain'. Each national heritage institution define, crawl and sometimes even preserve their own national web domain in a different way. Consequently, each national library or heritage institution has its own national domain crawl data collection with very specific local characteristics that differ from other national web archives.

If we want to conduct high-quality national and transnational research in the future that will help us to understand the history of (trans)national web domains and of transnational events on the web, then we have to map, study and understand these characteristics and differences between the laws, policies, practices and collections of the various national institutions. This information can be found in the text of the laws, policy papers, blog posts and other written sources which hopefully will be available for future research. How the rules are interpreted and put into practice depends on the people who have to work with them on a daily base and the technical, legal, personnel and financial frameworks within which they have to perform their duties.

The goal of my research is to map, study and analyse these difference between the national web archives of France and Denmark and the domains crawls of France and Denmark by comparing the legal deposit laws of the different countries, how the law was put into practice and shaped the library policies, practices and results of BnF and KB Denmark. Following the work of Emily Maemura (2022) I will focus on contextualising the differences between these national collections and the sociotechnical systems that generate data from the national web domain and preserve this data. The differences between the content of national web archives has a variety of causes: the history of heritage preservation in a given country, the size and nature of the national web domain, the legal basis for collecting national digital born heritage, the policy of the national heritage institution regarding the delimitation of the national web domain, the policy of the national heritage institution regarding the task view on the national domain crawl, the size of the staff and its technical and content expertise, the budget for doing web archiving or a domain crawl, the availability of technology and the possibilities of storage and availability. There are therefore a number of reasons for these differences between the content of national web archives:

1. Existing tradition of preserving national heritage and collecting born digital content;
2. Existing legal deposit law regarding born digital heritage (or the lack of such a law);
3. Policy of the heritage institute regarding collecting born digital material;
4. Used technique and technical possibilities, including outsourcing of the crawls;
5. Practices (organization, staff, budget, expertise);
6. Cooperation with academic research institutions or academics and other heritage organisations during preparation;
7. Use of the crawl data by academic research institutions or academics.

In the next section, I will explain these points in more detail and explain how I will examine them across national libraries.

1. Existing tradition of preserving national heritage and collecting born digital content

Each country has its own history of heritage preservation. Firstly, this has to do with state and nation building and the legitimation of ruler or state power that took place from the Middle Ages onwards. Older nation states such as France, Denmark and the UK, for example, have a long tradition of preserving printed information that is rooted in older state forms such as the monarchy and continued with the creation of the nation-state in the 19th century and led to the formation of national libraries or archives. This tradition is also continued in modern times by making the preservation of digital publications on the web part of the task of a national heritage institution. Younger nation-states lack this tradition, but there the need to underpin the nation-idea is often strong and therefore the need to preserve born digital heritage is felt also.

It is difficult to define the boundaries of a national web domain for a heritage institute based on collecting all publications from those who belong to the definition of a nation state or people. The web is a technical construct formed from an infrastructure and bits and bytes, leads a virtual existence and is only physically present in the analogue world because of data centres where the data is stored and data cables through which the information is transported. The web itself consists of number strings referring to web addresses. These addresses are divided into addresses with generic TLDs (Top-Level Domain) such as .com and .org and ccTLD (country code Top-Level Domains). Applying for a web address with the former can be done by anyone in the world, but applying for an address with a country code may sometimes be subject to specific legal requirements of the relevant state. For example, a web address may only be registered by a natural person who is a resident of a particular country or a company registered there. Nevertheless, unravelling the worldwide web to 'national' parts is a difficult and complex issue that scholars and web archivists alike ponder, e.g. geolocation of web addresses, language used on a site, name or origin of the owner or the appearance of a particular physical postal address.

2. Existing legal deposit law regarding born digital heritage (or the lack of such a law)

The legal basis for collecting digital born heritage is rooted in the national legislation of individual countries. This legislation, too, usually has deep roots in the country's (publication) culture, society and history. Most states around the world have legal deposit legislation that covers physical publications, according to which every publisher of a paper print with certain characteristics is obliged to transfer one or more copies to one or sometimes even several heritage institutions in the country. This legislation is usually extended to the digital domain of the web and translated into the right of heritage institutions to preserve at least public material of the national domain.

3. Policy of the heritage institute regarding collecting born digital material

In most cases, the legal deposit law gives only global directions on how, when, where and what should be collected by which heritage institution. What exactly the national web domain contains, how the national web domain should be delineated, when and how often a domain crawl should be conducted is usually only defined in the collection strategy of the heritage institution conducting the domain crawl. The question is therefore how the text of the legal deposit law is translated into concrete collection policies of the heritage institution and how these policies delineate the boundaries of the national web domain, what is and what is not included in the web domain, what criteria are used to determine whether or not certain content is part of the web domain, how often a crawl should be carried out and what technical, legal, personnel or other conditions that the crawl should meet.

4. Used technique and technical possibilities, including outsourcing of the crawls

Based on the collection policy, the domain crawl can be performed in different ways. To start with, one can choose to outsource the crawl by a commercial or non-commercial party or have the crawl performed in-house by the ICT department of the heritage institute. When outsourcing, a choice can be made to provide the seed list to the implementing party and leave the rest to the external partner or still have a say in how crawling is done. Other choices that can be made are the software used, the frequency (annual, ongoing or one-time), the settings (how many URLs, hops, crawl time, amount of data), the method of storage, etc. Ultimately, a choice will have to be made about the size, depth, scope, duration and what will be and will not be stored (e.g. no videos because of size).

5. Practices (organization, staff, budget, expertise)

The size and the content of a national domain crawl will be limited by the resources of the organization, like dedicated staff, allocated budget for doing a domain crawl and experience and expertise of the staff in conducting a domain crawl.

6. Cooperation with academic research institutions or academics and other heritage organisations during preparation or after the crawl

The purpose of a national domain crawl is not only to preserve for eternity cultural-historically important digital born heritage, but also to make resources available for history writing and research. Often, a national domain crawl at a national heritage institution is prepared on the basis of the insights from scholarly research, and the result of that crawl is often evaluated again by scholarly research. A national domain crawl is therefore often the result of collaboration between the heritage community and science.

7. Use of the crawl data by academic research institutions or academics

Who has the right to do research with the national domain crawl data and what can or may be done with the archived version of the national domain crawl? What tools can or may be used and what may be done with the data?

Based on points 1 to 7, it can be expected that the result of a domain crawl depends on a lot of factors. It is also to be expected that in each country the laws, policies and their implementation differ in practice. In the section that follows, using the features from points 1 to 7, I will compare the Royal Danish Library (KB) with that of the National Library of France (BnF).

DET KGL. BIBLIOTEK (ROYAL DANISH LIBRARY, KB)

1. Existing tradition of preserving national heritage and collecting born digital content

In both France and Denmark, heritage preservation by the state has a long tradition. Denmark is Europe's oldest monarchy and that form of state is closely linked to the preservation of printed publications. In 1482 the Copenhagen University Library was founded, only three years since King Christian I opened the University of Copenhagen. In 1648 King Frederick III established the royal library following the Legal Deposit Act in the same year, which continues to operate till today. In 1902 the State Library of Denmark opens in Aarhus. When Aarhus University is established in 1928, the State Library in Aarhus also becomes a university library. The integration of libraries is moving further when in 1939 the State and University Library of Denmark is made a national superstructure of education to the public libraries, which helps distribute knowledge to the Danish public libraries and further onwards into society. The Royal Library and the State and University Library of Denmark are merged on January 1st 2017 in the Royal Danish Library (KB), but are still based at two locations in Copenhagen and Aarhus. This library was called 'national library' for a week, but then it gained its old name 'Det Kgl. Bibliotek' back.[1]

The Danish web domain also has an early history and is among the oldest in Europe. The country code top-level domain .dk was created on 14 July 1987. The first Danish URLs dkuug.dk, diku.dk, bk.dk, ibt.dk, ifad.dk, lego.dk, mainz.dk and nordita.dk were registered in that year.

Like most web domains, parts of the Danish web were preserved from the moment the Internet Archive started regular crawling of the worldwide web in May 1996. Because of the technical possibilities at the time, only fragments of Danish sites from the early period have been preserved. Web archiving in Denmark started around 2001 with the project Netarchivet.[2] The name of this project was chosen based on the availability of the domain name, according to one of the founders. It is important to note that the project members came from libraries and research

1. https://tidsskrift.dk/magasin/article/download/66995/96507.

2. Birte Christensen-Dalsgaard et al., *Experiences and Conclusions from a Pilot Study: Web Archiving of the District and County Elections 2001. Final Report for The Pilot Project "Netarchivet.dk"*. (English version: February 2003). https://web.archive.org/web/20121114131014/http://netarkivet.dk/wp-content/uploads/webark-final-rapport-2003.pdf

institutes: Centre for Internet Studies (Niels Ole Finnemann and Niels Brügger), the KB (Birgit Henriksen and Søren Vejrup Carlsen) and the state library (Birte Christensen-Dalsgaard, Eva Fønss-Jørgensen and Harald von Hielmcrone). The reason for the start of web archiving during this period was the event of the Danish municipal elections in November 2001. Cooperation in web archiving between libraries and science and close contact with stakeholders such as politics and the publishing world are the reason for the success of digital born heritage preservation in Denmark.

2. Existing legal deposit law regarding born digital heritage (or the lack of such a law)

The development of legal deposit law in Denmark is also due to a long history of preserving publications and effective collaboration between libraries, universities, politicians and publishers. The Legal Deposit Act of Denmark has been introduced already in 1648. By law, all Danish printers must submit copies of their writings and books they print to The Royal Danish Library, which is established in the same year as the legal deposit law. This law, although in an updated form, still applies today. The first updated happened in 1902. In 1998, a new amendment took place, incorporating the collecting of CDs, videos and static Internet sites into the law.

When web archiving began in Denmark in 2001, it soon became apparent that legal obstacles were hindering the effective preservation of websites. Three laws were relevant in evaluating the legal framework for web archiving: the law on legal deposit, the law on intellectual property rights (copyright) and the law on personal data protection. The legal problem appeared to be that only the static portions of websites fell under the definition of a 'work' in the law on legal deposit and that under the existing law, entire websites containing dynamic content and video material could not be archived.[3] In this way, only a static snapshot of the Danish web could be made.

From 2001 to 2005, lobbying then took place to amend the law on legal deposit that was to enable web archiving. Niels Brügger of the Centre for Internet Studies got hold of a draft of the French law on legal deposit and translated this text from French into Danish.[4] This text eventually led to the amendment of the Danish law in 2005 and an extension of the definition of what can and may be collected.[5]

The essence of the 2005 legal deposit law is that all public available born digital material on the worldwide web published by Danes, in Denmark or aimed at a Danish audience in past, present or future may be retained by the KB.[6] In this, the form of the 'publication' is no longer

3. Birte Christensen-Dalsgaard et al., *Experiences and Conclusions from a Pilot Study: Web Archiving of the District and County Elections 2001. Final Report for The Pilot Project "Netarchivet.dk"*. (English version: February 2003). https://web.archive.org/web/20121114131014/http://netarkivet.dk/wp-content/uploads/webark-final-rapport-2003.pdf 17-18.

4. This was told to me by Niels Brügger in November 2022.

5. The Danish part of the Internet is defined as cultural heritage in the Legal Deposit Act (Act no. 1439 of 22.12.2004), effective from June 1st, 2005. https://www.retsinformation.dk/eli/lta/2004/1439#P8

6. Lov nr. 1439 om pligtaflevering af offentliggjort materiale af 22/12/2004. https://www.retsinformation.dk/eli/lta/2004/1439

important: both websites and social media can and may be preserved, but also moving images or other born digital material from the web. In addition, the Library has the right to archive web content behind a pay wall. Web content that is not on the web for the public but only for specific users will not be preserved and the owner does not have to give access to it. It is also important to stress that this right to archive is also valid retrospectively. Thus, the library can still add to the collection even third-party preserved material from the web that is no longer online.

3. Policy of the heritage institute regarding collecting born digital material.

The legal deposit law defining what belongs to the Danish web domain is deliberately broad and thus made as future-proof as necessary. This prevents the need to amend the law again in the near future because certain material cannot or should not be preserved. The framework of the law and thus the interpretation is determined by the Danish Ministry of Culture. In practice, the management of the national library determines the details of implementation in its policy, but ultimately it is the members of the web archiving team who make the final selection.[7]

Team management ultimately can adjust this selection, if necessary, but most of the time the web archiving team is self-governing. Ultimately, there is no hard steering from management: the prioritizing is done by the team itself and the final selection is also their own responsibility. The members of the team stressed that the selection process is a dialogue: it is a subject of a continuous discussion within the team. “Probably a Danish person has seen it online” can thus be a valid argument to harvest a site.

The web archiving policy making from the library was done in a bottom-up approach with multiple workshops with everybody involved in web archiving at the library participating and a facilitator that was not part of the team, that was responsible for facilitating the overall digital accession strategy — not only web archiving. The question was how can they archive as much as what is published on the Danish web, what Danish people visit on the web and what context has to be preserved to understand all of this, but as resources are limited, the library cannot archive everything it wants.

In principle, therefore, the library could archive almost the entire web on the basis of its policy, since almost all publications on the web are ultimately aimed (also) at a Danish audience. In practice, the selection is reduced to .dk domain websites, sites published on Danish territory and sites in Danish or created by Danes. There were projects like ‘WebDanica’ to find relevant content on non-dk.domains from links in broad harvests as well as manual nominations of .com-, .net, .org-sites, etc.[8] Social media also falls within the criteria of collection building, as long as the content is public available. The team focuses on social media accounts with lots of followers, comments, subscribers or lots of interaction.

It is also important to determine what will not to be archived or what is excluded deliberately from crawls. Ultimately, nothing from crawls is discarded, at most excluded from the index. As one of the members of the team said: “It’s not something we want to collect, we need to collect it.”

7. https://www.kb.dk/en/policies-and-strategies/det-kgl-biblioteks-strategy-accession-digital-cultural-heritage

8. https://github.com/netarchivesuite/webdanica

4. Used technique and technical possibilities, including outsourcing of the crawls

The Danish web archivists have done broad crawls themselves since 2001, including during the test period in 2001. However, archived websites have been purchased in bulk from the Internet Archive over the years in order to fill in the missing parts of the early and the later web (1996-1999 and 2000-2005). These early websites pre-date the legal deposit law but may nevertheless be included in the collection legally according to the adapted law of 2005. These early web collections will be ingested in the web archive so that researchers can study these as well.

The domain crawl starts with a list of URLs from the country code Top-Level Domain (ccTLD). The Danish ccTLD consists of 2.7 million unique domain names (2022). These were delivered in the past as a list, but are transferred to the library through an API now. The Danish domain outside of .dk is mapped by out links from the websites crawled in the broad domain crawl, language analysis of archived sites, lists of Danish companies abroad (23,000 sites), a list of foreign newspapers online with tag “Danish” and other techniques like IP-validation.

There has been made a link graph of the harvested material through time between 1998 and 2003. Still, it is difficult to compare different harvests from different years, as the crawling technique have varied from time to time. The robots.txt was respected during crawling in 2011, which lead to a drop in archived data of 40%. There has been a compression project in 2017 and therefore less material seemed to be harvested then in that year. The first true national domain crawl took place in 2005 after the legal deposit law was passed.

The national domain crawl which takes a snapshot of the Danish web domain takes place around four times a year and starts with a seed list of 2.7 million sites. Sometimes the running time per site can be as much as 2.5 months if it is not manually stopped when archiving irrelevant content or if the new domain crawl starts. After the broad crawl is ended, it will be started again. We can thus state that the Danish domain crawl is actually an ongoing process.

Next to the domain crawl, the team builds up a selective web collection from the following types of websites: almost all Danish news media online (crawling takes place in the range from 12 times daily snapshots to weekly crawls), websites and social media accounts of political parties, organisations and associations, ministries and agencies, selected profiles from social media, YouTube videos (for example weekly) and TikTok videos. It is important to note that most of these social media and YouTube crawls are done ad-hoc and manually, as there are many technical issues.

Also event harvests take place two or three events annually (for example parliamentary elections or during the Corona pandemic). Special collections based on certain themes are also build based on research requests, but the library does not built special thematic web collections, as most of the interesting materials is crawled anyway through other means and can be thematically extracted. It is possible to identify and isolate websites about events or special harvests if needed by using harvest ID´s in queries in SolrWayback. There have been made also some efforts to conduct selective crawls. Still: “The intention of the various strategies is that they, when combined, provide the best possible coverage of what is published on the Danish part of the Internet,” as is stated in a survey. It is therefore not necessary to built up special collections, as most of the relevant material will be crawled anyway and can be traced in the collection by SolrWayback queries.

It is important to note that there is a collection of 100-200 Danish websites which sites are so huge that these have to be crawled continuously to collect all data from the sites. We can rather say that the domain crawl is continuously finetuned to better follow trends and pinpoint interesting Danish content by focusing on the context. As the team noted: there is no such a thing as a perfect domain crawl. It is the result of policy, research, experiments, recrawls, quality control and lots of talk between the team members, researchers and experts.

5. Practices (organization, staff, budget, expertise)

The selection and archiving is done on two locations: at KB in Copenhagen and in Aarhus. The work is divided between curators, development, system operators and management.

- Curators (both "soft" and "tech savvy" curators) from the Department of Digital Cultural Heritage (approx. 2-4 FTE — depending on the jobs that needs to be done and if an event harvest is ongoing) and a program manager (1 FTE).
- Developers from IT Development Department (FTE will differ if there´s a specific project ongoing or it´s maintenance and minor development to fix bugs), so anywhere from 2.5 FTE to 0.2-0.5 FTE.
- System operators from IT Sys Department, depending on the work that has to be done.

Summed up:

- Maintenance and operation (keeping the systems running) approximately: 1.4 FTE.
- Curational effort/management etc.; 3-5 FTE.
- Only the Program Manager is working full time on a regular basis. Depending on prioritized projects developers might work full time for extended periods of time.

The skills of the members of the web archiving group can be summed up as: having a curious mindset, interest in history, various niche cultural areas of Denmark and from other countries, current affairs, and Danish society and internet culture in general. Of course the members need to have librarian and information specialist-type skills, search skills, knowledge of scripting and an understanding of the technical background of the web, website and web archiving. People who are good at web archiving love patterns and can see the patterns in the vast amount of messy data.

To be prepared for what is to come in future and preserve as much as possible of what is now online, there is a lot of room for technical experimentation with new media, social media and new techniques. One example is the use of customized game controllers during the quality control of the domain crawl. This makes it easy to adjust the byte size/-limits for 12K domains reaching max byte limit manually by moving up and down.

6. Cooperation with academic research institutions or academics and other heritage organisations during preparation or after the crawl

Apart from the staff, there is a kind of advisory committee of scientists and publishers and other members of the media (the Web/TV Editor group). It is an initiative of the cultural ministry that can help the team with input on online media trends. The members are nominated by KB.

7. Use of the crawl data by academic research institutions or academics

There is a lot of crawl data which can be studied by researchers of the Danish web.[9] Only the SOLR index alone is already 120-140 TB big. The library keeps a large amount of different log files, mainly of a technical nature. For some event based collections there has been made a semi-formalised initiation and (afterwards) evaluation process. The staff uses a Jira-based system for resolving issues. This can also be seen as documentation and certainly is used internally from time to time, but not by researchers. There is information about the terms of use on the website.[10]

BIBLIOTHEQUE NATIONALE DE FRANCE (NATIONAL LIBRARY OF FRANCE, BNF)

1. Existing tradition of preserving national heritage and collecting born digital content

The BnF has one of the oldest traditions of a legal depository with the aim of preserving the country's printed works. This tradition has its origins in the reign of King Francis I. and the change in thinking about kingship in France. According to tradition, the sovereign's God-given kingship was legitimized by coronation with a crown jewel. In the 16th century, this changed and the royal dynasty became the legitimation of kingly power. This put the emphasis on the historical roots of the dynasty and the sources that could confirm these ancient origins. With the help of jurists and historians, the king started collecting as many sources as possible that could support the dynasty's legitimacy. The legal repository was also created for this purpose. A copy of every printed work in the country had to be submitted to the king's library, the forerunner of the national library. With this, the world's oldest legal depository was born.

The ccTLD of France, .fr, was registered on 2 September 1986, one year before the Danish ccTLD and four months after .nl, the first ccTLD outside of the US. The French web domain has had a 'national' character from the beginning and is really used to demarcate the ‘borders’ of the French domain on the web. If someone wants to register a .fr domain, he or she must be a resident of the European Union or an EFTA member state (Switzerland, Norway, Iceland or Liechtenstein). It is also remarkable that due to Brexit, UK residents are not able to register new .fr domains anymore since 1 January 2021.

The ‘national’ or ‘state’ character of the digital sphere is also stressed by the digital ccTLD’s of its overseas regions and territories in the Americas and the Atlantic, Pacific and Indian Oceans which are domains under French administration:
.bl: CC TLD for Saint Barthélemy
.gf: CC TLD for French Guiana
.gp: CC TLD for Guadeloupe
.mf: CC TLD for Saint Martin
.mq: CC TLD for Martinique

9. https://www.kb.dk/en/find-materials/collections/netarkivet/research-access

10. https://www.kb.dk/en/find-materials/collections/netarkivet/applicant-declaration-netarkivet/applicant-declaration-netarkivet.pdf

.nc: CC TLD for New Caledonia
.pf: CC TLD for French Polynesia
.pm: CC TLD for Saint Pierre and Miquelon
.re: CC TLD for Réunion
.yt: CC TLD for Mayotte

The difference between the Danish and French legal conceptions of the to be harvested national web domain is the following. The Danes consider the web as a source of information for Danish citizens and this information must be preserved as cultural heritage for the future. The French consider the web as a national digital domain where certain state actors can exercise sovereign rights over. One of these rights is the preservation of the content of this domain as national cultural heritage.

2. Existing legal deposit law regarding born digital heritage

The legal deposit law in France started as a royal prerogative of the king, but now is it considered as an activity in the public interest. Since 2006, the legal deposit law concerning material of the Internet concerns all public content of French websites and online publications, including radio and television sites. Two separate heritage institutions are responsible for the preservation of born digital heritage from the web according to this law: the national library BnF, the 'Institut national de l'audiovisuel' (INA) and the Centre national du cinéma et de l'image animée (CNC). French websites are harvested by the BnF, French radio and television sites are preserved by INA and French movies by CNC.

The law does not require website owners to deposit their websites with the French heritage institutes designated by law. Instead, the institutes have the task and right to harvest what they deem of interest from the French web. The owners of the websites and the creators of web content have the duty to cooperate with the heritage institutions, to supply a copy of the content of the site or to grant access. They are not allowed to restrict the access of crawlers from the institutions. Also the registers of French websites have the duty to supply the identification details of website owners or producers to the heritage institutions.

The legal deposit law defining what belongs to the French web domain is not as broad as the Danish one, but the heritage institutions have more rights to harvest material behind paywalls. The law permits the harvest of all digitally written signs, signals, writings, images, sounds or messages of any kind communicated to the public by electronic means. This 'communication' can be published on.fr domain websites, but also websites with other extensions (.com, .org, etc.) registered by persons having French nationality, living in France, made by companies on French territories or websites produced on French territory are allowed to be harvested. The BnF does research on the extend of the French web, but this mapping of the French web domain is not exhaustive. The difference between Danish law and French law is that French law focuses on the production on the French web of the digital-born communication and Danish law focuses on the consumption of digital born communication by Danes.

Another important difference is the division between harvesting textual based websites and audiovisual content from sites and social media, which is divided between BnF and INA in France.

We will not describe in detail the harvest policies of INA. We will only note that this organization harvests also material from Twitter, YouTube and Instagram, among other audiovisual social media outings used by and produced by the French public. INA selects more than 14,000 French websites and 12,600 social network accounts for its content.[11]

3. Policy of the heritage institute regarding collecting born digital material

The BnF is entitled by law to decide which harvesting policy and practice it will choose to fulfil it legal task. It is therefore allowed to conduct a domain crawl and to harvest on a selective base. It is only obliged to notify the owners or producers of the sites which harvest strategy it uses to collect born digital material from a site.

The BnF has two collection policies to fulfil its task to harvest born digital material according to the legal deposit law. The first one is the yearly broad crawl of the French national web domain which aims at making a regular representative snapshot of the French web, like the Danish domain crawl. The scope of the crawl is defined by the law as described above.

The second one is a selective approach. It is directed at making 'targeted web collections' about topics depending on the actuality in French society. These collections consists of a selection hundreds till tens of thousands of sites which is made by librarians. The crawls of these collections can be more frequent and thorough and quality control can be done. The BnF considers these collections as representative for the French digital cultural heritage on the web. One of the special web collections which were created by BnF was the corona / COVID-19 collection. The creation was done by monitoring the web for certain keywords in the URL. In total 3,000 sited about this theme were identified, of which one third was harvested. The sites not selected were redirects, under construction or not active. The intention is that such collections will eventually be full-text searchable.

4. Used technique and technical possibilities, including outsourcing of the crawls

Both the Danish KB and the BnF use the Netarchive suite. As both institutes have regular status meetings, the used techniques, technical developments and results can be followed through the minutes of the meetings.[12]

The domain crawl of the BnF took 32 days in 2020, compared to 59 days a year earlier. Before the harvest started, the BnF ran tests on domain names (DNS-checks). It appeared that the French web domain grew with more than 1,15 million domains and shrunk with more than 725 thousand sites. These tests show the volatility of the French web and the need for preservation. The total amount of

11. https://www.ina.fr/institut-national-audiovisuel/collection-preservation-and-documentation-of-audiovisual-heritage

12. See the header 'statusmeetings': https://sbforge.org/display/NAS/

Social media is also selected and harvested by BnF. Following the TikTok crawl launched in 2022 on the theme of the French elections, BnF launched its first current TikTok harvest in March 2023. Till so far, 198 French TikTok accounts or French tags had been selected.[13]

5. Practices (organization, staff, budget, expertise)

The structure of the organization of the BnF and the web archiving team is different from that in Denmark. First of all, the broad crawl of the French domain is done by the Legal Deposit department at the BnF in Paris. The team consists of a head, two members and technical staff, in total seven people. The selection is done not only at BnF in Paris, but also by curators and collection specialists in various other legal deposit libraries in the regions of France. So in practice this team is much bigger and embedded in local institutions.

The BnF has started this cooperation with local experts in since 2004 when websites related to the general elections in France were selected. Since then, local legal deposit libraries have been involved in documenting the French web. Cooperation with regional partners also covers the selection and preservation of local digital born heritage. In 2013, the 'Alsatiques en ligne' of the National and University Library of Strasbourg represented the first collection of its kind. Since then, the Emile-Zola Montpellier Méditerranée Métropole Media Library and the Stanislas Library in Nancy have also selected sites of regional importance.

6. Cooperation with academic research institutions or academics and other heritage organisations during preparation

The BnF does cooperate with academic research institutions, like in the Respadon project, the RESAW (Research Infrastructure for the Study of Archived Web Materials) community and the WARCnet community.[14] It is not clear whether BnF works together with academics during the preparation of broad crawls.

7. Use of the crawl data by academic research institutions or academics

All of these born digital collections harvested according to the legal deposit legislation can be consulted in the Internet Archives, a platform accessible on the various BnF sites and in several legal deposit printing libraries.

The digital collections of archived materials which are collected by the National library of France within the framework of the Legal Deposit of Digital Materials can be used for research in two cases.

The first one, against which right holders cannot oppose, can only be claimed by research organizations and heritage cultural institutions – that is the National Library of France and some of its partners belonging to one of both categories mentioned afore.

13. https://sbforge.org/display/NAS/2023-03-07+Statusmeeting

14. See: Inès Carme, Le projet ResPaDon, what for? — looking back on a unique collaboration around French web collections; https://cc.au.dk/en/resaw and https://cc.au.dk/en/resaw.

The second one, open to any individual, can be refused by right holders (opt out option). If this is not the case, any individual will be entitled to undertake a TDM search on targeted digital collections under the condition that this individual is an accredited researcher.

If a researcher is indeed accredited, they can access the contents of the web archive at the library's location from their own laptop. It is also possible that the researcher visits another accredited library in France.

Access to collections entered through Legal Deposit for the purpose of data mining has to be lawful.

Distant access to the born digital collections by researchers, research organizations or any other body for the purpose of text or data mining is not possible yet.

Remote access to the web archive is given in 26 regional and national libraries:

1. Archives de la Martinique - Fort-de-France
2. Archives Départementales de la Guadeloupe - Gourbeyre
3. Bibliothèque Alexis de Tocqueville
4. Bibliothèque d'Étude et conservation - Besançon
5. Bibliothèque d'Étude et du Patrimoine - Toulouse
6. Bibliothèque départementale de La Réunion
7. Bibliothèque des Champs libres - Rennes
8. Bibliothèque du Patrimoine - Clermont-Ferrand
9. Bibliothèque Francophone et Multimédias - Limoges
10. Bibliothèque Mériadeck - Bordeaux
11. Bibliothèque municipale - Orléans
12. Bibliothèque municipale Part-Dieu - Lyon
13. Bibliothèque municipale Pompidou - Châlons-en-Champagne
14. Bibliothèque Nationale Universitaire - Strasbourg
15. Bibliothèque patrimoniale et d'étude - Dijon
16. Bibliothèque patrimoniale Villon - Rouen
17. Bibliothèque Stanislas - Nancy
18. Bibliothèque Toussaint - Angers
19. Médiathèque centrale Emile Zola - Montpellier
20. Médiathèque François-Mitterand - Poitiers
21. Médiathèque Jean-Levy - Lille

Remote access on BNF locations:

1. BnF - Arsenal
2. BnF - François-Mitterand
3. BnF - Maison Jean Vilar - Avignon
4. BnF - Opéra
5. BnF - Site Richelieu-Louvois

CONCLUSION

As we have seen in this report, each national library or heritage institution has its own national domain crawl data collection with very specific characteristics that differ from other national web archives. Both the KB and the BnF are among the most advanced and innovative web archiving heritage institutions in Europe. Nevertheless, there are major differences in the way the national domain crawl is defined in law, developed into policy and eventually implemented in the daily work processes. There is no telling which approach is better than the other. Each library aims to web archive as much of the national web domain as possible, but ultimately will have to wrestle with legal, policy, financial, technical and staffing frameworks. Ultimately, based on their own specific research question, the researcher of the future will make the judgment as to which collection is more suitable as a research source.

REFERENCES

Maemura, E. (2022). *Towards an Infrastructural Description of Archived Web Data*. WARCnet papers. Aarhus: WARCnet.

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