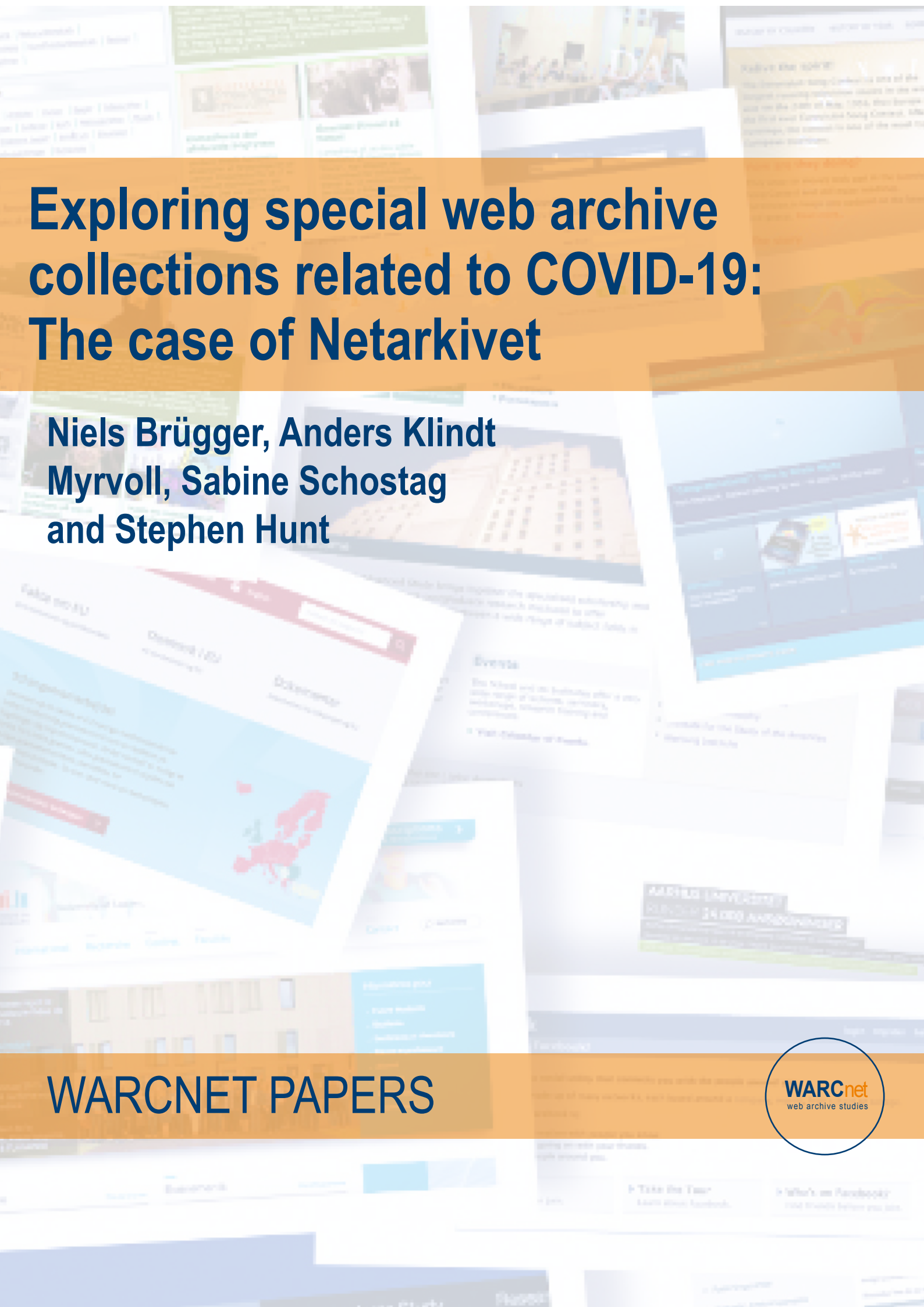
Exploring special web archive collections related to COVID-19: The case of Netarkivet

Niels Brügger, Anders Klindt Myrvoll, Sabine Schostag and Stephen Hunt



WARCNET PAPERS

WARCnet
web archive studies

Exploring special web archive collections related to COVID-19: The case Netarkivet

An interview with Anders Klindt Myrvoll, Sabine Schostag, and Stephen Hunt (Netarkivet) conducted by Niels Brügger (Aarhus University)

nb@cc.au.dk

WARCnet Papers
Aarhus, Denmark 2020

WARCnet Papers ISSN 2597-0615.

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

Published by the research network WARCnet, Aarhus, 2020.
Editors of WARCnet Papers: Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster, Michael Kurzmeier.

Cover design: Julie Brøndum
ISBN: 978-87-972198-5-0

WARCnet
Department of Media and Journlism 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* (2020)

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

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

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

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

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

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

Exploring special web archive collections related to COVID-19: The case of Netarkivet

An interview with Anders Klindt Myrvoll, Sabine Schostag, and Stephen Hunt (Netarkivet) conducted by Niels Brügger (Aarhus University)

Abstract: This WARCnet paper is the third in a series of interviews with European 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 archive collections. This publication explores how the national Danish web archive Netarkivet has collected the web activity related to COVID-19.

Keywords: web archives, social media, COVID-19, special collections, Denmark, Netarkivet

INTRODUCTION

This interview is part of the series of WARCnet Papers about how web archives have collected material related to the COVID-19 pandemic. The interview was conducted on 14 September 2020 with Anders Klindt Myrvoll (Programme Manager), Sabine Schostag (Web Curator), and Stephen Hunt (Digital Collection Manager) from the national Danish web archive Netarkivet.

Since 2005 the collection and preservation of the Danish part of the internet is included in the Danish Legal Deposit Law. The task is undertaken by Netarkivet at the Royal Danish Library (cf. netarchive.dk). Netarkivet is not accessible to the general public, but the archive is open online to researchers who have requested and been granted special permission to use the collection for specific research purposes.

Netarkivet uses four archiving strategies to collect and preserve as much of the Danish internet as possible: a) snapshot harvesting, where the entire top-level domain .dk and also Danish material outside this domain are harvested, b) selective harvesting, where app. 200-300 websites that are rapidly updated are collected, c) the event harvest, where focus is on collecting web activity related to an event (political, sport, terrorist attacks, etc.), d)

special harvestings, where e.g. a researcher would like to have set up a special web archiving.

As of today the holdings of Netarkivet is 691 TB.

THE REASONS OF THE SPECIAL COLLECTION

Did you conduct a special COVID-19 collect? And in case you did: why?

Anders Klindt Myrvoll: Yes. It is part of Netarkivet's tasks to document major events like a pandemic as a part of our obligations to fulfill the Danish Legal Deposit Law. An "event" is defined by the fact that it has so much public attention that it triggers new websites and is also largely treated on existing websites.

So, who decides when to start an event harvest? What would be the main criteria?

Anders Klindt Myrvoll: The group of curators working at Netarkivet decides if an event should trigger an event-harvest (big or small). We have short daily meetings in this forum where events can be discussed. The main criteria is maybe impact as a broad term, but can also be more narrow events like Google removing all Danish music from YouTube due to a disagreement on how to compensate Danish artists for their work (cf. http://cphpost.dk/?p=117112).

Sabine Schostag: Yes, events that generate extra activities on the internet will be harvested as event collections. Nowadays these activities primarily happen on social media platforms. Netarkivet's strategy for selective crawls covers all Danish news media, no need to focus on them in an event collection. The 'COVID-19 event Crawl' was part of an initiative run by the Royal Danish Library and other Danish cultural institutions on all aspects of COVID-19's influence on Danish society, especially during the lock-down period.

Stephen Hunt: Yes, events that are not already covered by our daily selective crawls.

What about the distinction between event harvest and selective harvests, are the borders clear?

Sabine Schostag: Selective crawls are ongoing in specific frequencies and depths. Frequency and depth depend on the update frequencies of the web domains to be crawled selectively. Our selective crawls cover all Danish news media, domains reflecting ordinary life in the Danish society and domains using the internet as an experimental platform. There is more and more focus on social media to be crawled selectively. Event crawls will be run on and about events generating temporary augmented activity on the Danish part of the Internet. Anyway, when searching in our archive, you can use the freetext search — no need to think about whether the content would come from selective crawls or event crawls.

THE SCOPE OF THE COVID-19 COLLECTION

What did you collect? Websites only? Social media (if so, which ones, specific hashtags, profiles, languages)?

Anders Klindt Myrvoll: We collected a wide array of different content. Below is a list of content specifically concerning COVID-19 and in addition to our normal harvests:

Websites (www): 2,300+. Mostly Danish sites but we also found sites with Danish relevance in Norwegian, Swedish, Greenlandic, Faroese, Icelandic, Finnish, German, Italian, Austrian, American, British, Chinese and New Zealand sites.

Podcasts: 17+

Facebook profiles: 250+ (and 1,700+ as a one time harvest via Archive-It from mid-September). Curated manually and consisting of profiles for politicians, political parties, journalists, opinion makers, COVID-19-related groups, satirical content, memes, and more.

Twitter: at least 2,200+ accounts (most are a part of our usual Twitter-harvest) and quite a few COVID-19 specific or related hashtags: #agfrfc, #autoritetstro, #COVID19dk, #Coronavirusdk, #hamstring, #lagkagehuset, #remdesivir, #skattely, #SkylandBeachCamp, #smittestop, #sommeridanmark.

Instagram: 400+ accounts.

Reddit: 1,800+ posts, and all Danish subreddits as part of our normal selective harvests.

TikTok: First we had difficulties even finding Danish profiles. Then we found a way to identify Danish profiles with COVID-19-related content. We didn't manage to harvest the TikTok-content, except a few videos, though.

YouTube: 300+ videos/channels.

Twitch: 5,000+ Danish channels that we haven't harvested yet. We saw a big increase in the amount of profiles after the lock down. We would have liked to get COVID-19-relevant content like Danish DJ Martin Jensen's DJ-sets on Twitch during the pandemic, www.twitch.tv/martinjensentv:

- Playing a DJ-set in an airplane hangar in Billund Airport (https://vafo.dk/artikel/dj-holder-online-koncert-martin-jensen-er-flyvende),
- 5-hours DJ-set in Parken (Denmark's national soccer stadium) 40,000+ capacity — playing for empty seats (https://www.fck.dk/nyhed/dj-martin-jensen-live-koncert-fra-telia-parken-denne-fredag),
- DJ-set at the frigate Niels Juel (https://olfi.dk/2020/05/08/verdensberoemt-dj-stoetter-veteraner-med-koncert-paa-fregatten-niels-juel/ - https://www.redbull.com/dk-da/martin-jensen-skovt%C3%A5rnet).

We will try to see if this COVID-19 relevant content can be obtained.

Could you elaborate a bit on what was challenging with this material?

Anders Klindt Myrvoll: The content was only available for a short period of time — maybe a week after it was published live for the first time (I didn't watch the event so it might have been shown just once). It's not available anymore. We didn't try to harvest Twitch before, so we hadn't figured out how to harvest this new platform before the need was there. We had to prioritize our resources and in the end we needed to focus on other content.

Stephen Hunt: One thing is to find the websites where content about COVID-19/Denmark is presented, but another thing is if we can actually harvest these websites. An example is e.g. Twitch where we can see an increase in activity because kids are sent home and have "more" time, but it hasn't technically been possible for us to harvest this website. Soundcloud is also a problem. The problem is that websites are focusing more on dynamic content, quick replies, large scale content and that requires more modern technology and our harvester is only a "static" harvester that collects/harvest URLs from source code. Dynamic content is not always found in the source code but generated when a user accesses a website or presses a button or scrolls down the page or plays the video/music content.

Sabine Schostag: Everything but news media content, as this is covered by the regular selective crawls. We focused on websites from public bodies such as municipalities, ministries as they changed their information on COVID-19 to the public frequently. Often they just update/updated one page, so it was important to capture these pages every day.

How did you identify in particular the social media to be included? — based on which criteria were they selected?

Sabine Schostag: Searches for keywords, hashtags, relevant persons' accounts.

How do you archive nationally something which is fundamentally global?

Sabine Schostag: We focused on COVID-19 in Denmark, this is the Legal Deposit angle (the so called Danica — content produced by Danes, in Danish or addressed to a Danish audience). Of course, sometimes, especially on social media it can be difficult to demarcate Danica. We also focused on news media reactions from other European countries and worldwide on what happened in Denmark during the pandemic.

Anders Klindt Myrvoll: That´s a really good question. We are obliged to focus on the descriptions in the Legal Deposit Law regarding Danish material. See http://www5.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html):

PART 3

Material published in electronic communication networks

§ 8. *Danish material published in electronic communication networks is subject to legal deposit. The legal deposit obligation is fulfilled by the legal deposit institution having access to request or produce copies of the material.*

(2) Material published in electronic communication networks are considered to be Danish when

1. *it is published from Internet domains etc. which are specifically assigned to Denmark, or*
2. *it is published from other Internet domains etc. and is directed at a public in Denmark.*

Excerpt from the Danish Legal Deposit Law.

For this event harvest we defined that we wanted:

- Danish web pages about Corona virus in Denmark (but not articles and theme sections from Danish news media — they are covered by ongoing selective collections, see also below),
- Foreign web pages about Corona virus in and concerning Denmark,
- Posts/groups on social media with Danish content related to Corona virus.

We have a limited amount of resources available and need to prioritize. So we focused on Danish relevance, and then hope other national libraries cover their "territory".

THE FRAME OF THIS SPECIAL COLLECTION

When did you start? When do you (plan to) stop? Capture frequency?

Sabine Schostag: In a sense, the story of COVID-19 and Netarkivet starts at the end of January 2020 — about 6 weeks before COVID-19 came to Denmark. A cartoon by Niels Bo Bojesens in the Danish newspaper *JyllandsPosten* (26 January) 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.

After that, the curators again focused on the daily routines in web archiving, as we believed that COVID-19 was a closed chapter in Netarkivet's history. But this was not the case. Suddenly, the virus arrived in Europe and on March 12, the Danish Government decided to lock down the country — all employees were sent to their home offices and borders were closed. The IIPC (International Internet Preservation Corporation) started an international event collection on COVID-19 in the middle of February 2020. We contributed with Danish content. The IIPC event collections give open access to the content, so in this way some Danish content can be accessed by the public.

Anders Klindt Myrvoll: We have ongoing discussions on when to stop. When do the event transform to a general condition, and what does that imply in terms of termination, frequency etc. The amount of dedicated new domains is getting smaller every day so we are spending less time to ensure all relevant domains are included and will most likely lower the frequency of the sites that are part of this event harvest.

So do you plan to stop the event harvesting because there are fewer new relevant domains, although the pandemic continues, or do you plan just to let the setup continue, but without monitoring it much?

Anders Klindt Myrvoll: Time will tell, but we are aiming at letting the setup continue but with less monitoring. If we don't, and terminate the event crawl, we might get less content, but as the event lasts longer many sites will be part of our broad crawls. We have many different tasks at the Royal Danish Library, even in the team working with web archiving, and there's an on-going prioritization of resources and tasks that needs to be solved. We definitely used way more hours in the start of the event than now.

Sabine Schostag: According to frequency: the main COVID-19 crawl runs daily. As this is a supplement to the regular selective news media crawls, we have to mention, that news media crawls run between 12 times a day (front pages), once a day and once a week — depending on the frequency of updates on the different sites. We are just discussing when to stop the COVID-19 collection, there are different opinions, one possibility would be to lower the crawl frequency.

How was quality control done on the collection (if applicable)?

Sabine Schostag: QA was conducted as samples.

Could you elaborate on that? Who was questioned?

Sabine Schostag: the QA was conducted by the curators, checking both crawl.logs and visually sample check in the archived content. A home made script listening problems from the crawl.logs helps identify problems crawling certain pages. Anyway, we only have ressources for sample QA.

What has been the amount of data? And the nature of the data (videos, tweets, etc)? And do you have a quantitative overview? Some figures were mentioned above, but how did you monitor the nature and amount of data while you were archiving? And did you encounter any issues, challenges, or limits related to the collecting activity?

Stephen Hunt: Dynamic websites (javascript based) are not always possible to harvest and playback with the 'look and feel' from the live web. Instead we sometimes manage to harvest content like videos and music files but these files are not found by accessing the web page URL and therefore a researcher will need information from the curators who made the harvest or search for the content themselves to actually find it. E.g. YouTube is not possible to playback correctly but to make sure we get the videos we use youtube-dl that can generate the video URLs that we then can harvest. Youtube-dl is a tool to harvest videos from websites and mainly developed for YouTube, but can also be used for other websites where there is video content. This is really not the right way, but our only way to ensure we get video content from the websites.

Could you elaborate on how you document this splitting up of things so that researchers can find the videos later?

Stephen Hunt: We always make sure to document the title, watch URL, video URL, harvested date and publisher. Plus we have also recently started collecting the metadata (json file) from youtube-dl that include all the metadata of the video. This is for now only documented internally and we would like to make these documents accessible for external users, but they are unfortunately not.

Sabine Schostag: We are not able to collect content from Facebook properly with our own tools — we have an account with Archive-IT for this part of the collection.

Anders Klindt Myrvoll: As a test, a few of us are manually using https://conifer.rhizome.org/ (Webrecorder-technology) to capture Facebook posts and all comments on prioritized profiles. We don't have a complete workflow to get the downloaded WARC-content from Conifer into Netarkivet but are working on it as we speak, including making sure it will fulfill all relevant preservation standards. Our strategy for collecting content favours automated processes, but also the need to challenge the boundaries of what relevant content we can actually get and how.

From a researcher point of view it will be a great advantage just to have this material in the collection, and then later potentially figure out how to fully ingest it in the web archive. Like in the good old print world where the vast majority of what the Royal Danish Library collected based on the Legal Deposit Law was not catalogued, but just stored and then when many years later a researcher wanted to study it, then it could eventually be catalogued.

Anders Klindt Myrvoll: This is a quite complex and interesting discussion. We download the data that we generate on platforms like these to our servers. But they are not part of the main web archive, are not indexed and can not be searched in and playbacked in our

access solutions, before they are ingested. There's a fine line between using too little and too much time on emerging web archiving technologies. We have to make sure that we get the most value possible from our decisions and work.

How easy was it to manage collecting activity from home, in case you did that?

Sabine Schostag: While Denmark came to a standstill, so to speak, Netarkivet's curators worked at full throttle on the COVID-19 event collection. Zoom became the most important work tool for the following 2½ months. In daily Zoom meetings, we coordinated who worked on which facet of this collection. To put it briefly, we curators had the corona virus on our minds. For our kind of work we need a laptop and a good connection to the internet. I had/have both at my "home office", so it worked fine for the technical aspect. Of course, there was the psychological aspect (not seeing the family, children and grandchildren was rather hard for me).

Stephen Hunt: This was actually quite easy. The way we work really didn't change from when we were working at the office and home.

Anders Klindt Myrvoll: In many ways working online and specifically introducing a 30 minute daily meeting with all Netarkivet employees (curators, IT-specialist etc.) brought us even closer as a team — and gave us motivation to really focus on getting the best results as possible for this groundbreaking event.

ACCESSIBILITY AND SEARCHABILITY

How is it (or will it be) possible to access and search these data?

Sabine Schostag: There are the same rules as for all Netarkivet content: you have to be a Danish researcher or you have to be associated with a Danish research institution for having the possibility to send an application for access to the archived content. The decision on whether you are granted access or not is based on individual assessment

Anders Klindt Myrvoll: When you are approved you get access via a Citrix-solution with two factor authentication. Playback is in OpenWayback and free text search in Blacklight (with playback in OpenWayback). We are working hard to get SolrWayback in production and this will give better and faster searchability and playback, and will include quite a few interesting tools, cf. https://github.com/netarchivesuite/solrwayback. All of Netarkivet's holdings are continuously, with a few months intervals, Solr-indexed to make the solution as swift as possible. Since the end of 2018 we've also had the possibility for researchers to get data dumps from Netarkivet for further research.

Are researchers already asking you about these collections, waiting to analyse them?

Sabine Schostag: there are at least two research projects on COVID-19 related themes who have got access to the archived material.

How did you communicate about this special collection?

Sabine Schostag: Press release (in the frame of the general COVID-19 cooperation with other cultural institutions), news articles on our websites kb.dk and netarchive.dk. Posts on The Royal Danish Library's Facebook account. Examples of the communication can be seen here:

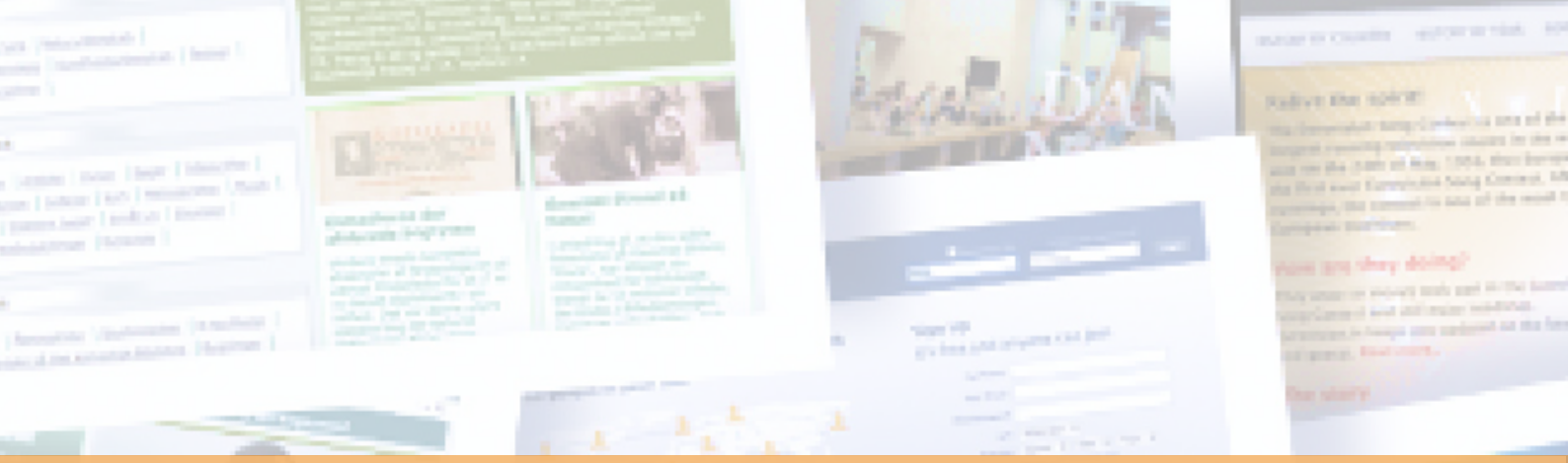
- https://kum.dk/nyheder-og-presse/pressemeddelelser/nyheder/corona-lukningen-af-danmark-skal-dokumenteres-og-bevares/1/1/
- https://www.kb.dk/nyheder/fortael-os-om-dit-liv-under-coronakrisen
- https://www.facebook.com/DetKglBibliotek/posts/3950195331687223

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

Stephen Hunt: We also use Archive-It as a supplement to our harvest, because it hasn't been possible to harvest Facebook since the start of 2016 because of CAPTCHA (Completely Automated Public Turing-test to tell Computers and Humans Apart). Archive-It has been more successful and shown better results, but because Facebook changes it's technology nearly daily and probably has a focus on preventing us from harvesting them we can't manage to harvest them by ourselves, and Archive-It for what I know has a dedicated team that works on changing configurations frequently so it's actually possible to harvest Facebook. The quality of the harvest is both good and very bad sometimes.

Anders Klindt Myrvoll: Using Archive-It we have, in some cases, harvested all posts on a profile as much as 4 years back in time. We get pictures and videos but unfortunately not comments.

Sabine Schostag: Informal cooperation with colleagues from the Royal Danish Library and the University of Copenhagen helping with the identification of URLs during the lock-down period.



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