Exploring special web archive collections related to COVID-19: The Dutch Web archive (KB)

Karin de Wild, Ismini Kyritsis, Kees Teszelszky and Peter de Bode

WARCNET PAPERS

WARCnet
web archive studies

Exploring special web archive collections related to COVID-19: The Dutch Web archive (KB)

An interview with Kees Teszelszky and Peter de Bode
conducted by Karin de Wild and Ismini Kyritsis

k.de.wild@hum.leidenuniv.nl

WARCnet Papers
Aarhus, Denmark 2021

WARCnet Papers ISSN 2597-0615.

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)

Published by the research network WARCnet, Aarhus, 2021.

Editors of WARCnet Papers: Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster, Michael Kurzmeier.

Cover design: Julie Brøndum
ISBN: 978-87-94108-04-1

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).

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)

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)

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

Exploring special web archive collections related to COVID-19: The Dutch Web archive (KB)

An interview with Kees Teszelszky and Peter de Bode conducted by Karin de Wild and Ismini Kyritsis

Abstract: This WARCnet paper is part of 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 Dutch web archive has collected web activity related to COVID-19.

Keywords: web archives, social media, COVID-19, special collections, The Netherlands, KB, Coronavirus

This interview is part of a series of WARCnet Papers about how web archives have collected material related to the COVID-19 pandemic. The interview is a result of two rounds: the first interview was conducted on 30 October 2020 with Kees Teszelszky (curator digital collections) from the National Library The Netherlands (KB); these answers were complemented by a second interview with Peter de Bode (Collection Specialist for web archiving, KB) that took place on 18 May 2021. As a national library, the KB is responsible for selecting, preserving and making available digital publications issued in the Netherlands, including websites. The main focus of the library is to collect and preserve websites that contain scientific and cultural content, as well as those websites that were at the forefront of technological developments at the time.

The Netherlands does not have legal deposit legislation, and therefore some form of consent is required from the rights holders to be able to collect websites. So instead of automatic harvesting of websites in large quantities, the KB has decided on a more selective approach in line with their collection policies. This also gives them the advantage of being able to pay more attention to technical details and to archive websites to the deepest level.

Each website consists of a large amount of individual files. To capture as many files as possible, the KB uses a Heritrix crawler that “wraps” these files into a “container”. This makes it easier to manage the archived version. Each individual file is described by metadata that gives information about the file format, time and date of the crawl and the

size of the file. The KB web archive has been available to the general public since 2011 in the reading rooms of the KB, and in addition data can be requested and is available for research purposes on the KB Lab site (e.g. “special web collection internet archaeology Euronet-Internet (1994-2017)”, “websites from the Chinese community”). Because of copyright restrictions, wider availability of the Dutch web archive has not yet been realised.

THE REASONS FOR THE SPECIAL COLLECTION

Ismini Kyritsis and Karin de Wild: Did you conduct a special COVID-19 collection? And if you did, why?

Peter de Bode: A special web collection is a window to the web collection as a whole and it gives an opportunity to highlight a specific event or subject from an internet point of view. I closely monitored the news for several weeks from January 2020 onwards, and when the first Dutch COVID-patient was identified I started collecting web content for a special web collection and for the Novel Coronavirus (COVID-19) collection of the International Internet Preservation Consortium (IIPC).[1]

THE SCOPE OF THE COVID-19 COLLECTION

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

Kees Teszelszky: Within the KB web archive, our aim is to capture a snapshot of the digital web culture in the Netherlands. We are the only web archive in the world which focuses exclusively on websites from the Netherlands, made by and for the Dutch and published in the Dutch language. However, we do not confine ourselves to the .nl domain; every website on COVID-19 published by a Dutch citizen can be selected. Although the only official language of the Netherlands is Dutch, our collection also includes websites that are written in other languages. We do not respect robots.txt. as we try to crawl sites as completely as possible. The crawl frequency is once a year, but it is possible to crawl more often if needed.

Peter de Bode: Websites for the special COVID-19 collection were crawled more regularly, sometimes weekly, monthly, bi-monthly or quarterly.

Kees Teszelszky: Within this special COVID-19 collection, we try to capture the influence of the Corona crisis on the Dutch Web culture. Our main focus is on websites that are an expression of a cultural change. So, when a site only reports on COVID-19 or related statistics, it is of less importance to us, but this data may still be preserved elsewhere. However, our main focus is on websites that are representing a cultural transition on the

1. For more information about the collection of the IIPC, see Bingham and Geeraert (2021).

web due to the global corona pandemic. We are mostly interested in websites that report on cultural phenomena in society. Often these phenomena dominate the news and social media only for a limited period of time; after that they disappear – and sometimes also from people's memory. Yet these sites can still be very interesting for future historians. We tried to establish a COVID-19 special collection that will become an important collection for historians interested in understanding the COVID-19 crisis.

Not everything is included within this collection. Sometimes this is caused by practical limitations. For example, online newspapers are not included as we are unfortunately unable to archive websites behind a paywall. We do archive online newspapers, like nu.nl, a Dutch online newspaper that reported extensively on COVID-19. We have crawled this on a daily basis. Also, government websites are not our main focus, as these are mostly preserved by the National Archive in the Netherlands. We do not have the expertise and capabilities yet to archive social media, but there are several scientific projects in the Netherlands that focus on preserving COVID-19 social media resources. We started a project this year which aims to identify social media links in the KB web archive.

So, what did we collect? Well for example, due to the lockdown in the Netherlands there was a problem with surplus food. We captured websites that contributed to the prevention of food waste, like the website benefrietjes.nl that was trying to prevent the waste of potatoes. A related event, which attracted a lot of positive media attention, was the online sale of the excess stock of 20,000 wine bottles by the nuns in Oosterhout. We also collected COVID-19 themed digital books, for example "The Story of Noor and the Coronavirus" by Monique Kerpen (Kerpen, 2020). And the KB has a special web collection on monasteries and religious orders in the Netherlands. Although monasteries are gradually vanishing in the Netherlands, monastic culture was regularly in the news during the COVID-19 pandemic. As many inhabitants of monasteries are old, their communities were at increased risk of being affected by the pandemic.

Peter de Bode: For the Covid-19 special collection, we selected various relevant websites, web pages and tags. There are various themes represented, for example traditions (e.g. the Queen's birthday, Remembrance Day); travel (e.g. Schiphol airport; the Holland-America Line Cruises); government (e.g. the Dutch Government, National Institute for Public Health and the Environment (RIVM)); economy (e.g. Netherlands Chamber of Commerce, Tax and Costums administration); medical (e.g. The Netherlands Red Cross, Lung Foundation Netherlands); sports (e.g. Dutch football (KNVB), Dutch Olympic Committee (NOC-NSF)); criticism (e.g. action group virus madness (Viruswaarheid)); education (e.g. Leiden University, Erasmus University); insurance (e.g. the Dutch insurer Unigarant); social (e.g. SupportYourLocal); and religion (e.g. Kerttijd.nl, a website for inspiration on how to celebrate Christmas)

How do you archive nationally something which is fundamentally global?

Kees Teszelszky: This is a major issue for all web archivists. In the World Wide Web the national web domains are connected more or less to other national domains (except

perhaps the domain of North Korea) and it is challenging to separate them from each other. Also, all human culture is global. Dutch culture is very well connected to and influenced by other global cultural phenomena. Still, we do sometimes miss important online trends and developments in the Netherlands if these are published in more or less closed communities or in languages we do not have access to. This is one of the reasons we started a special web collection of websites published by the Chinese community in the Netherlands with the help of Kitty Lin (Bode, de, Lin & Teszelszky, 2019).

Within the Netherlands, Dutch is the official language, but within the province of Friesland (in the North of the Netherlands) people also speak West Frisian. To capture this regional part of the national (Dutch) web domain, we are closely collaborating with a Frisian heritage institution "Tresoar". Together, we map the Frisian web domain and select important sites about Frisian culture that are published in the Frisian language. For the COVID-19 special collection, I contacted them to collaborate so that we would also be able to capture events within the Frisian domain. This brought an interesting problem: Tresoar offered us more than 600 websites, but at the moment our capture of the Frisian domain contains only 3046 websites. If we were to preserve all 600 websites related to COVID-19, this event would become too dominant within our capture of the Frisian web domain. The KB Web archive is a curated collection, so we decided to select the most important websites for our collection. Our resources are limited.

Peter de Bode: For this COVID-collection, the international perspective is not an issue for the collaborating countries. Each country contributes national websites and the IIPC COVID-collection as a whole is international.

THE FRAME OF THIS SPECIAL COLLECTION

When did you start? When do you (plan to) stop? What is the capture frequency?

Kees Teszelszky: When the first news reached us about the outbreak in Wuhan, we immediately started to monitor what happened within the Dutch web domain. It was important for us to wait until the pandemic clearly started to influence Dutch society and culture. On 27 February 2020, the first case of COVID-19 was confirmed within the Netherlands. This marked the moment the virus had reached the Netherlands, which gave us a reason to start a special collection under the title "Coronavirus COVID-19". On 29 February 2019, we started with selecting core websites and this selection has been crawled every month. We also selected some target crawls that occur once a week.

Peter de Bode: Target crawls are individual archived instances of specific websites. The capture frequency varies. For example, the minutes of parliamentary debates are crawled once, a website that changes its content may be crawled weekly, monthly, bi-monthly or quarterly. Initially the end date for the Coronavirus COVID-19 collection was set at 30

December 2020, but by then it became clear that the pandemic was at its peak. At the moment, collection development is still ongoing.

Kees Teszelszky: COVID-19 is not the first global pandemic, for example in 2009 we saw an outbreak of a new strain of influenza, commonly referred to as “swine flu”, which infected many people around the world. Around that time, the KB was also closely monitoring the Dutch web domain.

Peter de Bode: In that time, web archiving was still in a developmental stage with just one collection specialist. The swine flu or Mexican flu are not represented in our web archive in the same way as we are now capturing the current pandemic.

What has been the amount of data? And the nature of the data (videos, tweets, etc)? And do you have a quantitative overview?

Peter de Bode: As of 18 May 2021 the Dutch COVID-collection holds:

- 345 websites and web pages crawled by the IIPC and the KB;
- 96 websites and web pages crawled by the IIPC (for various reasons not crawled by the KB);
- 41 websites and web pages crawled by the KB (for various reasons not crawled by the IIPC).

Not all my suggestions were accepted by the IIPC.

How did you monitor the nature and amount of data while you were archiving?

Peter de Bode: There is hardly any monitoring during the archiving process. When a crawl is finished, I check the results.

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

Kees Teszelszky: We check the web archivability of the site before we crawl by doing a test crawl.

Peter de Bode: For this special web collection, the first crawl of a website is checked and subsequent crawls are checked during the regular quality control.

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

Kees Teszelszky: In our current online culture, social media is very important. Unfortunately, we are not yet able to archive social media content and we are not allowed to preserve it without notifying the owners. To capture some social media content for our COVID-19 special collection, I made screenshots on my telephone. Unfortunately, I can

only include this within my personal archive. Another restriction is that we cannot archive corona apps, so we collected websites with information about these apps or sites.

As a curator, I feel that I am not only a custodian, who keeps or looks after a collection. I also added to digital culture by tweeting comments on actual developments (see Figure 1).

Figure 1: Screenshot of a tweet by Kees Teszelszky during the COVID-19 pandemic

Peter de Bode: There are challenges, not just for this collection. For example, not all web technologies can be crawled (e.g. Javascript, embedded content, interactivity, filters). And a small number of website owners refused to be included in a web archive.

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

Kees Teszelszky: We started to work from home in March 2020. The transition was easy as we had already been used to it for years, but we miss spontaneous interactions with colleagues. The best ideas for collecting and web archiving start with a coffee (or a beer).

Peter de Bode: Indeed, no problem at all, I can concentrate better at home.

ACCESSIBILITY AND SEARCHABILITY

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

Kees Teszelszky: Due to copyright restrictions, the web archive of the KB is only accessible via a private network within the library building. All data is available for private use and personal study. It is not allowed to reproduce that data in any form (neither on paper nor digitally) without the consent of all stakeholders related to the websites. However, when the

source is acknowledged and all legal conditions are met, then any form of use is permitted by the Copyright Act.

Peter de Bode: The archived web pages with Dutch domains are freely available in the IIPC's Novel Coronavirus (COVID-19) collection. Among others, this selection includes national, commercial, and cultural websites.

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

Kees Teszelszky: Yes, we received several requests and we participated in several research projects on COVID-19 in the Netherlands. One of these projects is the "Archiving COVID-19 Communities" initiated by the CREATE Lab at the University of Amsterdam. Run by Julia Noordegraaf (Professor of Digital Heritage) and Tobias Blanke (University Professor of Humanities and AI), this initiative wishes to archive the corona experiences of citizens (individuals or communities) in the Netherlands. For this purpose, an online platform has been created, where citizens can share their personal experience during the corona crisis.

Furthermore, on behalf of the KB, I participated in the project "Navigating Stories in Times of Transition: The COVID-19 Pandemic as a Use Case" for the 2020 Accelerating Scientific Discovery (ASDI). In collaboration with Professor Gerben J. Westerhof from the University of Twente and, again, Julia Noordegraaf from the University of Amsterdam, we wish to develop eScience technologies to analyse digital storytelling and its evolution over time and across different media. KB also supports the NLnet foundation and Radically Open Security, the world's first not-for-profit computer security consultancy company, on their "Technical review facility to support trustworthy internet technology development in the fight against COVID-19".

How did you communicate about this special collection?

Kees Teszelszky: I have communicated the KB corona collection through a podcast interview in "Onder Mediadoctoren" (Teszelszky, 2020a). An edited version of this interview, explaining how I collected the material, is available at my blog post at Historians.nl (Teszelszky, 2020b). Also the Dutch newspaper NRC paid attention to our COVID-19 collection (NRC newspaper, 2020). Collecting strategies, selection approaches and collaborations with other cultural institutes (e.g. IIPC) regarding the corona collection are available on the KB's website.

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

Kees Teszelszky: We contribute a selection of websites to the IIPC's 'Novel Coronavirus (COVID-19)' collection. This collection is crawled by the IIPC and our collection is part of the IIPC crawls since March 2020.

Peter de Bode: The major reason to start our COVID-19 special collection was the call made by the IIPC, which is crawled by Archive-it. We also collaborated with the Digital Heritage Network in the Netherlands. Together we initiated a call for heritage institutions to help build a national collection about the coronavirus and its effects in the Netherlands.

REFERENCES

Bingham, N., Geeraert, F. (2021). *Exploring Special Web Archives Collections Related to Covid-19: The case of the IIPC Collaborative collection*. Aarhus: WARCnet Paper.

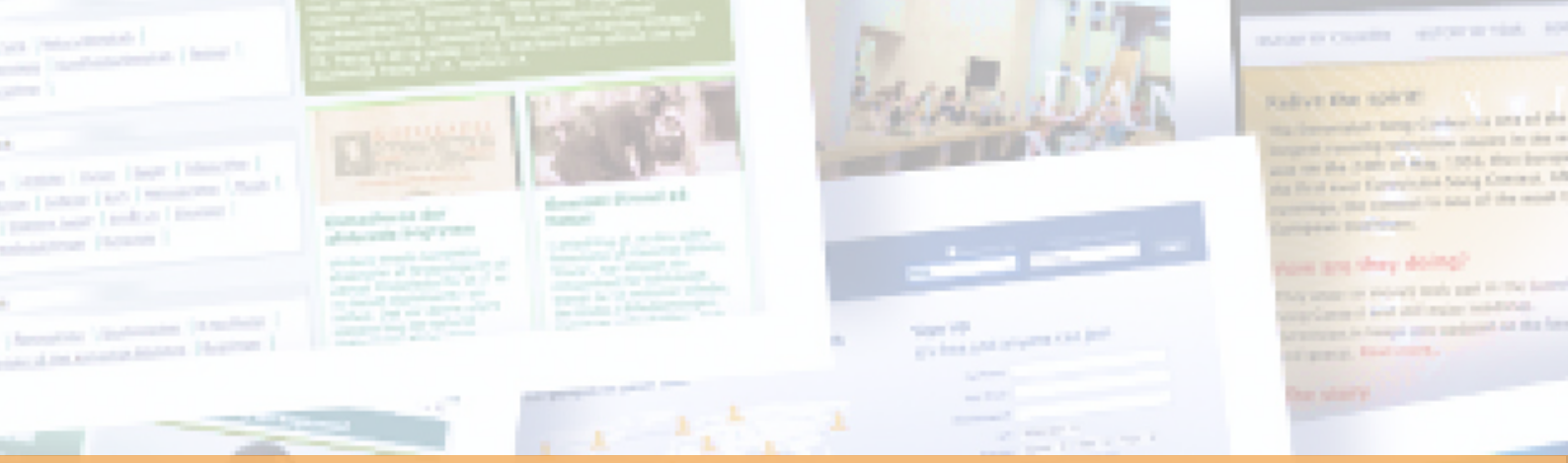
Bode, P. de, Lin, K. & Teszelszky, K. (2019). Chinese Netherlands web collection. *KB Lab: The Hague*. Retrieved from: https://lab.kb.nl/dataset/web-collection-chinese-netherlands

Kerpen, M. (2020) *The Story of Noor and the Coronavirus.*, Venlo: Hoera! Kindercentra & Jan & ko: creatief in onderwijs. Retrieved from: https://balansdigitaal.nl/wp-content/uploads/2020/04/noorenhetcoronavirus.pdf.

NRC newspaper. (2020). Coronacomplotten zijn óók digitaal erfgoed, *NRC newspaper*. Retrieved from: https://www.nrc.nl/nieuws/2020/11/09/complotten-bewaren-voor-later-a4019381

Teszelszky, K. (2020a). Podcast: wat bewaart de Koninklijke Bibliotheek in coronatijd? [Audio Podcast Episode]. In: *Onder Mediadoctoren: Afl 122 (Aantekeningen uit het ondergrondse 18)*. Retrieved from: http://doi.org/10.5281/zenodo.3842015

Teszelszky, K. (2020b). Wat te bewaren voor de toekomst in coronatijden? [Blog post], *Koninklijk Nederlands Historisch Genootschap*. Retreived from: https://www.historici.nl/wat-te-bewaren-voor-de-toekomst-in-coronatijden/?type=bijdrage



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