Chicken and Egg

Aasman, Susan; Bingham, Nicola; Brügger, Niels ; de Wild, Karin; Gebeil, Sophie ; Schafer, Valérie

Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections

Susan Aasman, Nicola Bingham,
Niels Brügger, Karin de Wild,
Sophie Gebeil and Valérie Schafer

WARCNET PAPERS

WARCnet
web archive studies

Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections

Susan Aasman (University of Groningen), Nicola Bingham (UK Web Archives), Niels Brügger (Aarhus University), Karin de Wild (Leiden University), Sophie Gebeil (TELEMME, Aix-Marseille Uni-versity) and Valérie Schafer (C2DH, University of Luxembourg)

valerie.schafer@uni.lu

WARCnet Papers
Aarhus, Denmark 2021

WARCnet Papers ISSN 2597-0615.

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 Collection*s

Published by the research network
WARCnet, Aarhus, 2021.

Cover design: Julie Brøndum
ISBN: 978-87-94108-05-8

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)

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*

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.

Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections

Susan Aasman (University of Groningen), Nicola Bingham (UK Web Archives), Niels Brügger (Aarhus University), Karin de Wild (Leiden University), Sophie Gebeil (TELEMME, Aix-Marseille University) and Valérie Schafer (C^{2}DH, University of Luxembourg)

Abstract: This report is the first in a short series of WARCnet papers which aim to provide feedback on an internal datathon conducted by Working Group 2 of the WARCnet project. It explores the creation of transnational merged datasets and corpora, based on seed lists, derived data and metadata provided by several web archiving institutions. The report highlights our first explorations of specially curated COVID web archives, in order to prepare an in-depth exploration of the issues, challenges, limitations and opportunities afforded by these heterogeneous datasets.

Keywords: COVID crisis, datasets, distant reading, datathon, seed lists, derived data, metadata, European web archives

> "If the question of the priority of the egg over the hen or the hen over the egg troubles you, it is because you assume that the animals were originally what they are now. What madness!"
>
> Denis Diderot, *The Dream of d'Alembert*, 1769 (our translation)

Working Group 2 (WG2) of the WARCnet project[1] is dedicated to the study of transnational events — both planned and unplanned, through web archives. Its work began at the same

[1] For more information on the WARCnet project, see: https://cc.au.dk/en/warcnet/about/. For more details on WG2, see: https://cc.au.dk/en/warcnet/working-groups/.

time as the early days of the COVID crisis, therefore the team chose to look at the COVID collections being created across Europe in web archiving institutions (Schostag, 2020). The pandemic provided an opportunity to analyse the way an unforeseen event is considered by European web archives and to examine how web archiving institutions organise themselves in the face of emergency, how they select, delimit and arrange their special collections, and how useful these datasets are for social science and humanities researchers.

A series of coordinated oral interviews following standardised interview guidelines was thus carried out with several archiving institutions and web archivists.[2] The aim of this series of interviews was to enable researchers to analyse the COVID data with an enhanced understanding of the chosen perimeters and parameters, archiving processes, type and format of the data etc. Our aim was also to provide a deeper understanding of what is required in terms of documentation and contextualisation for the present and the future, when "*Doing History in the Digital Age"* (Brügger, 2018) and to sustain the analysis of these special collections. Moreover, the work highlighted the diversity as well the commonality shared by disparate archival practices, making it possible to think about issues such as inclusiveness, values (Schafer and Winters, 2021), governance and the curational choices made by archival institutions, their means (human and technical) and many other issues related to Web archiving.

WG2 had ambitions to go beyond this "behind the scenes" approach, however rich and interesting it may be, and the second phase was to look at the collections themselves. Several possibilities are open to researchers when exploring topics based on web archives, the most obvious being to physically visit institutions such as Bibliothèque nationale de France (BnF), Bibliothèque nationale du Luxembourg (BnL), the UK Web Archive at the UK Legal Deposit Libraries,[3] etc., to look at the collections individually (few of these national institutions offer remote access,[4] except for the Royal Library of Denmark or arquivo.pt). However, travel restrictions due to the health crisis coupled with a motivation to think in terms of transnational corpora and to work collectively (rather than individually and on a national basis) led us to favour a different approach, namely, to ask institutions to supply us with what data they were able to in the context of a collective online event such as a datathon to be initiated for members of WG2. And this is where "the Chicken and the Egg" comes in: where to start the research process? We engaged in discussions on the various approaches which could be taken, between data-driven science and research-driven questions.

[2] Transcriptions of this series of oral interviews are available at https://cc.au.dk/en/warcnet/warcnet-papers/. You may find interviews with web archivists working at the French National Audiovisual Institute (INA), the French National Library (BnF), the IIPC (International Internet Preservation Consortium), the National Library of Luxembourg (BnL), Netarkivet (Denmark), the National Széchényi Library in Hungary, the UK Web Archive, the Swiss National Library and the Icelandic Web archive. Other interviews are scheduled.

[3] The UK Legal Deposit Libraries are the British Library, the National Library of Scotland, the National Library of Wales, Bodleian Libraries, Oxford, Cambridge University Library and Trinity College, Dublin.

[4] The remote access may be conditioned by a previous request and agreement with the library.

These are some of the questions that this WARCnet report aims to address, from the first steps of our requests to the web archiving institutions, to the intermediate results we obtained at the end of the three meetings held on 22 January, 26 January and 10 February 2021, while also taking into account the feedback from a presentation given during the general WARCnet meeting in Aarhus held on 21 April 2021.[5]

DATATHON APPROACH

Research in web archives have become increasingly transdisciplinary and the history of the Web can no longer be written solely by historians working in isolation. Both the vast scale and the technical complexity of archived web resources invites not only new research questions but also multiple challenges. As Nick Ruest et. al (2021) recently summarized, this challenge encompasses "size on the order of petabytes, billions of words, tens of thousands of images, all with murky metadata, provenance, and difficulty to access". Therefore, within this paper we propose the datathon as a model for effective transdisciplinary collaboration, idea generation and group learning. A datathon is a short-time but highly intensive meeting in which a group of participants works on a shared research problem. In our case, it allowed us to bring together web archivists and scholars with various disciplinary background (data science, media studies, (art) history, cultural studies).

The datathon approach is beneficial for the study of web archives. While data and medata is collected by web archives (and some derived data may also be produced), it is not yet readily available for researchers in machine-readable format. For access to the data, but also to become aware of the uncertainties and information gaps within datasets derived from web archives, it is essential that web archivists and scholars from various disciplinary backgrounds collaborate closely. In our case, we closely collaborated within a small group of experts (the authors of this paper, joined by author colleagues like Friedel Geeraert, Frédéric Clavert and Katharina Schmid at some point). The idea was that multiple-expertise viewpoints would aid in the development and critical review of the merged dataset. A transdisciplinary group can offer various perspectives on how to study data and design a dataset, as well as it can be beneficial in finding potential solutions to research problems. Continuous peer review is also productive for creating more reliable datasets. Multiple objective "eyes" can be valuable to identify inadequate data classifications, inconsistencies in the content or the limitations of the collections. Therefore, working with a diverse team in which some use their advanced technical skills taking the lead in some stages of the project, while others are more oriented towards theoretical reflections, or are able knowledgeable in web archival practices, can stimulate a high learning curve.

Our group adopted an iterative process and hosted several sessions of datathons for the creation of a dataset about Covid-19 special collections. The first two-hour datathon (25 January 2021) was organized to present the data provided by various web archives. The

[5] For more details on this WARCnet meeting, see: https://cc.au.dk/en/warcnet/events/view/artikel/london-2021/

aim of this meeting was to discuss research potential and limitations of the data and to explore possibilities for combining the data into a transnational corpus. On 28 May 2021, a follow-up two-hour datathon was held for a collaborative and interdisciplinary review of the created transnational dataset. The group reflected on quality issues and explored various forms of analysis of the dataset (both qualitative and quantitative). Before and after the datathons, participants continued to work in smaller groups to prepare for the next meeting.

COLLECTING DATASETS

Why and what we asked for?

By combining the expertise of web researchers and archivists, the WARCnet project aims to consider, stimulate, and facilitate transnational research in web archives with a European dimension. Therefore, one of our objectives was to create a concrete test bed to evaluate the possibility of creating transnational corpora. It was also a question of trying to push back the limitations of copyright and legal deposit by inviting institutions to think about the data they could provide outside their walls. We obviously did not expect to be able to access content, but derived data and metadata in most cases, however this would still allow us to test our own practices and approaches. Accordingly, the main goals were threefold:

(1) To create a sandbox and concrete test bed,

(2) To conduct a first round of analysis to determine what could be achieved with heterogeneous data and test how a shared corpus based on them could be created,

(3) To document our experience of working with heterogeneous, cross-national datasets, with a view to feeding back to web archives and documenting the process.

We therefore asked several WARCnet project partner institutions to provide us with COVID-19 related data. Some biases were induced by our request to the institutions. In some cases, we sent a rather vague request for data to be used in a WARCnet datathon — the words "data" and "datathon" already inducing a bias (Milligan et al., 2019), while in other cases we had a more precise request for metadata.

Several interesting aspects about this approach can be highlighted: in the case of some institutions such as the Bibliothèque Nationale du Luxembourg the data were directly retrieved online (seed lists were made available on the webarchive.lu website). With other institutions, our request was deliberately open; we asked to be supplied with whatever they could provide. For some requests, such as to the Royal Library of Denmark, the call was from the outset more focused on seed lists. Some requests were also the subject of queries from web archivists who sought clarification on the nature of our needs, and this open approach allowed for exchanges and refinement, as in the case of the INA (Institut National de l'audiovisuel, France), whose collecting activity focused on Twitter (with a first study by Blanckemane, 2020, which already provides vivid results). The datasets were in some cases provided with minimal information, while in other cases, such as that of the BnF, they arrived with substantial documentation (statistics, a description of the whole COVID-19 collection, etc.). Finally, our requests were made possible by the privileged relations created over several years with these institutions and the fact that they are also partners in the WARCnet project. In some cases, the use of the datasets was subject to precise conditions of use for a certain period and purpose. Thus, some datasets were only to be used and kept for the time of the internal datathon, which also raises questions about citation and the accessibility of a transnational corpus at a time when calls for FAIR Data[6] in the scientific community are increasingly important. In addition, other complementary data could certainly be obtained by refining the request to a more precise topic (we will revisit the case of the INA Twitter dataset, for which we requested a very small sample as the entire dataset INA preserved and could provide, would have been unmanageable) and by setting up a more specific convention and legal framework. Finally, it should be emphasised that these data are not exclusive and that other datasets and initiatives (numerous in the case of the COVID-19 crisis) have been carried out by researchers (i.e. a Twitter collection by Frédéric Clavert for example), by various GLAM and research institutions (see News Media Tweet Dataset from Universitat Autonoma de Barcelona), or with the collections created in Archive-It.

What we received

Several institutions responded positively to our call which allowed us to complement the datasets available online, in the case of the BnL (Luxembourg), which shared its seed list[7] and the *National Széchényi Library* (Hungary)[8] with several other datasets. We received data (mostly derived data and metadata) from the IIPC (the International Internet Preservation Consortium), the French National Library (BnF), the National Audiovisual Institute (INA), the UK Web Archive, the Danish Royal Library and the KB (Royal Library) in the Netherlands. Several elements are striking from the first exploration:

(1) Some data arrived with no contextual information, while others were well-documented by web archivists. The BnF for example attached very detailed documentation as well as statistics to help us study its data. We did not ask for detailed documentation, which may explain why some lists arrived alone, but the web archivists at BnF tried to foresee our needs when analysing their datasets.

(2) More generally, as mentioned earlier, there were enquiries by web archivists about our precise needs and research questions, in order to try to select the relevant data and provide a well-suited selection. Due to lack of clarity on what can and cannot be done with the data and questions about the format in which it should be shared, the process of sharing data is relatively underdeveloped in archiving institutions. Many national institutions operate under legal deposit regulations which include

[6] FAIR stands for Findable, Accessible, Interoperable, and Reusable (see Wilkinson et al. 2016 and Mons 2018).

[7] See https://www.webarchive.lu/covid-19/.

[8] For more details, see https://webarchivum.oszk.hu/en/webarchive/browse/browsing-in-the-event-based-subcollections/browse-coronavirus-epidemic-2020/.

stipulations preventing data being shared outside of the archiving institution. The UK Web Archive, for example was unable to share raw WARC files due to a combination of legal deposit restrictions and the complexity of extrapolating a subset of thematically grouped WARCs, as they are not stored this way in the file system. Instead, a spreadsheet was exported from the archives' Annotation Curation Tool, containing a list of collection metadata. A limited amount of data cleaning was carried out by the UKWA beforehand such as redacting the curator's names, however, due to limited time and resources, it was not possible to fully prepare the spreadsheet before exporting. It was, therefore, still quite untidy with several fields running across the columns as the CSV format separated text into separate fields when commas had been used in the original database. The IIPC Content Development Group were also happy to share metadata for their collaborative Covid-19 collection in the form of a spreadsheet containing the seed list and descriptive seed metadata (URL, Top-Level Domain, Title, Description, Website Type, Country of publication, Language(s)). Although not subject to legal deposit regulations, sharing the WARC files was still problematic due to the large size of the collection (3.6 TB) and it would have been challenging to find a WARC sample that would be truly "representative" of the whole collection. Again, a fair amount of preparation was necessary before exporting the metadata, as the nomination spreadsheet contained the identification of IIPC members and public contributors who selected the seeds. The web archivists also took the decision to exclude nominated seeds that had been rejected during the seed review process as they were not sure that this information would be useful to the researchers. Finally, the IIPC archivists had two requests; that at the conclusion of the datathon participants should delete the seed list and that any concrete outputs would be shared with the archiving institution/ consortium that had put the seed lists together.

(3) The data provided by INA, consisting of a collection of Tweets collected through the Twitter public API, was different to the other datasets in that it contained the text of the Tweets, or the content, furthermore, the files were in the JSON file format. An interview with Claude Mussou during our first meeting together with the WARCnet paper dedicated to the INA's special archiving during the COVID crisis (Schafer, Thièvre and Blanckemane, 2020) provided insight into INA's choice to focus on Twitter, in which the institute has real expertise, beginning in 2015 when it created a special collection dedicated to the French terrorist attacks. In the case of the Twitter collection, there were no seed lists or URLs, however there was access to the content. This allowed for keyword searches, network analysis, text mining etc. to be carried out on the sample, whereas the other institutions essentially provided an overview of the collection structure, such as seed lists and metadata, but not the content obtained. There did not seem to be an obvious method to combine the JSON dataset with the other datasets provided in the Excel or CSV format. Luckily, the JSON file could be converted into CSV, allowing its potential integration into a unified file after cleaning the various datasets, but it did not follow the same entries as the seed lists.

Beside the apparent interoperability of the datasets, other limitations appeared which led us to explore three preliminary questions:

(4) What are the strengths and limitations of the datasets?

(5) How could we combine the datasets and create a common corpus? Could we create a unified standard format for compiling all data in one data source?

(6) How could we search/analyse these data (e.g. distant reading, hyperlinks mapping, etc.)?

STRENGTHS AND LIMITATIONS

A first round of discussion focused on the strengths and limitations of the datasets. Even before trying to combine them, the aim was to address their salient points in concrete terms, while being aware that other limitations and assets would emerge in the course of the research process. In this context, the strengths and limitations did not only concern the datasets but – as we previously addressed – also the team's own capability, as the team was made up of heterogeneous levels of digital skills and data literacy (Milligan, 2020). For instance, being confronted with CSVs in the majority of cases but also with INA's JSON files stimulated us to explore techniques that could merge and combine the data and to discuss the issue of their interoperability. Besides this, there was also the question of the diversity of the preserved sources and formats and the need (or not) to isolate for example social network sites from websites, as well as the identification of gaps in the datasets that did not consider for example TikTok or other social media. Indeed, these lists also made it possible to identify absences and silences (or at least a kind of built-in, inherent/natural delay in acknowledging new emerging platforms and their societal impact) in the web archives.

Limitations related to research infrastructures and research data management were also obvious such as the computing power of our equipment and the slowness of conventional computers to compile the data. Questions about temporary storage, responsibility for and security of the data and its collective sharing also arose (e.g., which infrastructure to use, was the server secure? etc.).

Other very pragmatic considerations emerged from this first observation: a dataset such as the one provided by Hungary raised the issue of multilingualism since nobody in the group speaks Hungarian, making it difficult to read and understand the URLs. More broadly, the question of knowledge of the Hungarian web sphere arose. A researcher who does not know anything about this web sphere will find it difficult to address the issues of curation, the choice and representativeness of the dataset. Consequently, and despite clues in the dataset, the lack of knowledge of the context can be problematic. Although this was partly compensated by the interview conducted by Friedel Geraaert with Marton Németh, it remained a real obstacle. The question of documentation of the datasets also arose. Interviews are one means of deepening understanding, but the statistical data and precise documentation provided by the BnF also made it much easier to engage with the data. The key question, once it has been stressed that documentation of datasets is necessary, remains: whose task is it? (cf. the discussion in Brügger, 2018, 137-139) Should it be something done by web archivists alone or with the input of academics? What kind of documentation is needed? Can a common template be applied?

The question of a common template also arises with/when cleaning and structuring the various datasets. One of the advantages of such datasets is the ability to add most of the seed lists and information to an Excel file and to make comparisons in terms of the number of URLs collected, the main thematic areas to which the URLs relate, etc. However, very quickly one becomes aware that information that is part of one dataset is missing in other datasets (either in relation to others to allow comparisons, or more generally: for example, are websites regularly harvested or not? What is the period of crawling? etc.). Moreover, these elements provide information about the intent of the archiving institution but do not allow one to assess the quality of the captures and results. Nothing is communicated either, in terms of image files or MIME types, the datasets are exclusively textual, although without allowing for a precise idea of the content (except for the INA dataset) and the presence of visual and audio elements etc.

In addition, we had small datasets which may in some cases have provided us with distorted information. INA for example made the choice to use samples because there was no precise research question to inform a selection of hashtags and results, and otherwise it would have been necessary to deal with a mass of data of several million tweets.

This issue may also lead to another one: the choice to study a special collection raises the question of the reproducibility of the experiment in several respects. In brief, the current testbed project benefited from the fact that what the WARCnet team of WG 2 considered an event — the COVID-19 crisis — which was also acknowledged as an event by web archives and therefore numerous curated datasets already existed. In many other cases, this may not be the case and researchers must then delimit their collection based on the web archives general collections. However, it is also worth noting that although a pre-curated collection may already exist, it will not necessarily be consistent with how a researcher may want to delimit what is deemed relevant to cover the event and s/he may want to include supplemental material.

Finally, despite the enthusiasm of the group to engage with the data in order to explore them, there also remains, as we have underlined, digital literacy issues for Social Sciences and Humanities researchers who would like to approach the data. Moreover, an exploration of derived datasets and metadata, however useful it may be, is certainly not the most attractive approach for some researchers. There is also the issue of how to make these elements readable when it comes to sharing them to turn them, for example, into access points for neophyte researchers interested in the subject. This is also in line with the second question that the group decided to address: namely what to do with the datasets?

WHAT WE CAN DO AND WHAT WE WOULD LIKE TO DO

Compiling datasets in one data source

Although (most of) the received datasets are simply seed lists of what each web archive wanted to archive, and therefore contain very little information, it is possible to perform some analyses that can constitute results which can help investigate the possibility of studying web archives across borders.

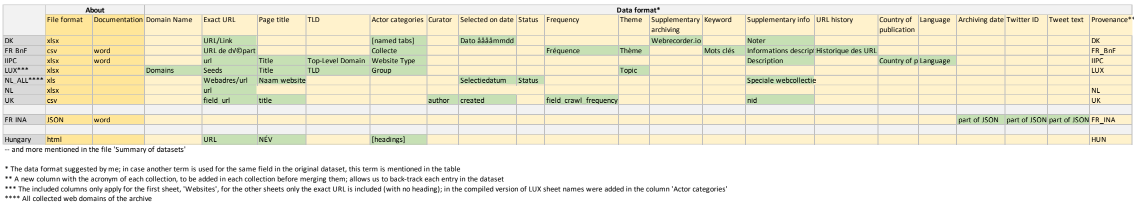
As most of the datasets include URLs, creating a unified standard format for compiling all data in one data source seemed an obvious place to start. The idea was therefore to:

- map the different formats (excel sheets (LU, DK), with different categories on tabs in the excel sheet (LU, DK), etc.),
- define what a unified format may include,
- ensure tracking of the provenance of the data, that is: which original dataset does each entry come from?
- operationalize this.

Data about the time span covered in the different collections also seemed relevant as the oral interviews highlighted different policies. Some collections started earlier than others, some are currently still running, others were defined for a precise period of time and had to end at the first wave of the crisis to focus on other collection and harvesting priorities (this was for example the case at the BnF, see WARCnet paper by Gebeil, Schafer, Benoist, Faye and Tanesie, (2020). As previously highlighted, the datasets we received had both similarities and differences. Except for the dataset from INA, all the datasets were seed lists, that is lists of the web domains that were selected to be archived. But even within the seven datasets with seed lists, the format of the data showed a high degree of heterogeneity regarding which type of information was included (cf. the Annex). Therefore, both cleaning the data and merging them were needed.

Since one of the aims of this pilot project was to investigate the possibility of studying web archives across borders, the first step was to get an initial overview of the structure of the datasets, to use this as a stepping stone to decide a common data structure and then to transform the datasets to fit this structure. This had to be done in a way that was flexible, robust, enrichable, and backtrackable: flexible, since it should allow for analysing datasets individually, more datasets, or even merging them; robust, since it must be something that can be handled in standard software, such as Excel, R, or similar; enrichable, since there may be a need to add information to the original datasets; and backtrackable, due to the requirement to identify the provenance of each domain name in the seed lists (individual or merged).

Table 1 provides a mapping of the datasets, with the type of data in each dataset marked by a green cell. It is worth noting: (a) that most seed lists were in comma-separated format (xlsx, csv, xls) while only one was in the form of an html page, (b) that all datasets were unique in their data structure, but that they also had types of information in common, (c) that they all shared one piece of information: the exact URL that was intended to be archived.



	About		Data format*																			
	File format	Documentation	Domain Name	Exact URL	Page title	TLD	Actor categories	Curator	Selected on date	Status	Frequency	Theme	Supplementary archiving	Keyword	Supplementary info	URL history	Country of publication	Language	Archiving date	Twitter ID	Tweet text	Provenance**
DK	xlsx			URL/Link			[named tabs]		Dato ååååmmdd				Webrecorder.io		Noter							DK
FR BnF	csv	word		URL de dV©part			Collecte				Fréquence	Thème		Mots clés	Informations descript	Historique des URL						FR_BnF
IIPC	xlsx	word		url	Title	Top-Level Domain	Website Type								Description		Country of p	Language				IIPC
LUX***	xlsx		Domains	Seeds	Title	TLD	Group					Topic										LUX
NL_ALL****	xls			Webadres/url	Naam website				Selectiedatum	Status					Speciale webcollectie							
NL	xlsx			url																		NL
UK	csv			field_url	title			author	created		field_crawl_frequency				nid							UK
FR INA	JSON	word																	part of JSON	part of JSON	part of JSON	FR_INA
Hungary	html			URL	NÉV		[headings]															HUN

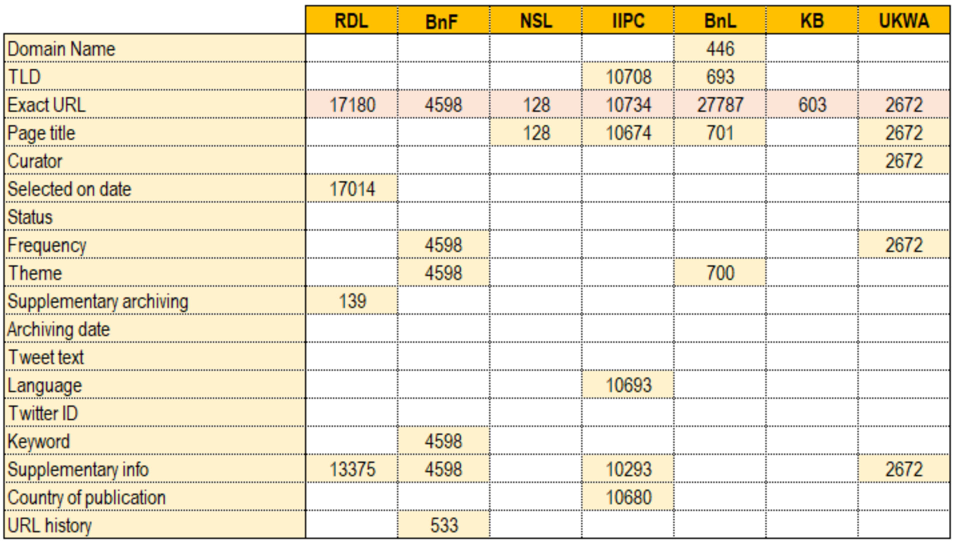
-- and more mentioned in the file 'Summary of datasets'

* The data format suggested by me; in case another term is used for the same field in the original dataset, this term is mentioned in the table
** A new column with the acronym of each collection, to be added in each collection before merging them; allows us to back-track each entry in the dataset
*** The included columns only apply for the first sheet, 'Websites', for the other sheets only the exact URL is included (with no heading); in the compiled version of LUX sheet names were added in the column 'Actor categories'
**** All collected web domains of the archive

Table 1: Mapping of datasets (full size in Annex)

Based on this initial overview, new column names were suggested to cover the totality of the information found in the datasets (heading row in Table 1), and to be able to backtrack to the original name of the corresponding column in the original dataset. The original name was added in the green cells, e.g. a new column named ‘Actor categories’ was suggested, which in the existing datasets correspond to ‘a named tab’ (DK), ‘Collecte’ (BnF), ‘Website Type’ (IIPC), ‘Group’ (LUX), and ‘headings’ (Hungary), respectively. Finally, a new column was added — ‘Provenance’ —showing the acronym of each institution or collection. This column had to be added for each collection before merging since it allowed backtracking to the original dataset and allowed for filtering of the dataset if only one collection was to be studied.

As the dataset was relatively small, it could be processed in Excel to cleanse the data. Since the exact URL was found in each dataset, this constituted the most valuable information for making cross-national studies (see figure 2). However, the exact URL may not have been needed for all studies, and may even have constituted ‘noise’, because it contained too much information. Instead, it would have been useful to have the domain name and the top level domain (TLD), information which could be extracted from the URL. Table 2 shows that some of this data was provided by web archives, but that it was incomplete. Therefore, we still extracted this information from the URL.



	RDL	BnF	NSL	IIPC	BnL	KB	UKWA
Domain Name					446		
TLD				10708	693		
Exact URL	17180	4598	128	10734	27787	603	2672
Page title			128	10674	701		2672
Curator							2672
Selected on date	17014						
Status							
Frequency		4598					2672
Theme		4598			700		
Supplementary archiving	139						
Archiving date							
Tweet text							
Language				10693			
Twitter ID							
Keyword		4598					
Supplementary info	13375	4598		10293			2672
Country of publication				10680			
URL history		533					

Table 2: Overview of the data entries provided by each web archive[9]

In the appendix, the process of cleansing the data, and extracting domains is explained step-by-step.

What can one study with these data?

Ideally, this kind of unified table and merged dataset may enable a first discovery of the special collections, for example, what is available, retrievable and searchable through European web archives. Limitations of search are addressed within the field (e.g. Winters and Prescott, 2019).[10] Analysing the seed lists may be a first entry point into web archives, as it would for example allow a researcher to study:

(1) Web archives archiving outside of their own ccTLD

- how many URLs in-/outside ccTLD?
- which ccTLD/gTLD outside ccTLD? — and how many of each?

(2) The types of actors

- Which categories of actors are archived? How many of each?
- Different categories are used, but there are also overlaps.

[9] The data was provided by the following Web archives: RDL (Royal Danish Library); BnF (Bibliothèque nationale de France); NSL (National Széchényi Library, Hungary); IIPC (International Internet Preservation Consortium); BnL (Bibliothèque nationale du Luxembourg); KB (Koninklijke Bibliotheek, The Netherlands); UKWA (UK Web archive).

[10] On retrieving lost data, web content and web sphere, the reader may also refer to Nanni, 2017, Huuderman et al., 2015 and Ben-David, 2016.

(3) New event-specific websites (one could run all URLs against the Internet Archive to check how many websites existed before the event started, and how many did not, and then have a closer look at the latter category).

These questions may be combined with broader research questions relevant to the WARCnet project as well as for the study of heritagization and web archives more generally, may it be with regards to inclusiveness, values and practices and how they are entwined and "negotiated" (Schafer et al., 2016):

(1) How to make an entry point for a researcher through European COVID collections? Why datasets may be useful to guide him/her?

(2) Can this kind of tables we produced highlight different methods/approaches/policies of creating COVID collections in European countries and more generally the practices of web archiving institutions as well as uncovering noises and silences in the collections?

From a cultural and governance perspective, it is very interesting to explore if web archives can have a political agenda comfort instance by testing web archiving institutions' experience, governance and practices with the reality of the datasets and the context of the data curation we had partially explored through oral interviews (Maemura, 2021). On the other hand, from a web archival perspective, we could explore how to integrate researcher feedback to adapt collection policies, especially when dealing with ongoing events like COVID-19?

Before analysing this dataset, it was essential to gain a better understanding of the data provided by each web archive. At this stage, we could create an overview of the URL's and domain names in the seed lists of each web archive (see figure 1). To further contextualize, some web archives provided documentation about their Covid-19 special collections (see table 3). Yet, the documentation the web archives provided was heterogenous and often too limited to be able to analyse the data (as discussed in the previous section 'What we received'). For historical data, it is essential that detailed provenance information is available.[11] Also in general, granular descriptive metadata about (items within) the collections offer essential contextual information, when this is missing it is very difficult (maybe even impossible) to draw meaningful conclusions (Di Pretoro and Geeraert, 2019).

[11] Using Linked Open vocabularies may offer mechanisms for syncing metadata across Web archives and making it interoperable. For that reason, Rhizome's Artbase (the Web archive for Internet art) describes their metadata and provenance in Linked Open Data. See, for example Rossenova, Wild and Espenschied 2019.

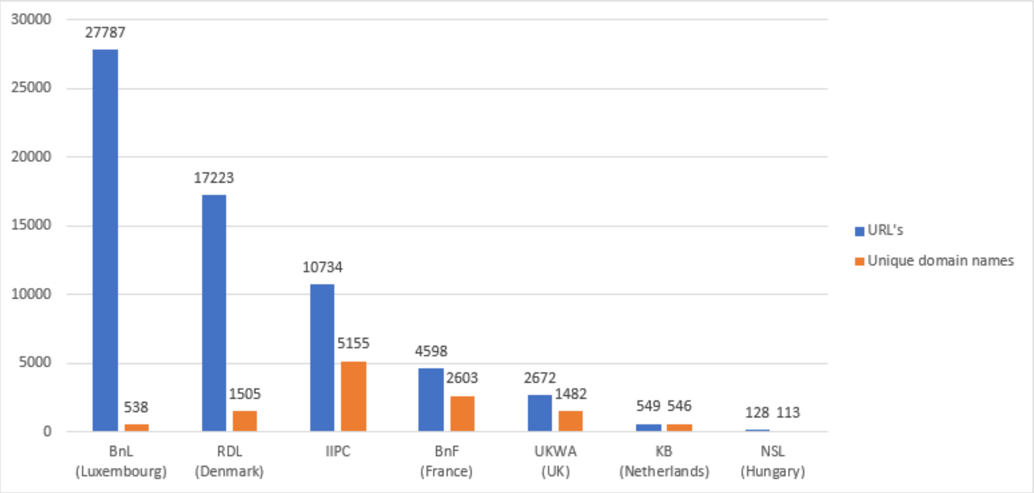
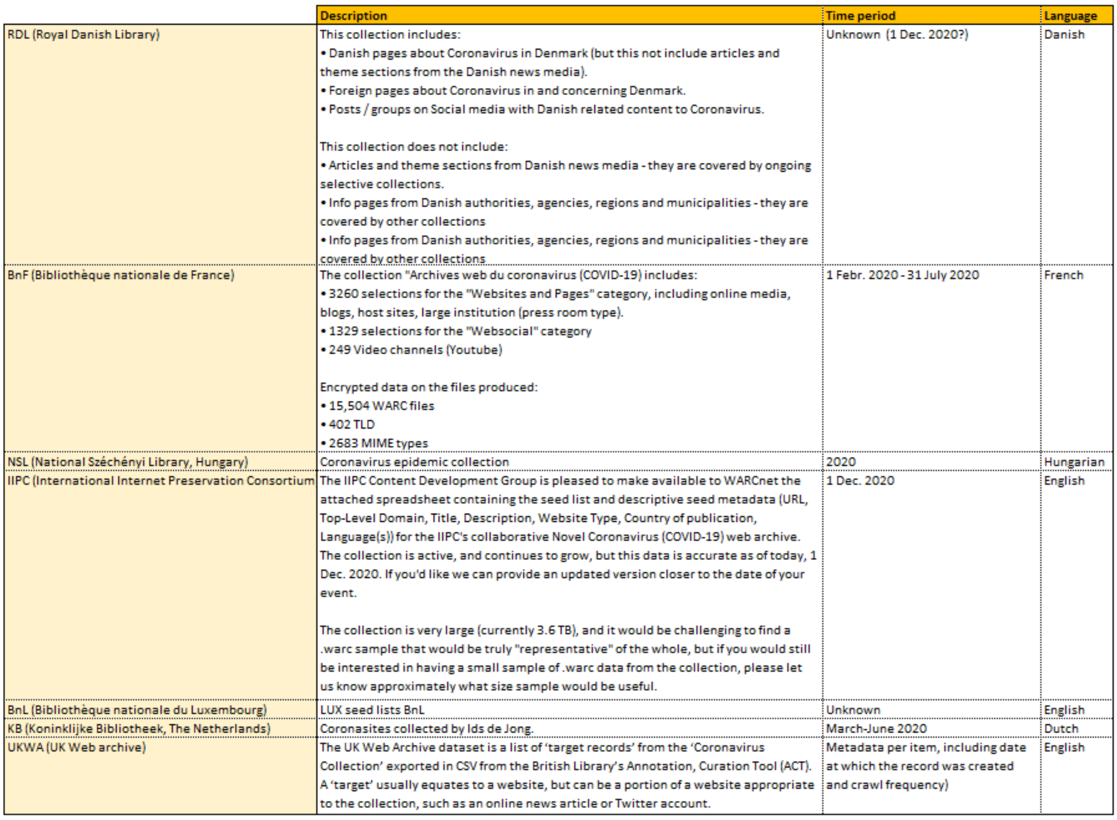


Figure 1: URL's and unique domain names in each web archive



	Description	Time period	Language
RDL (Royal Danish Library)	This collection includes: • Danish pages about Coronavirus in Denmark (but this not include articles and theme sections from the Danish news media). • Foreign pages about Coronavirus in and concerning Denmark. • Posts / groups on Social media with Danish related content to Coronavirus. This collection does not include: • Articles and theme sections from Danish news media - they are covered by ongoing selective collections. • Info pages from Danish authorities, agencies, regions and municipalities - they are covered by other collections • Info pages from Danish authorities, agencies, regions and municipalities - they are covered by other collections	Unknown (1 Dec. 2020?)	Danish
BnF (Bibliothèque nationale de France)	The collection "Archives web du coronavirus (COVID-19) includes: • 3260 selections for the "Websites and Pages" category, including online media, blogs, host sites, large institution (press room type). • 1329 selections for the "Websocial" category • 249 Video channels (Youtube) Encrypted data on the files produced: • 15,504 WARC files • 402 TLD • 2683 MIME types	1 Febr. 2020 - 31 July 2020	French
NSL (National Széchényi Library, Hungary)	Coronavirus epidemic collection	2020	Hungarian
IIPC (International Internet Preservation Consortium	The IIPC Content Development Group is pleased to make available to WARCnet the attached spreadsheet containing the seed list and descriptive seed metadata (URL, Top-Level Domain, Title, Description, Website Type, Country of publication, Language(s)) for the IIPC's collaborative Novel Coronavirus (COVID-19) web archive. The collection is active, and continues to grow, but this data is accurate as of today, 1 Dec. 2020. If you'd like we can provide an updated version closer to the date of your event. The collection is very large (currently 3.6 TB), and it would be challenging to find a .warc sample that would be truly "representative" of the whole, but if you would still be interested in having a small sample of .warc data from the collection, please let us know approximately what size sample would be useful.	1 Dec. 2020	English
BnL (Bibliothèque nationale du Luxembourg)	LUX seed lists BnL	Unknown	English
KB (Koninklijke Bibliotheek, The Netherlands)	Coronasites collected by Ids de Jong.	March-June 2020	Dutch
UKWA (UK Web archive)	The UK Web Archive dataset is a list of 'target records' from the 'Coronavirus Collection' exported in CSV from the British Library's Annotation, Curation Tool (ACT). A 'target' usually equates to a website, but can be a portion of a website appropriate to the collection, such as an online news article or Twitter account.	Metadata per item, including date at which the record was created and crawl frequency)	English

Table 3: Descriptions of the Covid-19 special collections

Consequently, we cannot provide in depth insights yet, but we can provide some examples of how one can retrieve preliminary entry points for further analysis. For example, we could discern how the Danish Royal Library collected outside their national domain, providing researchers with access to websites from various country domain names (see figure 2). The outliers in the data could be used as entry points for further research, for example why did they collect a relatively large number of websites from Germany? Were there more national web archives collecting the German country domain name (see figure 3)? We not only need more (meta)data to further explore these questions, but ideally, data visualisations should also be prepared within an interdisciplinary team. This more in-depth analysis will be the next step within our research group.

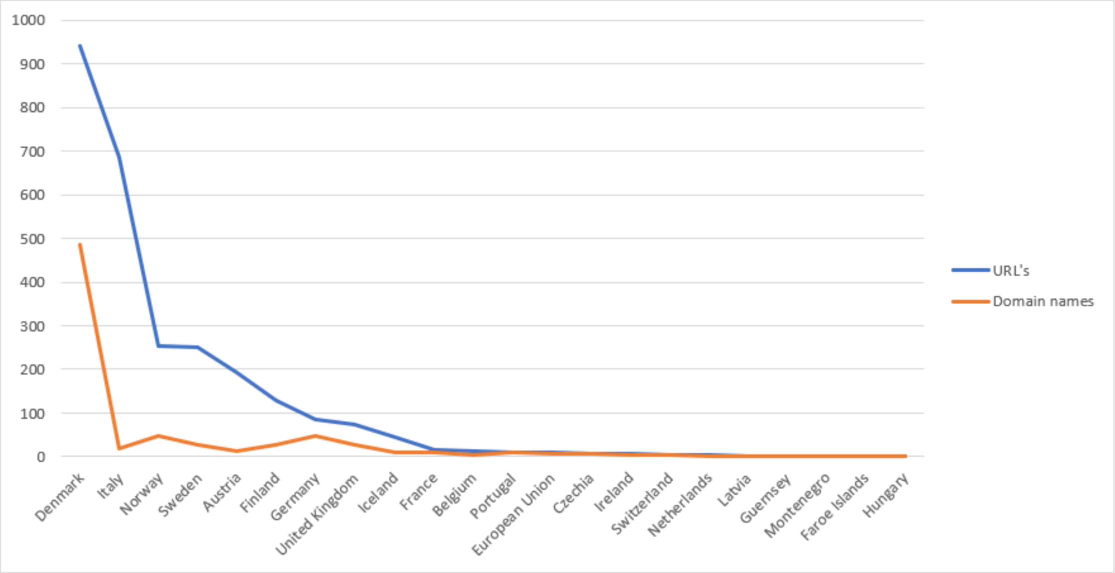


Figure 2: European ccTLD in the Covid-19 special collection of the Royal Danish Library

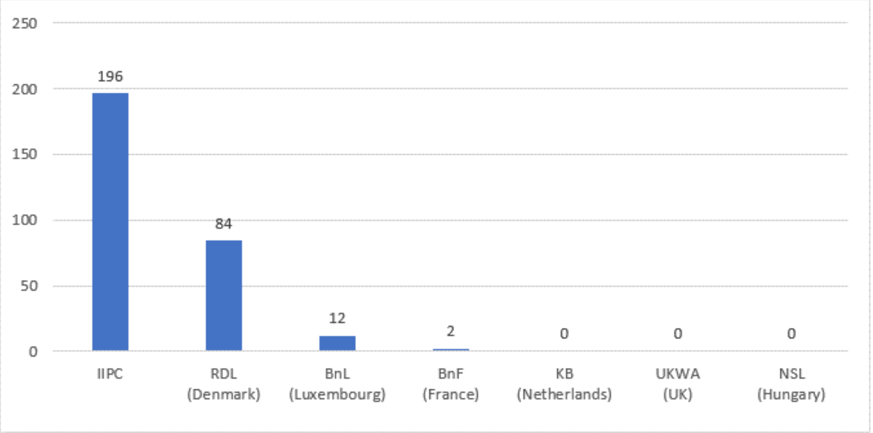


Figure 3: Total count of .de top level domain names in Covid-19 special collections

Of course, there are gaps between what we can and what we would like to study when it comes to a collective group, with several shared but also different interests and research questions. Furthermore, at some point, access to the content and WARC files is needed, when for example conducting research on a precise event within the web archives, for example, women, gender and COVID, or to retrieve online journals of lockdowns. The seed lists alone do not allow us to perform visual analysis (e.g., what type of visual communication about COVID is used of the web: e.g. virus visuals and use of colour). It may help to start an analysis of the COVID collections with reference to other pandemic collections, of governmental sites about COVID, or to follow the evolution of online museum collections and their communication during the pandemic crisis. The challenge of studying everyday life/amateur media practices and to combine different social media expressions providing insight into the ways people deal with COVID in their everyday life (incl. Twitter, YouTube, TikTok, Facebook and Instagram) is more complex and may necessitate other entry points, which are more visual or content-oriented.

CONCLUSION

This project provided a reality check which allowed us to gain an overview of national and transnational corpora available in web archives. It gave us the opportunity to combine metadata, derived data and seed lists to achieve a preliminary overview of some but also diversity of the datasets and to reflect on the choices made by web archiving institutions.

We have come to a few intermediary recommendations/conclusions that are strongly intertwined with daily practices of web historians.

Historians and archivists as natural partners

The emergence of web archives does not change this principle, but it does challenge existing knowledge about record keeping procedures, (trans)national datasets and research agendas. This study provides some insights and reflections about the issues related to processes of heritagization of unforeseen events through born-digital sources and “living archives” (Rhodes, 2013; Rollason-Cass and Reed, 2015).

The rhythms of data collection and the domains' selection made by web archiving institutions make it possible to identify approaches that are intended to be broad and inclusive, and that take place over several months, while the COVID crisis extends over time and the waves follow one another, which the oral interviews had already made it possible to understand. The budgets, the size of the teams, but also the urgency of other special collections or the greater or lesser capacity to integrate the social platforms, which have specific archiving characteristics, all play a role in the content (scope, number, duration) collected. Secondly, the way in which the data is documented, the choice of metadata and the entries by theme, date, etc., while relatively homogeneous, nevertheless present differences that also reflect the organizational and intellectual structure of the web archiving institutions. Legal issues such as access to data are also at the heart of the variety of approaches. Legal issues and institutional policies also come into play when coupled with questions related to the storage, sharing and preservation of data by researchers. They are also posed in terms of paradoxical injunctions in relation to open science, transparency, FAIR data, etc (Truyter, 2021).

Datathon as a way to share expertise and skills to be developed

One of the characteristics of our datathon was that we were both collectors requesters and explorers of these archives, in a peer-to-peer group with no defined leader at the outset. However, the roles were quickly and spontaneously divided, with the participants with more technical expertise guiding the others. Another characteristic is the extension in time of the process, which is not usual for this type of exploration. However, it has the merit of alternating phases of collective sharing and individual experimentation, although there is a risk of demobilization of participants between meetings.

The choice to have a critical approach to the data from the outset is linked to the participants' training, but also to the desire to avoid distant reading bias and to the need to mix datasets. Once the limitations and challenges of the data in terms of volume, size and format had become clear and the stage of exploring possible research problems had been identified, a second round allowed us to develop a first data analysis through distant reading of these datasets and to produce some visualisations (several examples will be further developed in another paper related to this datathon). At this point this is one of our main goals, in order to deepen our study and we will also expand on the first steps by accessing the IIPC Covid-collection content and WARC files through a collaboration with IIPC, Archive-It and the Archives Unleashed Team (Ruest et al., 2021), as a cohort in 2021-22 (Aasman et al., 2021).[12]

Transnational approach as key for web archives analysis

Our experiment demonstrates the possibility of developing shared corpora across European countries, at least based on metadata and derived data. This was feasible through a shared project between web archivists and researchers, and WARCnet provided the perfect frame for this kind of request, and an efficient test bed which is also to be deepened in other
WGs of the WARCnet project. It may provide an insight to researchers looking for collections and even open their minds to new studies when discovering datasets from other countries.

Finally, this type of research needs time and much more effort. Working with complex sources like web archival materials is still in an early stage. Once the limitations and challenges of the data in terms of volume, size and format have become clear and the stage

[12] The Bibliotheca Alexandrina is currently indexing the IIPC COVID collection into Solrwayback, which could also give a new angle to our project.

of exploring possible research problems has been identified, a second round with input from the data providers can be realised, as a typical iterative process that is both valuable and necessary involving close collaboration between web archivists and researchers. It is a fundamental project, or as noted by Richard Dawkins: “The chicken is only an egg’s way for making another egg”!

ACKNOWLEDGEMENTS

We sincerely thank all the libraries, web archivists and web archiving institutions which contributed to this datathon by providing us with datasets and documentation. Our sincere thanks go to the BnL (Luxembourg), the National Széchényi Library (Hungary), the IIPC, the French National Library (BnF), the National Audiovisual Institute (INA), the UK Web Archive, the Danish Royal Library and the KB (Royal Library) in the Netherlands.

REFERENCES

Aasman, S., Brügger, N., de Wild, K., Clavert, F., Gebeil, S. Schafer, V. (2021). Analysing Web Archives of the Covid-19 Crisis through the IIPC collaborative collection: early findings and further research question, IIPC netpreserve.org. https://netpreserveblog.wordpress.com/2021/11/02/analysing-web-archives-of-the-covid-19-crisis-through-the-iipc-collaborative-collection-early-findings-and-further-research-questions/

Ben-David, A. (2016). What does the Web remember of its deleted past? An archival reconstruction of the former Yugoslav Top Level Domain. *New Media & Society*, 18(7), 1103-1119.

Blanckemane, B. (2020). *Baromètre InaStat Septembre 2020. Épidémie de COVID-19 & Collecte Twitter*. Bry-sur-Marne: INA DL Web. https://f-origin.hypotheses.org/wp-content/blogs.dir/3864/files/2020/10/INADlweb-Etude-Twitter-Coronavirus.pdf

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

Di Pretoro, E., Geeraert. E. (2019). Behind the Scenes of Web Archiving: Metadata of Harvested Websites. Archives et Bibliothèques de Belgique - Archief- en Bibliotheekwezen in België, Archief, In P*ress, Trust and Understanding: the value of metadata in a digitally joined-up world*, ed. by R. Depoortere, T. Gheldof, D. Styven and J. Van Der Eycken, 106, 63-74. hal-02124714

Gebeil, S., Schafer, V., Benoist, D., Faye A. & Tanesie, P. (2020). Exploring special web archive collections related to COVID-19: The case of the French National Library (BnF). *WARCnet Paper*. https://cc.au.dk/fileadmin/user_upload/WARCnet/Gebeil_et_al_COVID-19_BnF.pdf

Geeraert, F. & Németh, M. (2020). Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary. *WARCnet Paper*.

https://cc.au.dk/fileadmin/user_upload/WARCnet/Geeraert_et_al_COVID-19_Hungary.pdf

Mons, B. (2018). *Data Stewardship for Open Science. Implementing FAIR Principles*. Boca Rota: CRC Press.

Huurdeman, H.C., Kamps, J., Samar, T., de Vries, A.P., Ben-David, A., & Rogers, R. (2015). Lost but not forgotten: Finding pages on the unarchived web. *International Journal on Digital Libraries*, 1-19.

Maemura, E. (2021). Towards an Infrastructural Description of Archived Web Data. Keynote at the Aarhus Warcnet Conference. https://youtu.be/uUdk76925D8

Milligan, I., Casemajor, N., Fritz, S., Lin, J., Ruest, N., Weber, M. S. & Worby, N. (2019). Building Community and Tools for Analyzing Web Archives through Datathons. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 19.

Milligan, I. (2020). You shouldn't Need to be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure. *WARCnet Paper*. https://cc.au.dk/fileadmin/user_upload/WARCnet/Milligan_You_shouldn_t_Need_to_be__2_.pdf

Moretti, F. (2013). Distant Reading, London/New York: Verso.

Nanni, F. (2017). Reconstructing a website's lost past – Methodological issues concerning the history of www.unibo.it. *Digital Humanities Quarterly*, 11(2).

Rhodes, T. (2013). A Living, Breathing Revolution: How Libraries Can Use "Living Archives" to Support, Engage, and Document Social Movements. Singapour, IFLA WLIC.

Rollason-Cass, S. & Reed, S. (2015). Living Movements, Living Archives: Selecting and Archiving Web Content During Times of Social Unrest. *New Review of Information Networking*, 2 (1-2), 241-247.

Rossenova, L., Wild K. de and Espenschied D. (2019). Provenance for Internet Art: Applying the W3C PROV model. In *Proceedings of the 16th International Conference on Digital Preservation (iPRES)*, Amsterdam, 2019, 297-305. ISBN 9789062590438.

Ruest, N., Fritz, S., Deschamps, R. Lin, J. & Milligan, I. (2021) From archive to analysis: accessing web archives at scale through a cloud-based interface. *International Journal of Digital Humanities*.

Schafer, V., Musiani, F. & Borelli, M. (2016). Negotiating the Wb of the Past. *French Journal for Media Research*, 6. http://www.frenchjournalformediaresearch.com/lodel-1.0/main/index.php?id=952.

Schafer, V., Thièvre, J. & Blanckemane, B. (2020). Exploring special web archives collections related to COVID-19: The case of INA. *WARCnet Paper*. https://cc.au.dk/fileadmin/user_upload/WARCnet/Schafer_et_al_Exploring_special_web_archives.pdf

Schafer V. & Winters J. (2021). The values of web archives. *International Journal of Digital Humanities.*

Schostag, S. (2020). The Danish coronavirus web collection – coronavirus on the curators' minds. *International Internet Preservation Consortium Blog*. https://netpreserveblog.wordpress.com/2020/07/29/the-danish-coronavirus-web-collection/

Truyter, V. (2021). Research Data Management and Sharing Practices of Researchers in Web Archives Studies, Engaging with Web Archives (EWA) Conference, 1 September 2021. https://ewaconference.com/ewa4dh-2021/ewa4dh-programme/

Wilkinson, M., Dumontier, M., Aalbersberg, I., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship, *Sci Data,* 3, 160018.

APPENDIX

This appendix shows the steps needed for cleansing the data and extract the domains.

To illustrate this, the URL 'https://www.acl.lu/en-us/news/voyages-loisirs/voyages-et-transports' includes the domain name (www.acl.lu), the second-level domain name ("en-us") as well as other levels ("news', etc,). For the computer to be able to read this as data, the URL is deconstructed by parsing the data into multiple columns:

> Select all URLs in the dataset (select the column or shortcut "Ctrl+Shift+down arrow key") and copy and paste it in a new sheet tab. Click the *Data tab* in the ribbon, then look in the *Data Tools group* and click [Text to Columns]. In step 1 of the wizard, choose *Delimited* then click [Next]. A delimiter is the symbol or space which separates the data you wish to split, so in this case we would like to delimit the data on "/". Click [Finish].

Some domain names begin with the prefix "www.". This is removed using the "Find and Replace" features in Excel:

> Copy and paste the domain names in a new column and name this "domain_names_without_www". Select the data in this new column and press "Ctrl+H", or go to *Home* > *Editing* > *Find & Select* > *Replace*. In the *Find what:* box, type "www.". Enter nothing in the *Replace with:* box and press [replace all].

The IF function is used to only keep the second-level domain names of certain websites, which makes it possible to only keep the second-level domain names of social media websites:

```
=IF(COUNTIF(Lookup!A:A;D2);K2;"")
```

> *(Within this example, the domain name is in column D and the second-level domain name in column K. There is a worksheet "Lookup", where column A includes a list of websites of which the second-level domain name is relevant, e.g. twitter.com, facebook.com, tiktok.com.)*

To remove duplicate domain names from URLs, it is possible to use dictionary keys and values in Python (key will be domain name, and value will be overwritten if the key already exists). Yet, since this file includes a column “Domain_names", it is also possible to cleanse it within Excel itself.

> The domain names were copied in a new column “unique_domain_names”. The following code was entered in the developer (*Developer* > *Visual Basic* or shortcut “Alt+F11”):

```
Sub RemoveDuplicates()
'UpdatebyExtendoffice20160918

  Dim xRow As Long
  Dim xCol As Long
  Dim xrg As Range
  Dim xl As Long
  On Error Resume Next
  Set xrg = Application.InputBox("Select a range:", "Kutools for Excel", _
                    ActiveWindow.RangeSelection.AddressLocal, , , ,
, 8)

  xRow = xrg.Rows.Count + xrg.Row - 1
  xCol = xrg.Column
  'MsgBox xRow & ":" & xCol
  Application.ScreenUpdating = False
  For xl = xRow To 2 Step -1
    If Cells(xl, xCol) = Cells(xl - 1, xCol) Then
      Cells(xl, xCol) = ""
    End If
  Next xl
  Application.ScreenUpdating = True

End Sub
```

Another essential element in the URL is the top-level domain name (e.g. “.com”). The following formula is used to extract the top-level domain from the domain names:

> =RIGHT(C2;LEN(C2)-SEARCH("$";SUBSTITUTE(C2;".";"$";LEN(C2)-LEN(SUBSTITUTE(C2;".";"")))))
>
> *(Within this example the domain names are in column C.)*

This formula contains several steps, the first of which is to find the number of periods within the URL (LEN(B2)-LEN(SUBSTITUTE(B2;".";""). As there are often multiple periods in a URL, it then tries to substitute the last period with a character that is not often found within an URL, in this example “$” (SUBSTITUTE(B2;".";"$"). Now the computer is able to find this position (SEARCH("$"). It is now possible to use the RIGHT() function to extract the characters before the "$", in other words the top-level domain. To add the period again to the top-level domain:

=CONCAT(".";F2)

(In this example, the top-level domains without a period are in column F. While it is also possible to add this to the RIGHT() formula above, in our data set the top-level domain names with a period were included in a new column.)

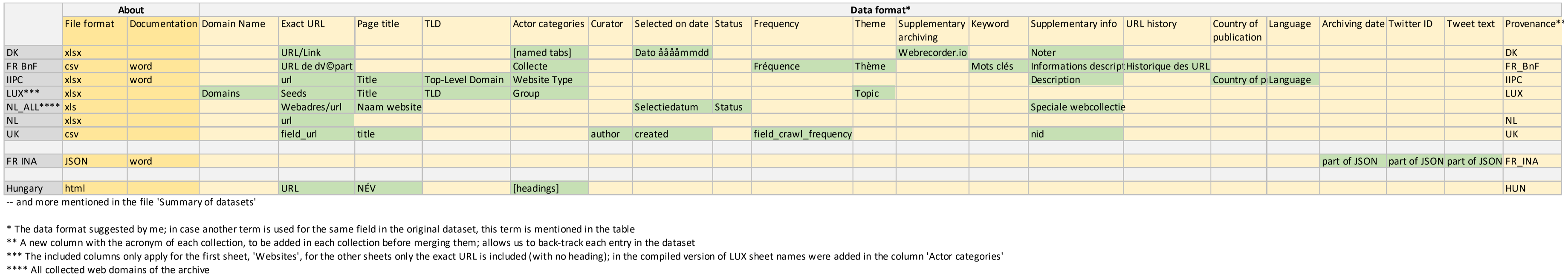
Top-level domains can convey information about the intended use of the website. IANA (Internet Assigned Numbers Authority), who manages and approves new top-level domains, distinguishes various groups, like generic top-level domains (gTLD), historically the generic domain names that are now sponsored by designated organizations, or country code top-level domains (ccTLD), generally used or reserved for a specific country. To further expand the dataset, extra data is scraped from Wikipedia and added to each URL, including the geographical locations:

The scraped data was pasted into a new sheet tab named “Lookup”.[13] Unintended whitespaces before a data element were cleaned up using the formula:
=SUBSTITUTE($V4;" ";"";1)

In the worksheet that includes the top-level domains, a new column “countries” was created and the following formula included:
=IF(INDEX(Lookup!Y:Y; MATCH($G2;Lookup!U:U;0))=0;"";INDEX(Lookup!Y:Y; MATCH($G2; Lookup!U:U;0)))

(In this example, the top-level domains with a period are in column G. The scraped data in the “Lookup” tab sheet also contains a top-level domain in column U and the country can be found in column Y.)

ANNEX



	About		Data format*																			
	File format	Documentation	Domain Name	Exact URL	Page title	TLD	Actor categories	Curator	Selected on date	Status	Frequency	Theme	Supplementary archiving	Keyword	Supplementary info	URL history	Country of publication	Language	Archiving date	Twitter ID	Tweet text	Provenance**
DK	xlsx			URL/Link			[named tabs]		Dato ååååmmdd				Webrecorder.io		Noter							DK
FR BnF	csv	word		URL de dV©part			Collecte				Fréquence	Thème		Mots clés	Informations descript	Historique des URL						FR_BnF
IIPC	xlsx	word		url	Title	Top-Level Domain	Website Type								Description		Country of p	Language				IIPC
LUX***	xlsx		Domains	Seeds	Title	TLD	Group					Topic										LUX
NL_ALL****	xls			Webadres/url	Naam website				Selectiedatum	Status					Speciale webcollectie							
NL	xlsx			url																		NL
UK	csv			field_url	title			author	created		field_crawl_frequency				nid							UK
FR INA	JSON	word																	part of JSON	part of JSON	part of JSON	FR_INA
Hungary	html			URL	NÉV		[headings]															HUN

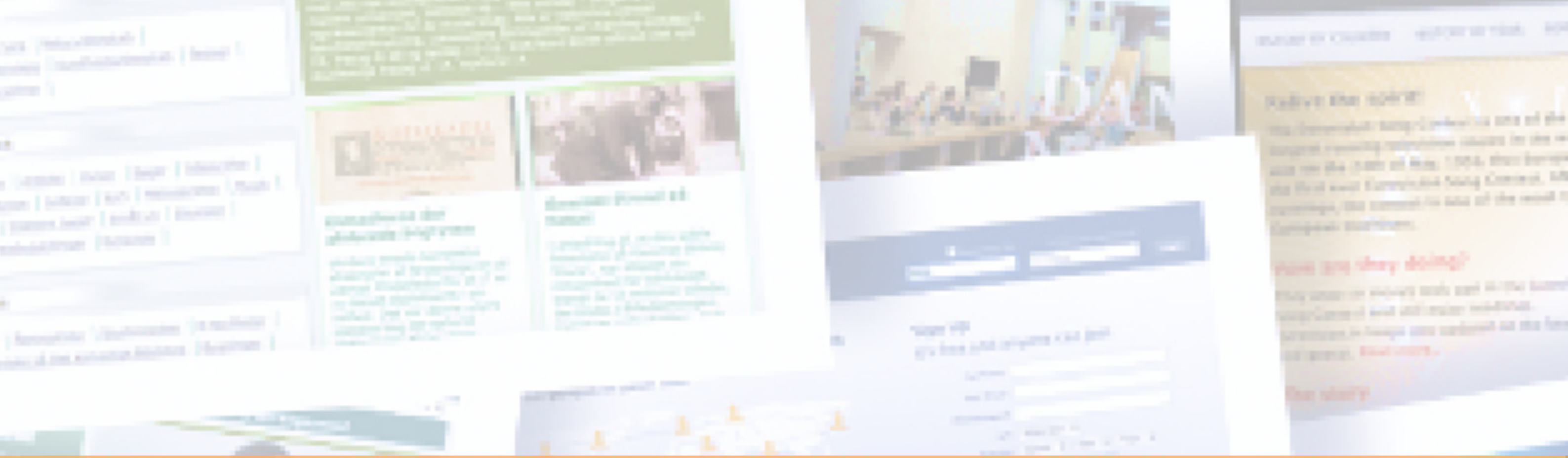
-- and more mentioned in the file 'Summary of datasets'

* The data format suggested by me; in case another term is used for the same field in the original dataset, this term is mentioned in the table

** A new column with the acronym of each collection, to be added in each collection before merging them; allows us to back-track each entry in the dataset

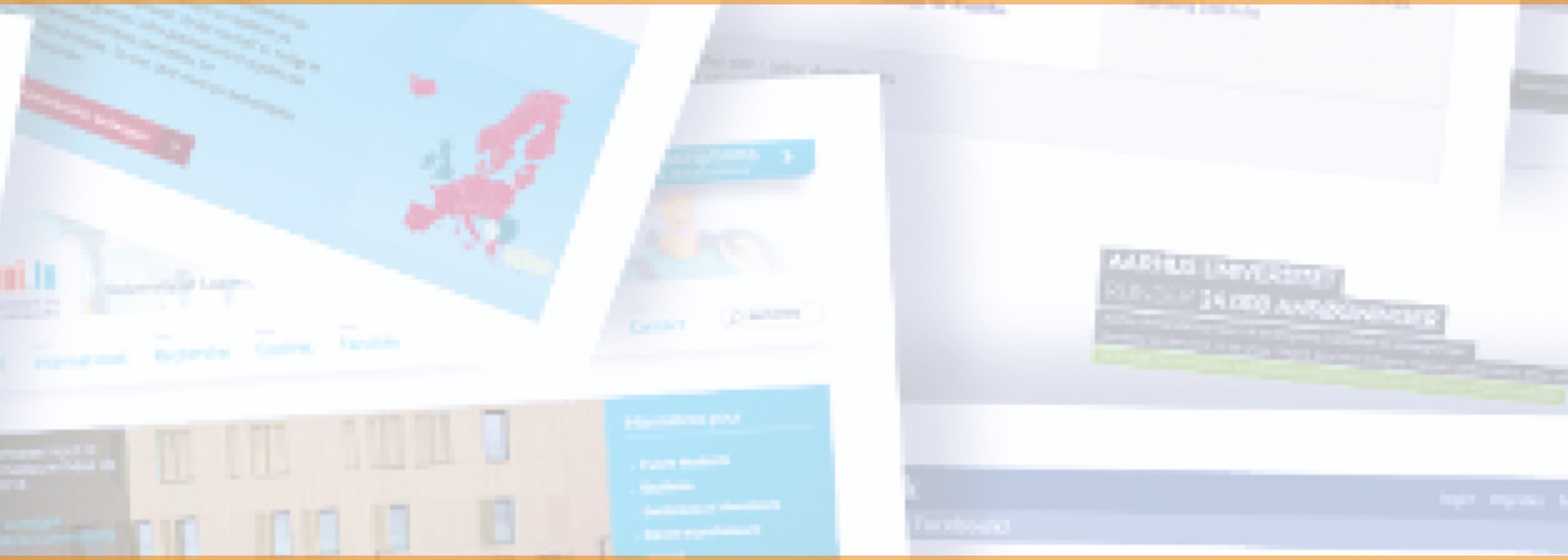
*** The included columns only apply for the first sheet, 'Websites', for the other sheets only the exact URL is included (with no heading); in the compiled version of LUX sheet names were added in the column 'Actor categories'

**** All collected web domains of the archive



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