Towards an Infrastructural Description of Archived Web Data

Emily Maemura

WARCNET PAPERS

WARCnet
web archive studies

Towards an Infrastructural Description of Archived Web Data

Emily Maemura

emaemura@illinois.edu

WARCnet Papers
Aarhus, Denmark 2022

WARCnet Papers ISSN 2597-0615.

Emily Maemura: *Towards an Infrastructural Description of Archived Web Data*

Published by the research network
WARCnet, Aarhus, 2022.

Cover design: Julie Brøndum
ISBN: 978-87-94108-08-9

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).

This WARCnet Paper has gone through a process of single blind review.

DANMARKS FRIE
FORSKNINGSFOND
INDEPENDENT RESEARCH
FUND DENMARK

WARCnet Papers

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Towards an Infrastructural Description of Archived Web Data

Emily Maemura

Abstract: Recent efforts in the web archiving community have focused on supporting research engagement through building new tools and interfaces in order to access collections. As focus shifts from browsing archived sites to the 'big data' analysis of WARC files, I argue that new frameworks are needed to understand the systems and processes underlying these datasets. This paper explores how to describe and characterise archived web data by contextualising collections within the sociotechnical systems that generate that data. Different modes of description are compared, revealing how the characterisation of data through computational means can be supplemented with infrastructural methods for describing the histories, organisational policies, relationships, and other catalysing events that shape web archives. Future directions are explored, outlining how tools and services might be configured to support a more holistic description of collections.

Keywords: Web archives, research methods, infrastructure studies

INTRODUCTION

As the creation and use of web archives has grown over the past decades, increasing attention, effort, and resources have been devoted to the development of tools and facilities for analysing these collections. This paper considers and compares emerging research infrastructures supporting web archives analysis, studying sites in Denmark and in Canada during active development and testing to support scholarly use of archived web data. Rather than evaluating the development from the practical perspective of user testing, I take the perspective of infrastructure studies in order to consider how the web archives research infrastructures in these settings include both computing facilities and data collections, but also encompass the organisational systems and people involved, including librarians, curators, and technical developers. While description of web archives collections has primarily focused on characterising datasets, I explore how research analysis must take into account forms of documentation and description of the practices and processes that relate data to specific organisational arrangements and the systems that produce and sustain them. Particularly, I argue that new frameworks (extending from the field of infrastructure studies) are needed to inform description of these collections beyond characterising datasets, to also include the context of data curation systems, processes, procedures, and organisational settings within which collecting takes place.

ESTABLISHING FACILITIES AND FRAMEWORKS FOR ANALYSING COLLECTIONS

Over the past decade, major advancements have been made in developing tools and interfaces to support web archives research. This includes significant improvements to search and access interfaces, allowing research users to gain new insights into the data found within collections. However, as I outline and argue here, more robust frameworks are needed in order to support research engagement with web archives by contextualising and describing not only the data artefacts generated, but also the systems and practices that shape collections.

As described in a previous literature review (Maemura, 2018), several key reports were published from 2010 to 2014 on the subject of web archives and research engagement, calling for the field to invest in search and analysis interfaces, supporting analysis at the scale of 'big data.' On the one hand, these initial reports addressed a practical and technical need to support research with large collections, since institutions like national libraries have now established web archiving programs for many years or decades, and their web archives collections have grown to very large volumes of data (i.e. hundreds of Terabytes). Additionally, it has been noted that supporting research engagement at this scale necessitates new methodologies, moving away from previous approaches focused on browsing and rendering individual web pages and web resources. This shift from 'document-centric' to 'whole-collection' approaches has been described previously by Hockx-Yu (2014) and Ben-David & Huurdeman (2014). This kind of 'paradigm shift' in methods is at the core of recent collaborations pairing researchers with developers and curators as many libraries establish facilities for high-performance computing and collection indexing and analysis.

For example, the past decade has seen several collaborations within the web archives community focused around these whole-collection approaches. From a technical perspective, many efforts have been devoted to the development of tools for processing and presenting archived web data at scale. Examples include powerful search and indexing tools like the UK Web Archive's Shine search and Solr index (https://www.webarchive.org.uk/shine), as well as the Archives Unleashed Toolkit, which can generate a set of standard research-centric derivatives (https://cloud.archivesunleashed.org/derivatives). Additionally, the standard WebARChive (WARC) file format has become central to the set of tools, interfaces, and high-performance computing facilities emerging for this kind of computational humanities work with web archives collections. Beyond these technical developments, other challenges of research engagement have also been discussed, as curators and researchers have encountered various legal or policy barriers when working with web archives collections, as well as ethical aspects (Vlassenroot et al., 2019; Bingham & Byrne, 2021). For example, the access policies for national web archives collections often mirror the closed ‘reading room’ access policies that had been developed for counterpart print collections, and digital collections like web archives are beholden to similar copyright restrictions. Despite these challenges, collecting institutions have made significant developments in establishing facilities to meet the needs of researchers in whole-collection analysis.

With these recent efforts in mind, I observe that the modes of analysis provided by these facilities centre on specific data artefacts and the data fields found in WARC files. For instance, in their analysis of the Danish Royal Library’s Netarchive collection, Laursen & Møldrup-Dalum (2015) present several key visualisations comparing changes in the collection over time, showing a count of top level domains, the relative amount of data by media type (e.g. audio, image, text, video and 'other' formats), as well as HTTP response codes (e.g. 200 OK, 301 Moved Permanently, 404 Not Found). This example illustrates some of the different analyses that can be performed with big data approaches, with an emphasis on characterising the ‘raw data’ underlying archived web pages. Description of datasets computationally performed in this manner relies heavily on the data fields available in WARC files and the related standardised data formats of HTTP transactions, URLs, and HTML webpages. While a focus on WARC files in these kinds of analysis is useful in promoting standardisation and interoperability, the resulting characterisations present a limited view of archived web data. When studying collections, researchers must also ask what elements are missing when description of web archives collections is being increasingly delegated and deferred to computational tools and systems

Recently, researchers have been addressing the limits of these computational methods and tools, which are now being actively probed and questioned by many in the web archives community. It is becoming more and more clear that, in addition to facilities for processing ‘big data,’ researchers require additional information about the processes used for collecting and curating, seeking out ways to describe and characterise web archives beyond what is available in computational interfaces. For example, in studying the North Korean web domain as captured by the Internet Archive, Ben-David and Amram (2018) address the Internet Archive’s crawler as a ‘black box’ whose specific logics and rules are largely unknown to users. Milligan (2019) similarly describes the challenges and unknowns of using web archives as a historian, acknowledging that the scale of collections is not amenable to detailed description but calling for access to certain key components such as selection algorithms or “the top-level decision-making or objectives that may have led to a given site being collected can help us at least write informed histories” (p. 86). Brügger (2018) additionally advocates that specific documentation be made available to researchers, including key metadata generated during initial collecting or subsequent processing with a focus on information to aid “web historians who are trying to establish the provenance of what is in a web archive, what should have been archived, why it did end up there, and when” (p. 141). Providing access to datasets and computational analysis tools is no longer sufficient; in order for researchers to perform rigorous analyses of archived web data, they require description and contextualisation of datasets that address the selection and curation decisions applied at various stages of collecting and processing.

In summary, while methods and facilities for analysis have shifted from document-centric to whole-collection approaches, the concepts and ideas informing the description of a web archives collection have not yet caught up with this shift. Overall, practical development of facilities for analysis centre description of a collection on computational characterisation of standard data formats. However, for researchers, analysing these collections demands an understanding of the processes and transformations enacted upon that data by various people and systems. My primary interest is in establishing an approach that outlines how to describe, contextualise, and characterise data in relation to web archives research infrastructures, and accounting for the mediation and intervention of various organisations, people, and systems in the archiving process.

Several potential frameworks are presented in past work, which may be considered for the description of collections, contextualising them more broadly beyond computational analysis of data artefacts. For example, from the perspective of collecting institutions, the Open Archival Information System (OAIS) reference model provides a framework outlining the people and systems involved in web archiving. Sierman and Teszelszky (2017) apply these existing OAIS concepts to describe the KB-NL collection and its selection processes, focusing on the context information necessary for the ‘designated community’ of the archive’s future users. While the OAIS functional model of processes and informational model both note the need for recording context information for the archive’s designated community of users, overall the OAIS model primarily addresses what information is needed but not how to enact this work of description. An alternative approach is presented by Maemura et al. (2018), which outlines several dimensions of the sociotechnical context in which decisions are made.

In my study presented here I develop a new approach which focuses on empirical settings and the nuances of what actually occurs in collecting practice by drawing connections to infrastructure studies, which provides key concepts and frameworks for relating data to sociotechnical systems. In this literature, ‘infrastructures’ comprise complex assemblages of people and technologies that depend on site-specific contingencies as well as evolving standards and practices – for example, a historical study of the Museum of Vertebrate Zoology at Berkeley addresses the collection’s underlying information infrastructure that relies on the use of standardised paper form to record details of specimen collection, supporting cooperative work among amateur collectors, farmers, trappers and traders throughout California, ultimately enacting the vision of the museum’s director and founder (Star & Griesemer, 1989). Analysing infrastructures through this lens, rather than focusing on technical facilities, draws attention to the human dimensions that bind together work performed by multiple interconnected systems. Though the technological modes and means of web archives collections differ greatly from a natural history museum in the early 20th Century, I adopt this framework to similarly draw attention to the practices, standards, and social relationships that shape web archives and their data. I find the lens of infrastructure studies particularly useful to describe and contextualise collections within their specific social, historical, political, and cultural settings.

In particular, as I argue and demonstrate here, these perspectives from infrastructure studies are useful to highlight data's relationships to the collecting and management systems used for archiving and supporting research applications. I apply this perspective to emphasise how these infrastructures generate, sustain, and support the use of tools and data, through often-invisible decisions and unstructured documentation, which are vital to understanding and reading meaning into these collections. As I describe in more detail below, this approach is particularly useful to consider how web archives are shaped both with and both with and beyond computational or algorithmic logics.

ADOPTING AN INFRASTRUCTURAL APPROACH

In order to study the sociotechnical contexts that shape web archives data, my work adopts concepts and approaches from the field of infrastructure studies. This approach begins with a key distinction, focusing less on 'research infrastructure' as computational facilities and more on 'research infrastructure' as an assemblage of people, organisational practices, resources and contingencies. While everyday terminology stresses infrastructure as a technical substrate (i.e. roads, electrical grids, water) or essential service,[1] I build on the analysis of information infrastructures as developed by Bowker and Star (2000) that focuses on identifying the key relationships that often reside in the background and bringing them to the fore. Their conception of infrastructure is centred on any arrangement of people, standardised methods or practices, and commonly developed concepts or artefacts that support distributed work towards a shared goal (Star & Griesemer, 1989; Star, 2010). With the focus on a common goal of knowledge production, I also consider this 'research infrastructure' of web archives to span settings of collecting and archiving, and scholarly research and analysis.

While not strictly defined in the literature, an 'infrastructure' as an object of study has been characterised by several qualities or dimensions such as embeddedness, transparency, 'reach or scope,' and 'embodiment of standards' seen in the relationships and arrangements of people and practices (Star & Ruhleder, 1996; Bowker & Star, 2000). Additional detailed description of an information infrastructure is provided through a prominent example from Bowker and Star (2000) which centres on the healthcare setting and addresses the development and evolution of the International Classification of Diseases (ICD). While the ICD was developed as a classification system to standardise assembling mortality and morbidity data worldwide, the authors note how cultural context impacts what is ultimately recorded on a death certificate. They illustrate this with an example of clerks entering data into a medical database who may differently interpret and categorise abortion: where User A considers it a crime, User B views it as a routine

[1] Sandvig (2013) provides a useful overview of varying definitions and the evolution of infrastructure studies more broadly.

medical procedure, and these differences impact User C's subsequent compiling of statistics from that database. In this instance, the infrastructure not only comprises elements such as design of the database and use of a standardised classification system, but also the legal and cultural context informing User A and User B's decisions in categorising are equally important as part of this infrastructure. I translate Bowker and Star's approaches with a broad consideration of context that informs how data artefacts are constructed, examining when or why each value is entered in a given data field, and extending these ideas to the ways data are recorded and interpreted from web archives.

Adopting this approach from infrastructure studies requires recognizing specific and situated sociotechnical relationships in any given setting, and exploring how decisions are made within the particular contingencies and constraints of a given context (Star & Ruhleder, 1996). Methodologically, empirical work is essential in order to explore contingencies in different settings, and to reveal how generic solutions cannot be applied universally. Star (1999) describes the principles for conducting an 'ethnography of infrastructure' and focuses on sites where particular activities can be observed: decision-making, processes of standardisation, or tinkering and tailoring to develop solutions. Not limited to traditional forms of ethnographic fieldwork from anthropology, methods for studying infrastructure also include historical analysis, interviews, observations, and systems analysis.

Embracing this importance of empirical investigation for infrastructure studies, my work centres on fieldwork in two key settings where research infrastructures are being developed for collecting and analysing archived web data.[2] The first phase of fieldwork took place in Denmark in early 2018 where I studied the Royal Danish Library's work, and their longstanding national legal deposit web collection, the Netarchive (http://www.netarkivet.dk). The collection has captured the Danish web domain since 2005, and I focused in particular on the Royal Library's recent partnership with researchers from Aarhus University's NetLab to collaborate on a pilot project for testing the National Cultural Heritage Cluster high-performance computing infrastructure with their project "Probing a Nation's Web Domain" (Brügger et al., 2019, 2020). The second phase of fieldwork took place in Canada during the latter half of 2018 where I studied the research infrastructure being developed by the Archives Unleashed project. With Archives Unleashed, my observations focused on one of the 'datathons' facilitated by their team and hosted by Simon Fraser University in Vancouver. These datathons are two-day hackathon-style events where local and international participants collaborate in small ad-hoc teams of researchers, librarians and developers (as has been previously described in Milligan et al., 2019). In that setting, I subsequently conducted interviews with librarians who created the collections that were analysed at the datathon, including Archive-It event- and subject-based collections from the University of British Columbia (UBC) and University of Victoria libraries (https://archive-it.org/organizations/734 and https://archive-it.org/home/uvic).

[2] I am grateful for the funding support provided by the SSHRC Canada Graduate Scholarship and Michael Smith Foreign Study Supplement, as well as the University of Toronto School of Graduate Studies Research Travel Grant that allowed me to conduct this ethnographic fieldwork.

In terms of the research design, I position myself alongside this work and these projects, studying the challenges that arise as data are transferred between systems, to new social and technical contexts, in order to understand the selection and curation decisions made by different people and tools. In addition to interviews and participant observation at these sites, I extend the 'ethnography of infrastructure' to include document analysis of project plans, work plans, and other organisational artefacts, as well as performing systems analysis and modelling to study the entities and data structures employed by various systems.

My goal is to explore and reveal how specific choices are made for web archives collections, and how decisions are made within the particular contingencies and constraints of a given context. With this perspective, I analyse and consider data from web archives collections, looking beyond WARC files as the central components or artefacts of concern, and beyond the automated logics of the web crawler. By studying the supporting infrastructure of web archives, I ask how we might better contextualise WARC data by describing and documenting these decisions made at various points in archival processing. I explore where and how decisions are made with algorithms and computational logics, and which decisions are located outside of these logics through other kinds of interventions in sociotechnical systems made to curate, aggregate, select, extract, manipulate, or transform data. I adopt the lens of infrastructure studies in order to emphasise the people involved in data curation, and I argue that the invisible work that is excluded from formally standardised data representations can only be studied through 'an ethnography of infrastructure' (Star, 1999). I believe this account of collecting is a vital and necessary complement to the statistical or computational readings of data provided by current tools with more complex, complicated, and nuanced understandings of sociotechnical factors that shape these collections.

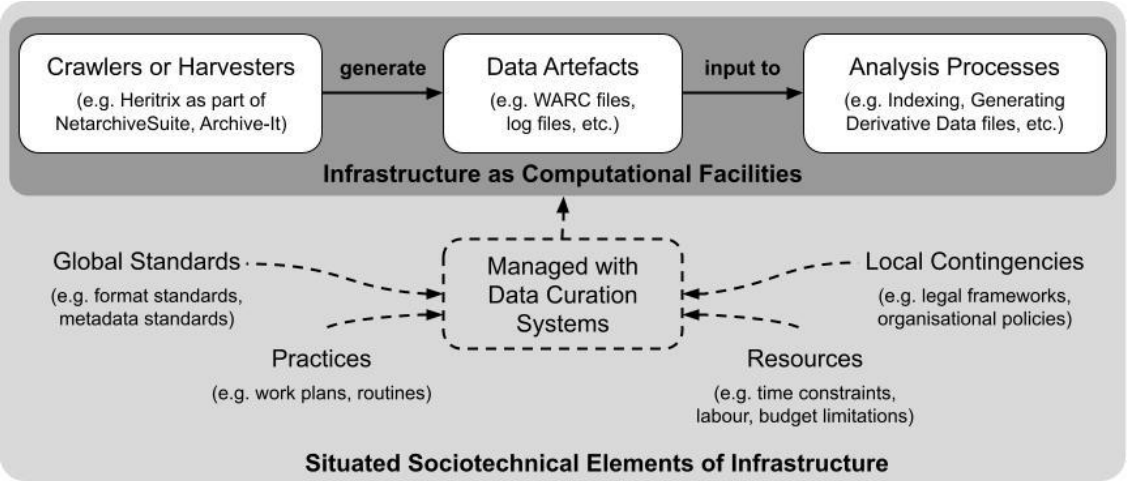
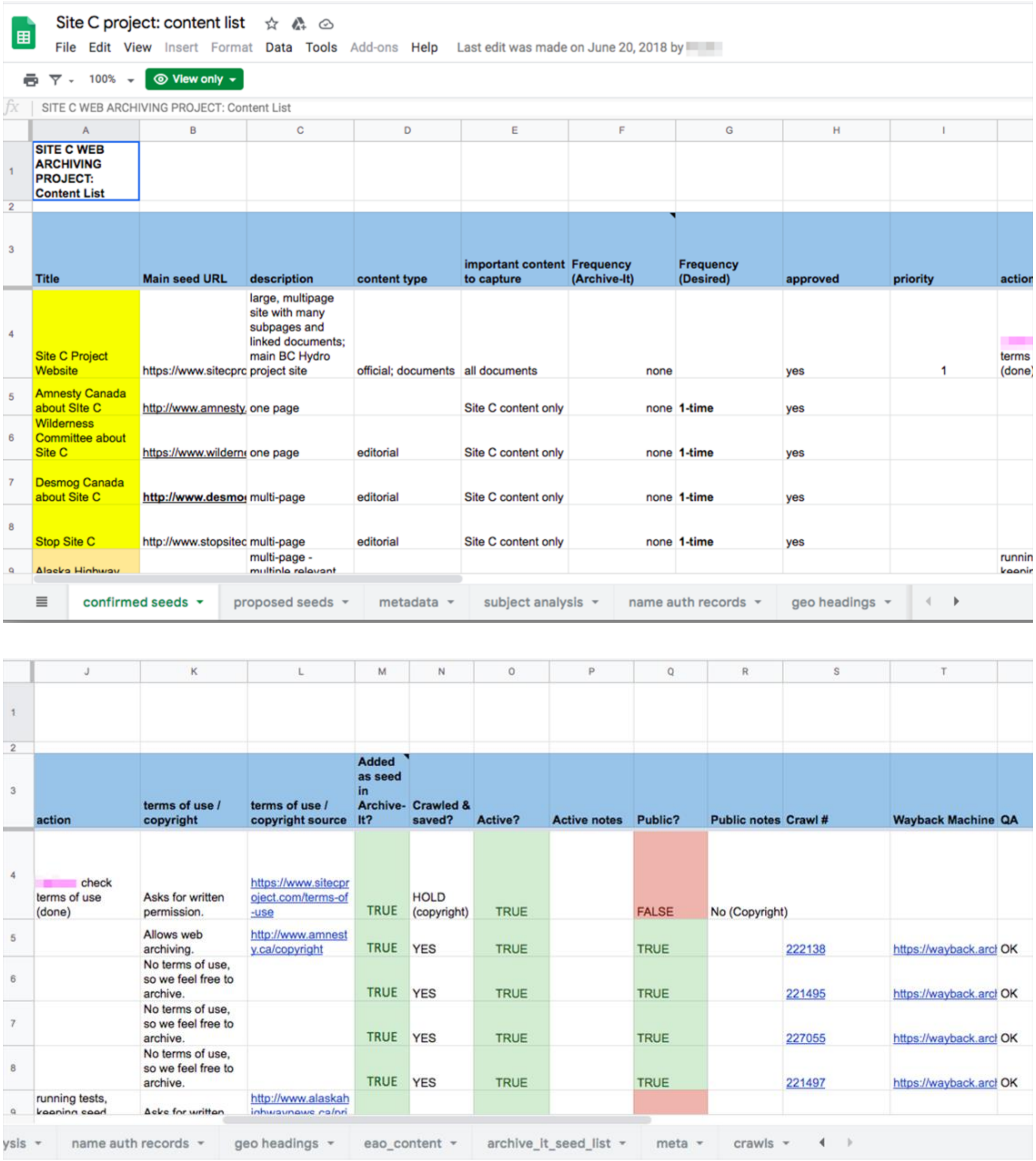

Crawlers or Harvesters
(e.g. Heritrix as part of NetarchiveSuite, Archive-It)
generate
Data Artefacts
(e.g. WARC files, log files, etc.)
input to
Analysis Processes
(e.g. Indexing, Generating Derivative Data files, etc.)
Infrastructure as Computational Facilities
Global Standards
(e.g. format standards, metadata standards)
Managed with Data Curation Systems
Local Contingencies
(e.g. legal frameworks, organisational policies)
Practices
(e.g. work plans, routines)
Resources
(e.g. time constraints, labour, budget limitations)
Situated Sociotechnical Elements of Infrastructure

Figure 1: Diagram of technical, procedural view of 'Infrastructure as Facilities' focused on data artefacts and the broader 'Situated Sociotechnical Elements of Infrastructure' characterising the context of collecting and analysis.

EXAMINING SITUATED DATA CURATION SYSTEMS AND THEIR CATEGORIES

While the study of infrastructure addresses context broadly, I focus my discussion here on one key aspect: the importance and prominence of data curation systems that support collecting practices. The use and ubiquity of these systems is a central finding from my ethnographic study and fieldwork, where I discovered that curators and librarians have developed their own site-specific systems used for data curation. In studying these different collections, I observed that while collecting processes vary between settings, each has developed their own systems to manage collection development, and these systems exist separately from harvester or crawler software. Though they can take different forms such as the spreadsheets and step-wise processes described below, these data curation systems are used in the process of collecting to manage, mediate, and intervene in the crawler's automated work. These systems are also designed to fit within the needs and contingencies of their specific context, and are necessary because crawlers don't always run smoothly, automatically, or as expected, and to manage the overall amount and quality of data being ingested. In my analysis here, I focus on describing an example of a data curation system used in each setting, in order to reveal: first, how a collection is managed at each site, and second, how these systems are designed to fit within the needs and contingencies of a specific context.

One example of a data curation system comes from the University of British Columbia Libraries, whose web archives curators use a Google Sheets spreadsheet template as a central part of the process of collection development. During my fieldwork, one librarian showed me the completed spreadsheet that had been created for the collection around the Site C project, that has captured web pages around the proposed hydro dam in Northern British Columbia, and the related controversy and response from local environmental activists and First Nations communities (this collection is accessible on the public Archive-It site: https://archive-it.org/collections/7588). The overall collection development spreadsheet is composed of several different individual sheets, each used at different times within the collecting workflow. For example in later stages, the spreadsheet is used to manage subject headings, name authority records, and geographic metadata for each seed. In terms of data practices, this shared spreadsheet document is useful to coordinate multiple processes in this one place, including processes such as: documenting seed URLs, adding notes to perform Quality Assurance checks, annotating and adding metadata, and assigning tasks to different members of the library team. Overall, this spreadsheet is central to the collecting process, used to structure and record curation decisions and judgments.



Site C project: content list

Last edit was made on June 20, 2018 by [redacted]

SITE C WEB ARCHIVING PROJECT: Content List

	A	B	C	D	E	F	G	H	I	
1	SITE C WEB ARCHIVING PROJECT: Content List									
2										
3	Title	Main seed URL	description	content type	important content to capture	Frequency (Archive-It)	Frequency (Desired)	approved	priority	action
4	Site C Project Website	https://www.sitecprc	large, multipage site with many subpages and linked documents; main BC Hydro project site	official; documents	all documents	none		yes	1	[redacted] terms (done)
5	Amnesty Canada about SIte C	http://www.amnesty.	one page		Site C content only	none	1-time	yes		
6	Wilderness Committee about Site C	https://www.wildern	one page	editorial	Site C content only	none	1-time	yes		
7	Desmog Canada about Site C	http://www.desmo	multi-page	editorial	Site C content only	none	1-time	yes		
8	Stop Site C	http://www.stopsitec	multi-page	editorial	Site C content only	none	1-time	yes		
9	Alaska Highway		multi-page - multiple relevant							runnin keepir

	J	K	L	M	N	O	P	Q	R	S	T	
1												
2												
3	action	terms of use / copyright	terms of use / copyright source	Added as seed in Archive-It?	Crawled & saved?	Active?	Active notes	Public?	Public notes	Crawl #	Wayback Machine	QA
4	[redacted] check terms of use (done)	Asks for written permission.	https://www.sitecproject.com/terms-of-use	TRUE	HOLD (copyright)	TRUE		FALSE	No (Copyright)			
5		Allows web archiving.	http://www.amnesty.ca/copyright	TRUE	YES	TRUE		TRUE		222138	https://wayback.arcl	OK
6		No terms of use, so we feel free to archive.		TRUE	YES	TRUE		TRUE		221495	https://wayback.arcl	OK
7		No terms of use, so we feel free to archive.		TRUE	YES	TRUE		TRUE		227055	https://wayback.arcl	OK
8		No terms of use, so we feel free to archive.		TRUE	YES	TRUE		TRUE		221497	https://wayback.arcl	OK
9	running tests, keeping seed	Asks for written	http://www.alaskahighwaynews.ca/pri									

Figure 2: Excerpt from UBC's Google Sheet for managing the Archive-It "B.C. Hydro's Site C Project" collection.

Looking more closely at the spreadsheet and how it structures practices and decisions, an excerpt is shown in Figure 2, representing the primary sheet recording "confirmed seeds" which lists the seed URLs that have been input to Archive-It to be crawled. UBC's curators record a wide range of information by using the sheet's 22 different columns. For example, these spreadsheet fields are used to group and classify web materials, as seen in the addition of descriptive categories defining a seed's "content type" in a field that distinguishes editorial articles or official government documents. To aid the iterative process of crawling and quality assurance used with Archive-It, curators also use specific spreadsheet fields to note "important content to capture" and "priority" highlighting areas where curators might take more time and care in QA reviews. Copyright is a central concern for the library, reflected in spreadsheet fields that provide a summary of the terms of use for each seed URL, and link to the page listing copyright restrictions where noted for a given website.

The spreadsheet is not only used to add or annotate information about a seed URL, but also provides a space for curators to negotiate categories and mediate pre-determined parameters defined for crawling with Archive-It. For instance, two columns help manage crawl frequency, with one listing the options available in Archive-It (e.g. daily, monthly, quarterly), and another for desired frequency. Curators can use this 'desired frequency' field to note their own level of crawl frequencies that aren't available within Archive-It's standard settings, such as 'every two weeks.' Additionally, other columns are used to check and interface with the Archive-It system, and its standard set of inputs and outputs. One column titled 'Added as seed in Archive-it?' performs an automatic check against another sheet within this spreadsheet that contains the direct output 'seed report' file from Archive-It by using a VLOOKUP formula. The graduate student who created this spreadsheet template added a text comment for this column to describe the formula, and to explain how future curators must perform additional checks for formatting errors with URLs in the case that any seed returns a 'FALSE' value. Overall, the affordances of the spreadsheet are useful since curators can structure and manipulate formulas in the spreadsheet to suit their needs, in a way that is more malleable than the rigid rules and code of the crawler and harvester.

What I want to highlight with this one example from UBC is how it both embodies and documents specific contextual factors and site-specific contingencies for that setting's sociotechnical infrastructure. The spreadsheet is designed to track and manage curation decisions, and it's also built in response to a certain set of conditions, characterising that collection's context. In terms of technical systems, it's designed around Archive-It and its data structures and entities. Curators are also concerned with legal constraints, particularly copyright issues and fair dealing which restricts what university libraries in Canada can or can't collect. Organisational factors are managed through this sheet, including assigning tasks to certain people, within given work plans and available human resources, all of which only works since most of these collections have fewer than 100 seeds. There's also a sociotechnical aspect of managing data budgets which are associated with real costs to the library for the archive-it subscription, leading librarians to closely monitor data volumes used for each web archives collection, and to set priorities for which seeds or sites should get more attention, or are most important to capture in full. Even the form that this document takes as a Google Sheet is itself reflective of the library's organisational context—curators had formerly stored similar collection development information on their intranet, but since they rely heavily on graduate student workers in their web archives collection development, they needed to find something more flexible for sharing and contribution (since graduate students don't have the full-fledged IT credentials to sign into the Intranet). The spreadsheet's role and value is multiple: it serves as a tool supporting certain collecting practices, a record of where, when and why decisions are made for a certain seed, and also provides insights into the sociotechnical context for UBC in which certain categories and classifications for seed URLs have been developed.

A second example of a curation system comes from the Danish Netarchive, whose curators and crawl engineers use their own system for managing and curating large-scale data from their quarterly Broad Harvests. These Broad Harvests are one of three core collecting strategies, and include materials from over a million domains or seed URLs, where one harvest can include over 35TB of data (Laursen & Møldrup-Dalum, 2017). To manage this scale of collecting, curators rely on some built-in features of the Netarchive Suite harvesting software, and its design that uses a two-step process for large-scale crawls. In Step 1, a small data limit (in the range of 10 MB) is applied to all domains, which are crawled over several days. In Step 2, Netarchive Suite will use the stored results from Step 1, and re-crawl only those domains that exceeded that initial small data limit. Approximately 80,000 of these larger domains will be captured with a higher data limit set in Step 2, and it can take several weeks to complete. While also supported by the use of tabular data and an Excel spreadsheet, the data curation system developed for the Netarchive collection is focused on this consistent set of steps followed for each Broad Harvest.

For instance, in speaking with curators, I learned that determining the Step 2 limit for each domain is not an entirely automatic process and requires review by curators who aim for a balance of collecting complete sites, but not capturing unwanted data. These limits are set with the aid of a CSV report exported from Netarchive Suite and explored and analysed using Excel. One curator specifically manages these per-domain data limits, and he described the process as follows: first, he imports the CSV report into Excel, resulting in a very large file of over 12000 rows. He then looks through the information from each domain, and assigns individual limits, anywhere from 1GB to 12GB (at 2GB increments). He bases these decisions on rules of thumb, judging whether a domain’s limits should be raised based on the number of pages or resources captured for that domain, or the average file size. He also includes special limits ending in “99” for any domains that he knows provide ‘bad data’ i.e. full of crawler traps, shopping sites, calendars. For these sites, instead of noting a limit as “4GB” he would instead set a limit as “3999MB” and that would be a signal in the future not to change that limit and that these domains do not need to be reviewed again. While the spreadsheet is used to sort and filter domains, the curators makes these judgments to raise limits based on his own experience and expertise.

Similar to the UBC example, this process used to manage crawling for Broad Harvests reflects the specific sociotechnical context of the Netarchive. The procedure for setting domain-specific data limits uses a report output from the Netarchive Suite software in conjunction with sorting and filtering in Excel in order to track and manage data captured during these the large scale crawls. This procedure is only possible in the Netarchive's context, and was in a sense developed in response to their specific set of conditions. For instance, the Royal Library's dedicated IT staff affords them technical capabilities to develop and maintain the custom-designed Netarchive Suite software. Broad Harvests of the entire Danish web domain are possible within the library's legal context, and their mandate through the legal deposit law to do this large-scale collecting. Additionally, in terms of organisational culture, the same people or staff members have worked on these harvests over the years, developing their own expertise to assess site limits at scale by filtering, using metrics and aggregate views of data with the simple and malleable tools in Excel. The relative stability of the Netarchive's staff over time also allows for the implementation of internally applied codes like the “99” numbers for setting limits. Again, this data curation system illustrates the way that various sociotechnical factors can converge to determine where, when and why site-specific categories for archived web data are developed.

As these brief examples illustrate, comparing these systems used for data curation reveals how each is closely interrelated to its setting. They additionally reveal the wealth of contextual information that exists in site-specific data curation systems that exist beyond data recorded in the standardised fields found in WARC files. Very different curatorial concerns are reflected in these examples (e.g. the need to locate copyright information for UBC, compared to the need for computer-identifiable outliers with the ‘99’ limits in the Netarchive’s large-scale collecting), and these descriptions above highlight some of the key assumptions and pre-conditions in each setting that determine what is possible. The Netarchive aims for consistency in the longitudinal captures of its national collection and has legal mandate and duty to capture the Danish web domain, which has demanded investment in IT developers and creation of the Netarchive Suite software. In contrast, academic libraries like UBC must adapt to more off-the-shelf systems and develop practices that work with limited-term student contracts. The variations between curation practices in each setting can be seen to relate to each site's constraints such as affordances of software, available staff, time and resources, which all converge to determine how work happens and which data are crawled and collected.

Additionally, this analysis begins to reveal how some categories and decisions are only located within these curation systems and are not reflected in the collection's standardised data artefacts like WARC files – and importantly are not directly available to subsequent computational processes. Yet, these categories are also important for subsequent research and analysis of a collection. For example, categories in UBC’s spreadsheet such as ‘priority’ seeds and ‘important material to capture’ are primarily used to guide the curators in collection development and are not subsequently translated or recorded into any of the standardised data files from Archive-It like WARC files, indexes or crawl logs. Similarly, data fields such as copyright restrictions, or quality assurance notes are primarily written and accessed only by curators. In the Netarchive's case, while data fields like the “99” codes are recorded in the Netarchive Suite harvesting system, the meaning behind those codes is only discussed and documented internally among curators. As a result, many of these categories and judgments used in the curation process take shapes and forms that aren’t amenable to the processing enabled by the set of tools that have been recently (focused on WARC files) and therefore remain largely invisible in interfaces used for access and analysis of archived web data. In contrast to categories like media type, date and time, domain name, and HTML text that are recorded in WARC files, these decisions made in the spreadsheet are located outside of WARC but also determine how WARC files are constructed, which seeds go where, with what settings applied.

With an infrastructural approach, I want to highlight the importance of these categories and classifications applied by curators and embedded in the data curation systems they develop. I argue that web archives can benefit from finding new ways to consider the organisational and cultural contexts in which data selection and curation takes place. I also acknowledge that complete transparency isn't always possible or desirable. Some categories and practices might be difficult to share more widely because of the situated nature of that information, such as some fields in the UBC spreadsheet that include curator names. In the Danish context, there are additional legal limits on which data can be shared and how, due to data protection laws. However, some sense of the different categories being tested, negotiated, and added (including when, where, how, and by whom these data were generated) is important for researchers to know—even if complete documentation cannot be made available to them. For example, researchers could benefit from knowing that several ‘content type’ categories (‘editorial’ or ‘government document’) were used to guide seed selection for the Site C collection. Both researchers and curators must work together to foster a greater awareness of these different processes of classification, and the kinds of categories and decisions made possible within the specific organisational and cultural contexts of their research infrastructures.

FUTURE DIRECTIONS FOR AN INFRASTRUCTURAL APPROACH

These brief examples from UBC and the Netarchive highlight how collecting and curating decisions are made in the context of site-specific systems, categories and data practices. These findings highlight some of the limits of whole-collection research methods centred solely on computational analyses of standard data fields from WARC files. While the web archives community benefits from collaboration and interoperability of tools built around standards like the WARC, it’s important to address other dimensions of infrastructure that emphasise local or situated variations, contingencies and constraints. Even where some degree of universal standardisation is desired, Bowker and Star (2000) have noted that infrastructure is ‘fixed in modular increments, not all at once or globally,’ meaning that the evolving and dynamic nature of infrastructure is reflected in the heterogeneity of archived web data collected in different settings over time. My research stresses the importance of sociotechnical judgments that curators make both with and outside of crawlers and harvesting systems. In addition to the algorithms that follow more strict computational logics and rules, curators contribute to collection management with decisions such as comparing and checking against standard outputs from crawler software, applying heuristics or rules of thumb, and navigating organisational policies or legal requirements. Systems like the UBC collection management spreadsheet and the Netarchive's Broad Harvest set of steps are important evidence of these curatorial judgments. For both researchers studying collections and practitioners like curators or developers supporting their use, these findings reinforce the need to look beyond the standardised data fields of WARC files; the 'data artefacts' produced by crawlers must be described and contextualised. As a result, important curatorial categories, judgments and meanings that shape collection decisions are not always directly available to standardised readings of data through computational tools and interfaces.

Addressing this need to describe and account for data's site-specific sociotechnical context, I argue that when building and configuring web archives research infrastructures, we must take into account the situated or local nature of data. This work therefore leads to additional questions and provocations for how these ideas can be extended in future work, such as: What does it take to describe data in context? How can the development of future tools be informed by the infrastructural approach? How can we build infrastructures that account for 'local' data artefacts, practices and data curation systems, and centre on situated solutions? In this closing section, I highlight two practical ways that embracing an infrastructural approach that can inform the future development of web archives, by proposing how new developments by collecting institutions and researchers might re-centre these local categories and practices.

Building Tools for Non-Standard Data

As a first practical step, I propose that web archives researchers, curators and developers look more closely and critically at how tools and research infrastructures are being developed for standardised forms of data. Current tools are useful for managing how very large scales of data are processed through indexing, as well as plotting trends based on standard data fields in the WARC file. However, when relying on these tools to aggregate and synthesise datasets from different institutions or covering wide spans of time, key differences within the data can be rendered invisible. These tools are designed around standard structures of data, and focusing only on these standard data artefacts cannot account for organisational contexts with disparate collecting practices, or varying cultures, workloads and labour. I consider here how tools can be developed and configured to support new readings of data that emphasise more situated or 'irregular' data and practices.

For example, looking more broadly at the literature from infrastructure studies, Loukissas (2019) provides lessons and insights such as to 'look at the data setting—not just the data set.' Work like Loukissas' temporalities application (developed for analysis of collections from the Digital Public Library of America) embodies this approach by analysing the different date formats in a collection, representing all numerals as dashes to highlight the variation and patterns in how dates and times are annotated. Applying a similar approach to web archives data, there is potential to highlight the variety of metadata formats and schemes that have been recorded in a collection over time, to pinpoint where, when, and who this data comes from. In the case of the Netarchive, analysing different metadata fields might lead to the discovery of several domain-specific limits set to end in "99" for the Netarchive, leading a researcher to look more closely at where, why, when these unique numbers for limits were set and how those domains may be grouped or classified in future analysis. Recognising these patterns in date formats and other metadata can highlight where seemingly unrelated data may have similar origins. In this way, visualisations can be used to reveal the unique processes, systems, or idiosyncrasies at certain points in time in the broader 'archival infrastructure' and can be used to consider data from multiple collections, as well as revealing insights about web and Internet infrastructure like metadata created by certain Content Management Systems. By configuring tools to specifically reveal anomalies, we can highlight the importance of non-standard data, and what hidden meanings or opaque processes it can represent.

Bringing Context to the Forefront in the Research Process

As a second practical step, I consider how new work on 'research engagement' might re-envision entry points into web archives collections. In particular, I believe it's essential to focus on descriptive and qualitative understandings of practices for collecting and curating data at early stages of a research investigation. Individual curators can be invaluable sources of information about data's relationships to their specific settings and research infrastructures. While many researchers may begin to explore a collection through data portals and search interfaces, discussions with curators and librarians can also serve as a starting point and may help researchers navigate local orders and organisational schemes for data in order to construct more robust search queries in their subsequent analyses.

One example of this approach that's already being implemented is seen in the "Probing a Nation's Web Domain" project, a collaboration between Aarhus University's NetLab and the Danish Royal Library (Brügger et al., 2019, 2020). Their early work on the project included discussion with curators to understand which WARC files from which broad harvests should be used in the corpus being constructed for their research analysis. Prior to data analysis, the project team met with curators to determine which data to include, accounting for variations in the Step 1 and Step 2 duration over the ten-year span of the collection's quarterly crawls. The team had to define specific rationale and document their reasoning for which data to include from which time frames, and their work represents an exemplary model for collaboration between curators and researchers, and the kind of documentation and decision-making necessary to select data based on local categories and arrangements. However, it's important to note that this collaborative approach is also a result of the unique configuration of the Danish context, possible because of the resources available there, and necessary due to the restrictions imposed by Denmark’s data protection law, which prevents researchers from accessing raw data directly. The unique conditions within the Danish context impose restrictions on researchers, but also lead to more careful selection and discussions with curators since direct or wholesale data movement from the entire collection is not possible. While this exact model may not translate to other contexts, it can serve as a useful example for future research to develop and document clear selection criteria considering the specifics of an individual web archives collection.

Considering how to foreground an understanding of data’s context in research approaches more broadly, I propose that this requires a combination of investment and dedicated resources, development of new roles, and embracing new models of what web archives research projects entail. On this last point, Ogden & Maemura (2021) describe one such model, proposing several initial steps that researchers may take in working with a collection, which are necessary prior to starting any in-depth computational analysis of data. This early phase of research includes activities such as orientating to the collection’s specific context (within a longer organisational history, a set of policies, available resources, legal framework), auditing the data and available documentation that describes and characterises the collection, and finally, constructing new concepts and conceptual understandings of that collection and its data which can further influence or guide computational analyses. Additional findings from my doctoral research highlight the need for new skillsets and roles on the research team. In particular, I consider how future projects might include ‘data ethnographers’ to complement data scientists or data engineers, i.e. team members dedicated to assembling documents and investigating practices that contextualise data from WARC files. Just as the development of tools and interfaces for research engagement marked a significant shift, these new approaches to supporting research through description of data systems and practices will require significant investment of resources, time and efforts from the web archives community.

CONCLUSION

In order to facilitate researcher engagement with large-scale web archives collections, recent efforts have focused on computational tools for analysis and description. However, understanding how data are curated during archiving and collecting processes requires more information than these tools alone can provide. As a result, researchers are recognising a need to consider the algorithmic processes and computational logics of crawlers and harvesters, and are seeking out log files, metadata, indexes or other documentation to further contextualise datasets. I argue here that an infrastructural perspective highlights the additional need for description that considers the messy and situated forms of data curation and documentation in any given setting, which are increasingly important to understand selection and exclusion decisions applied to a collection.

Through my analysis here, I have explored how the specific configurations of sociotechnical contexts shape data, including how people, organisations, processes and systems work together to determine the composition of a web archives collection. By studying the data curation systems used in Danish and Canadian settings, I highlight when and where individual curators intervene in computational processes by applying their own data categories and classifications. Extending this work to implement an infrastructural approach to description requires reconsidering tool development and allocating resources in order to foreground the situated, irregular and local nature of data, focusing less on universal interoperability and aggregation of data at scale. Overall these findings reinforce the importance of identifying and describing how data are shaped by converging sociotechnical forces, enabling researchers to better understand new kinds of data categories and relationships that can inform their analyses.

REFERENCES

Ben-David, A., & Amram, A. (2018). The Internet Archive and the socio-technical construction of historical facts. *Internet Histories*, 2(1–2), 179–201.

Ben-David, A. & Huurdeman, H. (2014). Web Archive Search as Research: Methodological and Theoretical Implications. *Alexandria*, 25(1):93–111.

Bingham, N. J., & Byrne, H. (2021). Archival strategies for contemporary collecting in a world of big data: Challenges and opportunities with curating the UK web archive. Big Data & Society, 8(1), 2053951721990409.

Bowker, G. C., & Star, S. L. (2000). *Sorting Things Out: Classification and Its Consequences*. MIT Press.

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

Brügger, N., Laursen, D., & Nielsen, J. (2019). Establishing a corpus of the archived web: The case of the Danish web from 2005 to 2015. In N. Brügger & D. Laursen (Eds.), *The historical web and digital humanities: The case of national web domains* (pp. 124–142). Routledge/Taylor & Francis Group.

Brügger, N., Nielsen, J., & Laursen, D. (2020). Big data experiments with the archived Web: Methodological reflections on studying the development of a nation's Web. *First Monday*. https://doi.org/10.5210/fm.v25i3.10384

Hockx-Yu, H. (2014). Access and Scholarly Use of Web Archives. *Alexandria,* 25(1):113–127.

Laursen, D. & Møldrup-Dalum, P. (2017). Looking back, looking forward: 10 years of development to collect, preserve, and access the Danish web. In Brügger, N., editor, *Web 25*, pages 207–227. Peter Lang, New York.

Loukissas, Y. A. (2019). *All Data Are Local: Thinking Critically in a Data-Driven Society*. The MIT Press, Cambridge, Massachusetts.

Maemura, E., Worby, N., Milligan, I., and Becker, C. (2018). If These Crawls Could Talk: Studying and Documenting Web Archives Provenance. *JASIST*, 69(10):1223–1233.

Milligan, I. (2019). *History in the Age of Abundance?: How the Web Is Transforming Historical Research*. McGill-Queen's University Press, Montreal.

Milligan, I., Casemajor, N., Fritz, S., Lin, J., Ruest, N., Weber, M., & Worby, N. (2019). Building Community and Tools for Analyzing Web Archives Through Datathons. *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 265–268.

Ogden, J., & Maemura, E. (2021). 'Go fish': Conceptualising the challenges of engaging national web archives for digital research. *International Journal of Digital Humanities*. https://doi.org/10.1007/s42803-021-00032-5

Ruest, N., Lin, J., Milligan, I., and Fritz, S. (2020). The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 157–166, Virtual Event China. ACM.

Sandvig, C. (2013). The Internet as Infrastructure. In W. H. Dutton (Ed.), *The Oxford handbook of Internet studies* (1st ed, pp. 86–108). Oxford University Press.

Sierman, B., & Teszelszky, K. (2017). How can we improve our web collection? An evaluation of webarchiving at the KB National Library of the Netherlands (2007–2017). *Alexandria: The Journal of National and International Library and Information Issues*, *27*(2), 94–107. https://doi.org/10.1177/0955749017725930

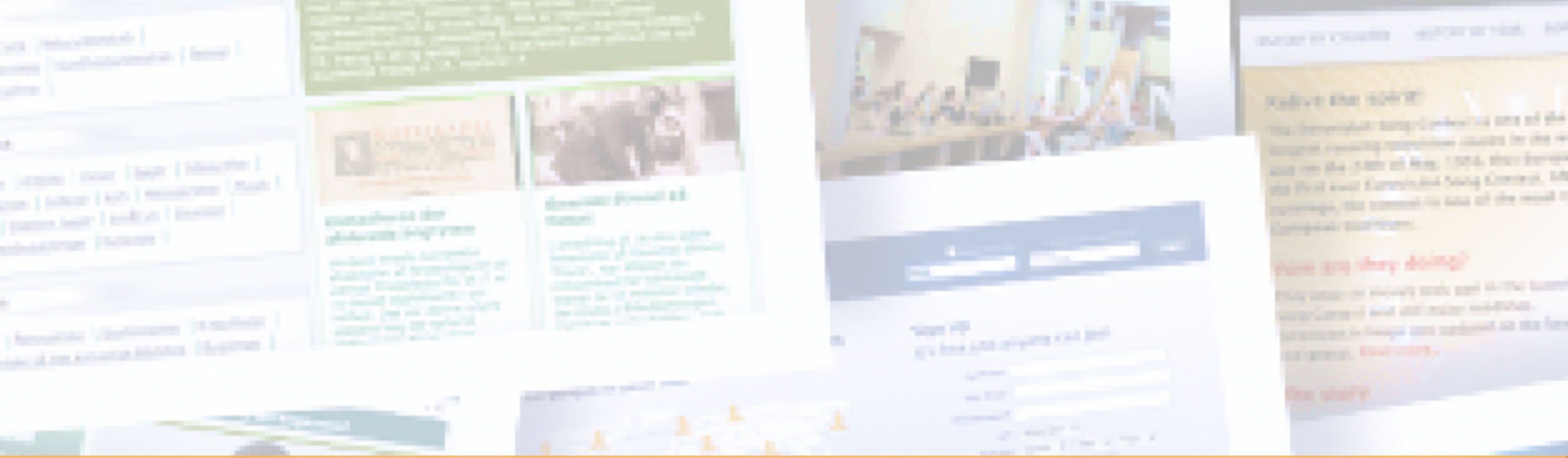
Star, S. L. (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, 43(3):377–391.

Star, S. L. (2010). This is Not a Boundary Object: Reflections on the Origin of a Concept. *Science, Technology & Human Values*, *35*(5), 601–617.

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, *19*(3), 387–420.

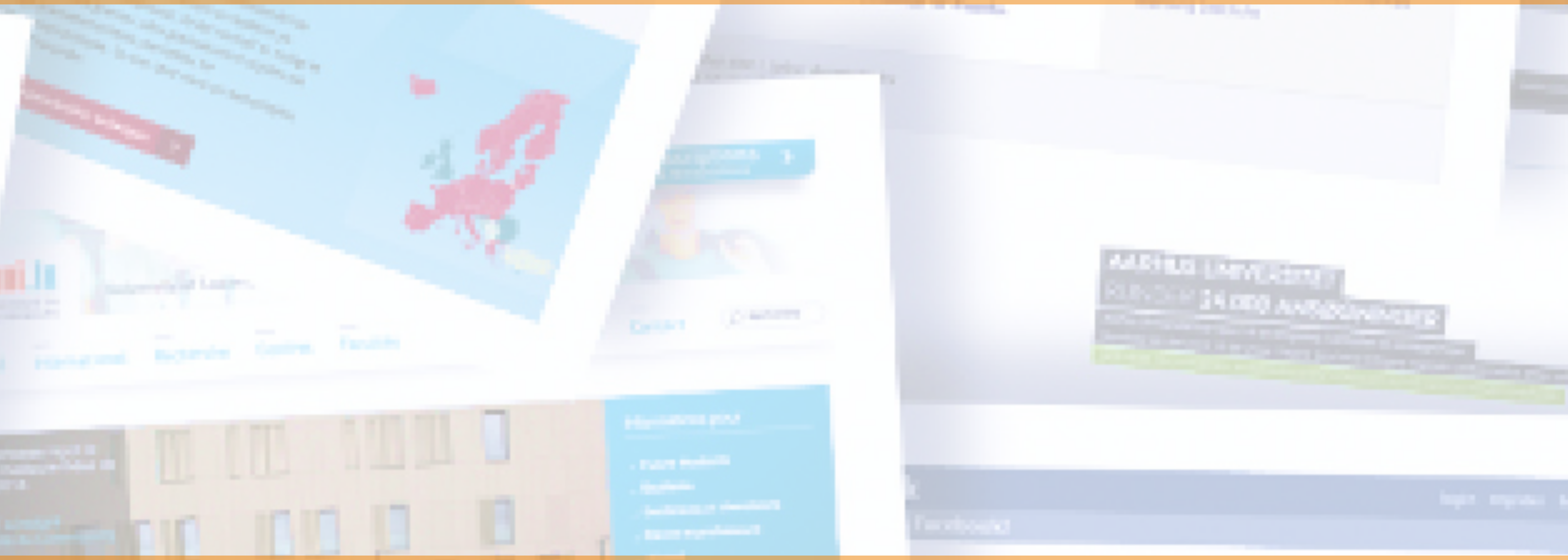
Star, S. L., & Ruhleder, K. (1996). Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research*, *7*(1), 111–134.

Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. International Journal of Digital Humanities, 1(1), 85–111.



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 2020-23, 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