If These Crawls Could Talk: Studying and Documenting Web Archives Provenance

Emily Maemura
Faculty of Information, 140 St. George St., University of Toronto, Toronto, ON, Canada M5S 3G6. E-mail: e.maemura@mail.utoronto.ca

Nicholas Worby
University of Toronto Libraries, 130 St. George St., 4th Floor, University of Toronto, Toronto, ON, Canada, M5S 1A5. E-mail: nicholas.worby@utoronto.ca

Ian Milligan
Department of History, University of Waterloo, 200 University Ave. W., Waterloo, ON, Canada N2L 3G1. E-mail: i2millig@uwaterloo.ca

Christoph Becker
Faculty of Information, 140 St. George St., University of Toronto, Toronto, ON, Canada M5S 3G6. E-mail: christoph.becker@utoronto.ca

The increasing use and prominence of web archives raises the urgency of establishing mechanisms for transparency in the making of web archives to facilitate the process of evaluating a web archive's provenance, scoping, and absences. Some choices and process events are captured automatically, but their interactions are not currently well understood or documented. This study examined the decision space of web archives and its role in shaping what is and what is not captured in the web archiving process. By comparing how three different web archives collections were created and documented, we investigate how curatorial decisions interact with technical and external factors and we compare commonalities and differences. The findings reveal the need to understand both the social and technical context that shapes those decisions and the ways in which these individual decisions interact. Based on the study, we propose a framework for documenting key dimensions of a collection that addresses the situated nature of the organizational context, technical specificities, and unique characteristics of web materials that are the focus of a collection. The framework enables future researchers to undertake empirical work studying the process of creating web archives collections in different contexts.

Received October 13, 2017; revised January 22, 2018; accepted March 19, 2018

 DOI: 10.1002/asi.24048

Introduction

As the live web changes over time, web archiving aims to preserve a record of the web's past. Web archives have been identified as important sources for research, and especially for emerging work in the digital humanities (Brügger, 2016; Gomes & Costa, 2014; Winters, 2017). Examples of this work can be found in recent anthologies that reflect a diversity of approaches to studying web history with web archives (Brügger, 2017; Brügger & Schroeder, 2017). However, enabling research and engaging scholarly use of web archives remains a challenge, as highlighted in a number of initiatives and reports (Dougherty & Meyer, 2014; Hockx-Yu, 2014; Meyer, Thomas, & Schroeder, 2011; Nielsen, 2016; Stirling, Chevallier, & Illien, 2012; Truman, 2016). Emerging work approaches this challenge through the development of tools or solutions for specific use cases (Farag, Lee, & Fox, 2017; Fernando, Marenzi, & Nejdl, 2017).

These new tools and methods allow researchers to analyze web archives data in new ways. For researchers to have confidence in the validity of their findings, however, they must also understand how the collection was created. In particular, given the expansive nature of the live web, researchers must grasp what is included or excluded from collections, judge how they are representative of their object of study, and determine whether an emerging narrative from

these data is indicative or anecdotal. Previous work has tested the coverage of web archives against materials from the live web (Ainsworth, Alsum, SalahEldeen, Weigle, & Nelson, 2011), but these quantitative approaches do not address why, when, or how archival absences occur. Both social and technical factors influence what is captured and preserved for the future, but for any given collection these are often opaque; at best, decisions and constraints are documented inconsistently. Few have addressed the specific ways in which researchers could understand the technical processes, systems, and computing environments used—as well as the human judgments and decisions that shape the archive. This remains a key barrier to the use of web archives for research purposes.

To engage scholarly uses of web archives and support researchers, we must first study web archives practice to understand and communicate the impacts on resulting collections. This study thus asks: How can the sociotechnical process of creating web archives collections be systematically structured and documented? We address this by comparing the process of creating three different collections within one institution and identifying the key decisions that shape the collection and its future use. We seek to explore a new understanding of "web archives provenance," which we consider to broadly encompass what users need to know about how a collection was made as they use, analyze, and make inferences from these aggregations.

Background

Web archiving can take many forms, as described in various introductions to the field (Brown, 2006; Brügger, 2005; Masanès, 2006; Niu, 2012; Pennock, 2013). Client-side archiving through "crawling" captures the web without direct access to the files stored on a server, and is a common approach in the cultural heritage community. Web crawling was originally used for search engine indexing and has been developed for archiving to automate aspects of discovery and capture of web resources. A crawler essentially starts with a list of URLs (a "seed list"), then visits the first webpage on the list and downloads the content and resources linked from that page or embedded within it (such as image or video files). The crawler also identifies any links on that page and can add them to the list of URLs. It then moves on to the next URL on the list, downloads content, and so forth. Several organizations began crawling the web in the mid-1990s, and in 2003 the International Internet Preservation Consortium (IIPC) was formed, with founding members including the nonprofit Internet Archive as well as national libraries and archives (http://netpreserve.org/about-us). Currently, this approach has extended to an increasing number of institutional web archiving initiatives, especially in colleges and universities (Bailey, Grotke, McCain, Moffatt, & Taylor, 2017; Costa, Gomes, & Silva, 2017; Truman, 2016).

The most popular, widespread web archives crawling technology used today is the *Heritrix* web crawler, developed by the Internet Archive and other IIPC web archiving partners. *Heritrix* provides a graphical user interface or can be used on the command line (Mohr, Stack, Ranitovic, Avery, & Kimpton, 2004). The results of crawls are stored in a WebARChive (WARC) file, an ISO standard format developed by the IIPC. The WARC files can then be used in conjunction with Wayback software which renders the archived resources as a webpage in the browser, commonly seen in the "Wayback Machine" service of the Internet Archive, and also available as "Open Wayback," an open source project supported by IIPC members (https://github.com/iipc/openwayback/). Although the *Heritrix* crawling software is free and open-source, it requires computing infrastructure and expertise that may be unavailable in many contexts. Alternatively, there is increasing use of outsourcing some of this infrastructure through archiving-as-a-service, with options available from various providers. This is the model used by many cultural institutions that have adopted *Archive-It*, the subscription branch of the Internet Archive that provides a web-based dashboard interface for the *Heritrix* backend. Collections are stored by the Internet Archive, although partner institutions have the right to download the web archival content locally.

Over time, new tools and interfaces have also been developed to access web archive resources. Although early methods and interfaces such as the Wayback Machine have focused on individual "site biographies" (Nanni, 2017; Rogers, 2013), the current trajectory of tool development focuses on studying collections at scale, analyzing the WARCs of an entire crawl, or aggregating files from multiple crawls and taking these as datasets and an object of study. For example, indexing and search functions have been implemented in beta for the Wayback Machine, and the SHINE interface for the UK Web Archive has been developed through case studies with researchers in the BUDDAH project (https://buddah.projects.history.ac.uk/). Provision of interfaces for data analytics or visualizations, and of web archives APIs, is a major focus of current tool and interface development efforts to support research use (Meyer et al., 2011; Padilla, 2016).

This shift to analyzing whole collections leads to several further challenges and considerations when aggregating data from different crawls. Previous work has investigated how web crawling is an imperfect method of capturing web resources, but some limitations are easier to identify and predict than others. For example, because the crawling process can run indefinitely as new URLs are constantly added to the list, crawl parameters are usually set to determine what is or is not added to the list, or limiting the amount of time the crawler will run. Crawler technologies also have known limitations. For example, trade-offs of additional time and set-up may be required to capture web materials with JavaScript content (Banos & Manolopoulos, 2016; Brunelle, Kelly, Weigle, & Nelson, 2016).

When analyzing a web archives collection, a key challenge is documenting and communicating these limitations, and the nature of absences may not be clear to users. Hockx-Yu (2015) notes that a "resource not in archive" message

DOI: 10.1002/asi

can hide many different technical or legal rationales from the user. Jackson (2015) further notes that this challenge requires collaboration between web archivists and web archives users. This kind of work requires a major shift for many researchers used to working with nonweb-based sources, a reflection on the methods they use, and a critical assessment and systematic documentation of the tools and interfaces they use to access web archives (Ben-David & Huurdeman, 2014; Maemura, Becker, & Milligan, 2016).

Although a significant volume of work has emerged in the form of web archiving surveys, in-depth case studies of collections are still scarce. Previous work begins to address these considerations of selection and scoping of a crawl, and how decisions are then reflected in the resulting web archives collection. Milligan, Ruest, and Lin (2016) compare two collections with varying selection processes, calling for more self-reflective practice and consideration of these choices. Emerging research focuses on the situated practice of web archiving and the activities that are involved in web crawling and constructing a collection. Summers and Punzalan (2017) explore the interactions between the individuals creating web archives and the systems or automated agents used. Their work highlights that the process of selection and scoping is collaborative work between human and machine actors and requires a sociotechnical perspective. Ogden, Halford, and Carr (2017) further this study of practice with an ethnographic approach of the Internet Archive, developing the concept of web archival labor to encompass the knowledge work of human and nonhuman actors involved in collecting and maintaining web archives.

We similarly focus our study on the practice of web archiving, aiming to address the current lack of transparency in communicating the processes of web archiving to web archives users. We see this as a first step toward the design of a system for documenting, systematically, how web archives are made, how this is communicated to users, and how this impacts the kind of inferences made from whole-collection analyses. We use the term "web archives provenance" here to broadly encompass what users need to know about how a collection was made. To expand on this concept, we briefly introduce here select discussions of provenance, focusing on two fields: archival theory and data curation.

Archival theory takes archival provenance as its foundational principle, understood as an essential relationship wherein the records represent and provide evidence of the creator's activities. Douglas (2017) notes how the understanding of archival provenance has evolved, and particularly, how the postmodern turn in archives has led to an expanded notion of provenance, addressing changes to the body of records over time. This work has also called for a more self-reflexive approach to archival practice, acknowledging the active role of the archivist as another "creator" and the power of the Archives as an institution in shaping the records (Cook, 2001; Meehan, 2009). Some within archival theory have called for a new vision of provenance that embraces a broad understanding of the context of records creation, additionally capturing the societal or community context, as well as actions taken by the archivist and subsequent interpretation by users or researchers (Bastian, 2006; Millar, 2002; Nesmith, 2005). A key challenge for archival theory is reconciling an expanded notion of provenance with a system of archival description; such a system seeks to represent the body of records and facilitate access to archives users (Duff & Harris, 2002; Yakel, 2003). It is also worth considering the distinction Douglas (2017) makes between "provenance as process" and "provenance as principle" in archival theory. Provenance as process focuses on operationalizing, tracing where records come from, and is often connected to systems of arrangement and description. Provenance as a principle broadly guides archival practice and addresses the essential question of what an archival aggregation represents. In archival theory, the principle of provenance is constantly critically interrogated, as it defines what is important and essential information to capture and maintain.

Data curation generally considers data provenance as tracing a dataset to its originating computational processes. This is a major concern for data sharing and reuse, and Borgman (2015) describes that when researchers use data collected by other people at different times and places, this "data distance" requires formal knowledge representations to communicate important aspects about how the data came to be. Data provenance in a computational sense is often associated with eScience and frameworks like *Research Objects* from Bechhofer et al. (2013) that focus on identification and traceability of datasets, code, and derivatives by encapsulating them together to enable reuse, replication, and reproducibility. However, other forms of knowledge representation are required to understand the origins of data from social science and humanities perspectives. For example, social studies of science and critical approaches to data like those found in Gitelman (2013) address the challenge for data to shift between local and global infrastructures (Edwards, Mayernik, Batcheller, Bowker, & Borgman, 2011; Ribes & Jackson, 2013; Wallis et al., 2007). The origins of data have also become a prominent concern in social science research with digital methods, especially with the use of "big data," where the processes behind data collection are often opaque (boyd & Crawford, 2012; Helles & Jensen, 2013; Ruppert, Law, & Savage, 2013). Similar questions arise in the digital humanities, where critical examination addresses the construction of digitized texts and how they are transformed into datasets for analysis (Bode, 2017; Trettien, 2013). This shift toward new methods of text analysis, and approaches that take media as data, has led to the development of humanities-focused data curation initiatives (Flanders & Muñoz, n.d.; Posner & Klein, 2017).

Although there is little consensus on the meaning of "provenance" within or across these different fields, each can still provide useful perspectives to inform the work within web archives. These foundational concepts of provenance, and their relation to current discussions of web

DOI: 10.1002/asi

archiving practice, are expanded further in the discussion on new perspectives below.

Research Design

Although greater transparency can enable researchers to evaluate a collection and have confidence in the conclusions drawn from web archives data, this must also be balanced against the time and labor required to generate documentation. Our approach is to first understand how a collection is made, and the different decisions involved, to determine which parts can and should be documented. We compare three selected collections to structure the key steps and decisions made in the process of their creation. The collections are located in an academic research library and use the *Archive-It* service to conduct the crawls, but they differ in scope, timeframe, and mandate. Data collection for this study included: conducting collaborative workshops involving the web archivist and web archives researchers; examining available documentation; inquiries to colleagues previously involved in these collections; and examining the actual systems used and the datasets created in the web archival process. We initially structured our investigation around phases from existing models for creating web archives collections (Bragg & Hanna, 2013; Niu, 2012):

- **Appraisal and selection** includes selection of materials to be captured, and the list of "seed" URLs where the crawler starts.
- **Scoping** decisions determine which resources will be captured, through limiting to specific domains or media type, as well as crawl duration, or crawl frequency.
- **Acquisition or data capture** begins as the crawler runs and accesses each web resource from the live web servers. For *Archive-It* users this process is mostly a "black-box" but reports and logs are generated by the crawler.
- **Organization and storage** has become standardized through use of the WARC file format, supplemented by separate index files. For *Archive-It* users, storage is constrained by data budgets allocated by annual subscription. The management of files and servers is part of the service.
- **Quality assurance** reviews check archived material for "quality and completeness." This can include review of crawl logs and reports to identify the size and number of resources captured by domain and anything not captured due to scoping rules or time cutoffs of the crawl. In addition, individual archived pages can be viewed in the browser to identify missing resources or "leaks" to versions of resources from the live web (not recorded in the web archives collection).
- **Description and metadata** facilitate access and discovery of archived resources. Different levels of detail are possible, because metadata can apply generally to a collection or distinguish individual sites or pages—more granular metadata requires greater time and effort for description.

We used these stages to begin discussion of the workflow in practice and locate key decisions that impact the composition of each web archives collection. These discussions capture the perspectives of a librarian who develops web archives collections, a historian who uses web archives as sources for their research (and is also involved in developing tools), and system design and digital curation researchers. Our aim is to use this framework to describe what happens, highlight when choices are made, and trace their impact.

Description of Cases

All collections are from the University of Toronto Libraries (UTL), where web archiving began in 2005 with *Archive-It*. For the first 8 years, web archiving at the Library existed in a liminal "pilot-phase" with inconsistent staffing and undocumented policy. Web archiving is slowly becoming more integrated with the Library's more general acquisition practices. Part of the shift to web archiving as a higher profile activity was prompted by changes to the dissemination of Canadian government publications: following the end of print depository services (which distributed free print government information to Canadian libraries), almost all levels of government in Canada disseminate government information only in electronic formats and increasingly in HTML. Web archiving has therefore become the primary means for UTL to acquire government information from provincial and municipal governments.

Since 2013, web archiving occupies a substantial fraction of UTL's government documents librarian's role, and is also supported by a few other staff members and student employees. With inconsistent staffing, earlier collections tended to be automated crawls of entire sites rather than dynamic event-based captures that require constant monitoring, selection, and crawling. Currently, staffing for tasks like production crawls and quality assurance can vary from collection to collection. In an effort to mainstream web archiving and make collection decisions more transparent, a new collection development policy was created in 2016. The new policy requires submission of a formal proposal for potential collections that provides documentation on scoping, the frequency of crawls, and long-term plans.

The three collections studied here range in scope and date from different times in UTL's web archiving program, as described below and summarized in Table 1.

- **The Canadian Political Parties and Political Interest Groups (CPP) Collection** is a longitudinal collection established in 2005 as one of the earliest collections in the trial phase of UTL's participation in the *Archive-It* project. The collection consists of quarterly crawls of entire sites of Canadian federal political parties and loosely defined political interest groups. Finding information about the early phases of the collection is particularly challenging due to retirements and transition of web archiving duties between four different librarians.
- **The Toronto 2015 Pan Am & Parapan American Games Collection** captured webpages relating to the sporting events that took place in the summer of 2015. The collection was developed to support the study of "mega-events" and planning by host municipalities across southern Ontario. The

DOI: 10.1002/asi

TABLE 1. Overview of the three collections studied.

	Canadian Political Parties	Pan Am Games	Global Summitry Archive
Collection timeframe	October 2005 to present	February 2015 to December 2016	June 2016 to present
Crawl frequency	Quarterly (every 3 months)	Combination of daily, weekly, monthly and one-time crawls	TBD, based on timing of summit events
Crawl duration	3 days	Varies widely by crawl (from hours to days)	5 days (currently test crawls only)
# of active seeds	62 seeds	434 seeds	167 seeds
Total data archived[a]	>900 GB >29,000,000 documents	>100 GB >3,500,000 documents	>400 GB >5,000,000 documents
Crawl limits and rules specified	Ignore robots.txt	Ignore robots.txt Block twitter.com URLs for "lang=?"	Ignore robots.txt

[a]As of October 2017.

collection focuses on documents and individual pages as well as a handful of full domains, collected over several months leading up to the games. The collection was resourced with a great deal of summer student hours for selection and quality assurance.

- **The Global Summitry Archive** is a collection of webpages of global summits and meetings such as the G20. It is a collaborative project in development with an academic research group outside of the library. It focuses on web materials for individual events, but is also intended to be longitudinal, as these events are ongoing. Almost all web archiving tasks are performed by the research group's staff and research assistants (RAs), and the Library provides consultation and training.

Analysis of Cases

Our discussions revealed that the process of conducting a crawl was iterative, making it difficult to clearly delineate between phases of appraisal and selection, scoping, data capture, storage, quality assurance, and description. It was instead useful to map out the workflow followed to conduct crawls generally. We use the CPP collection as a fleshed-out example below.

Much of the "appraisal and selection" work for the CPP collection was completed in 2005. The rationale for developing the initial seed list was not well documented, as this collection began as a pilot project, but it was based on the federal political party registry listed on the *Elections Canada* website (http://www.elections.ca/content.aspx?section=pol&dir=par&document=index&lang=e). The rationale for which "political interest groups" to include has not been documented. The seed list is updated periodically to match changes in the *Elections Canada* registry, but changes to this registry have not been documented over time. Before a new seed is added, an assessment and test crawl are completed. The librarian first visits the new site and views it in the browser using developer tools to identify any signs or snippets of code that are likely to cause a problem for the crawler based on their own previous experience or known issues like JavaScript. Until the data capture process begins, the exact size and structure of a website and its associated resources may be unknown, so a test crawl is often used to estimate the size of that specific site. After the test crawl is complete, the librarian reviews the crawl reports and identifies any issues, such as calendar pages, which are known "crawler traps"—a set of regular expression rules or parameters may then be used to exclude these pages for subsequent crawls. The data collected from these test crawls are not stored, so many test crawls can be completed if necessary. After this assessment, the new site is added to the seed list for the next production crawl with any additional rules. The production crawl runs automatically every quarter. After the crawl is complete the librarian reviews production crawl reports generated by *Archive-It* for errors or unexpected outcomes. These could include resources left in the queue due to crawl time limits, or large amounts of data from a single site. These reviews can be supplemented by more extensive quality assurance checks by student assistants. Any missing resources identified can then be targeted for a separate "patch crawl" to fill in those resources in the collection after the main production crawl has run.

A similar workflow is used for all three collections and is generalized in Figure 1. This highlights the three different stages of selection and capture. Preproduction includes test crawls to determine and refine a collection's size and scope. Production crawls capture all web resources in the collection. Postproduction quality assurance may require patch crawls to fill in any web resources missing from production crawls. Note that, in practice, multiple iterations may be required for one or more of these individual activities or stages. More specific findings about the curatorial decisions in this workflow are described below.

Finding 1: Scoping and Considerations of Data Budget Are Ongoing Concerns

The three collections vary widely in terms of initial scoping and development of the seed list because each has a different mandate and aims to capture different kinds of material. Scoping work for both the Pan Am and Global Summitry collections involved researchers searching and compiling a list of web resources, whereas CPP is mostly based on the existing *Elections Canada* registry. For a collection such as Pan Am, a great deal of effort was spent in developing seed lists to target individual pages and documents. This included work by summer students as well as

DOI: 10.1002/asi

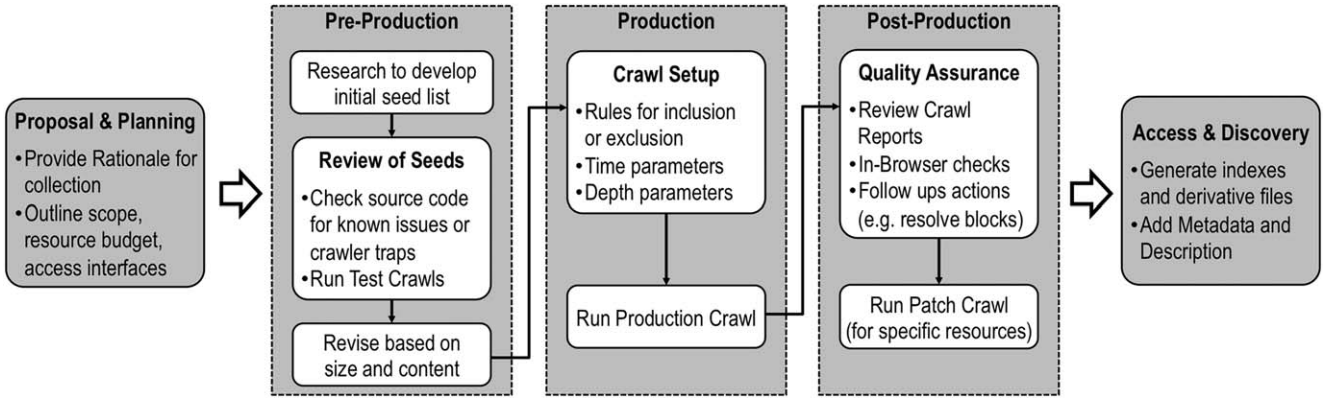

FIG. 1. General workflow for creating a web archives collection at UTL.

consultation with a Planning and Urban Studies professor, requiring a significant amount of scoping work before test crawls could begin.

It was apparent in all the cases that scoping is inherently tied to the data budget, that is, the maximum amount of data captured for all crawls (for all UTL collections) as determined by the annual subscription fee of the *Archive-It* service. UTL's new policy for initiating a collection manages budgets by requiring a formal proposal outlining the intended scope for each new collection. However, the impact of scoping cuts across all phases, and choices relating to the data budget are made and remade throughout different parts of the process. For example, test crawls for the Global Summitry Archive revealed many embedded YouTube videos, resulting in bloated crawls. With the data budget in mind, crawl parameters were adjusted to exclude the capture of video files and limit the collection to the intellectual content of interest to the research team. Late-stage scoping decisions can also arise during QA reviews, when it may be determined that some sites are of less interest and they are removed from a collection.

Finding 2: In Crawling, Unforeseen Issues Are Negotiated in Unpredictable Ways That Require Better Documentation

In the *Archive-It* system, much of the crawling process is automated, but the process of crawling can also require active management and interaction. Because the crawler is an actor in network transactions, its actions can provoke reactions from the administrators of sites being crawled. In effect, the live web can (intentionally or unintentionally) resist the process of archiving, leading to different possible responses from the web archivist conducting the crawl.

For example, with the CPP collection all candidate pages from the Liberal Party of Canada were blocking the Internet Archive (IA) crawler for a production crawl in 2015. Coincidentally, this was during the weeks preceding the 2015 federal election, and in spite of the busy election season, the librarian managed to contact someone from the party's web team who explained that the content management system (CMS) used for candidate sites confused the requests from the IA crawler as a security threat. They had to manually unblock IA's IP range to allow UTL to continue capturing the sites.

A different kind of resistance to crawling is found with robots.txt files. These files can be included in a website to specify how search engine indexing robots interact with the site; for example, specifying parts of the site to exclude from indexing, defining a time limit between requests to the server, or blocking a crawler altogether. Unlike the case of a CMS blocking the crawler, robots.txt files can be ignored and overridden. For archival crawling, the decision to adhere to robots.txt can depend on the institution's policy and legal framework. UTL chooses to ignore robots.txt files, but the crawler can still be impacted by crawl delay times, meaning that certain sites may not be crawled in full. Delays can be adjusted or overridden, but this requires consultation with *Archive-It* support staff and is not a typical dashboard setting.

Finding 3: Decisions Interact and Evolve Over Time in a Changing Context

Through our discussions we discovered that our initial approach to capture individual decisions fails to address a collection's broader context as well as changes over time. The *Archive-It* lifecycle model from Bragg and Hanna (2013) addresses aspects like policy and high-level decisions for vision, objects, resources, and workflow. We highlight here the ways that high-level organizational decisions impact each collection, and how they define a "decision space" in which curatorial choices are made. The timeline in Figure 2 summarizes relevant changes over time for the UTL web archiving program, including the organizational context of UTL, the wider legal and political environment, and the ongoing developments of the *Archive-It* system.

The adherence to robots.txt files, mentioned above, is expanded here to understand the impact and interactions of contextual factors. Early versions of *Archive-It* did not provide a choice to ignore robots.txt—this feature was introduced in 2010 (L. Donovan, personal communication,

DOI: 10.1002/asi

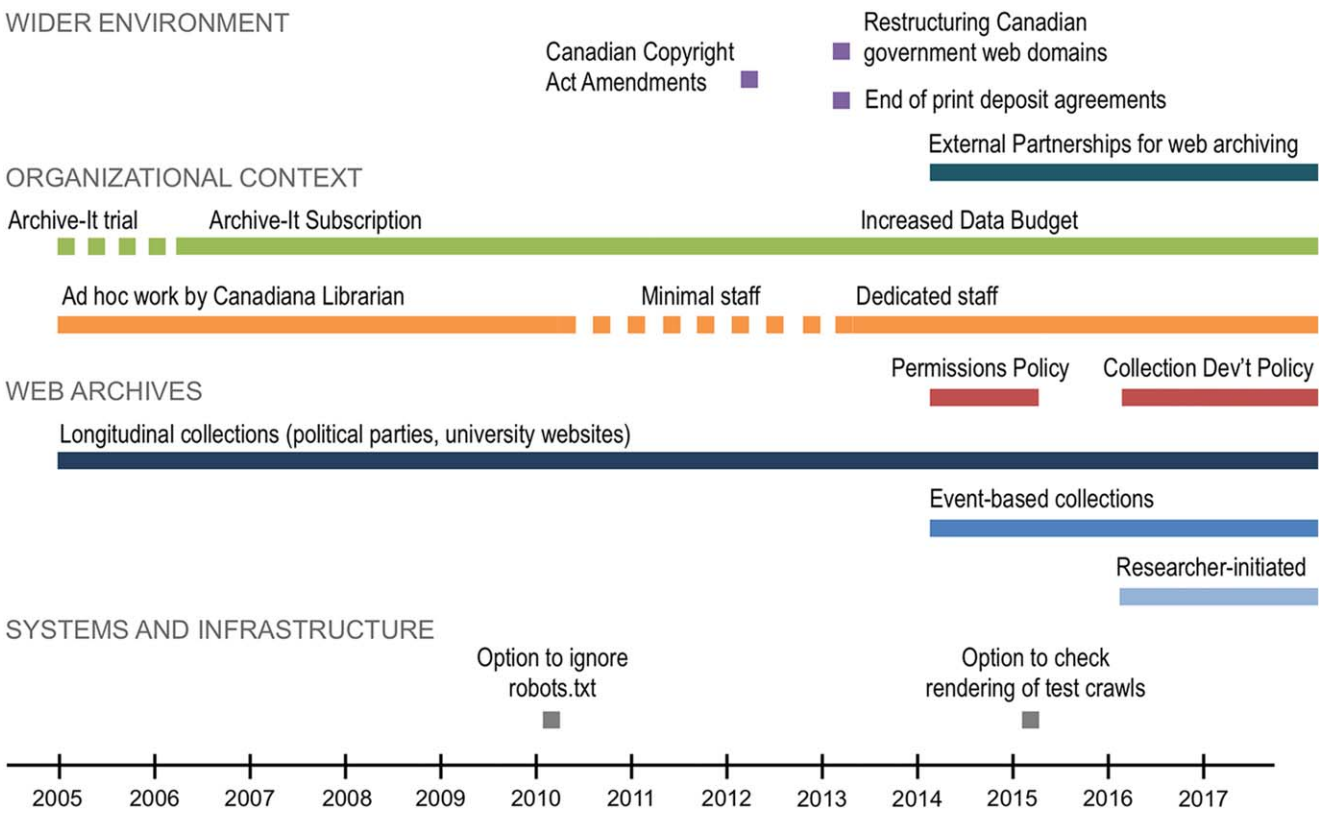

FIG. 2. Timeline of changes to context of web archiving at UTL. [Color figure can be viewed at wileyonlinelibrary.com]

October 10, 2017). When introduced at UTL, permissions were initially requested from each site administrator, then a new policy was adopted to send opt-out notifications only when robots.txt files were encountered. This was prompted by the scale of the seed list for the Pan Am collection, which involved a breadth of individual documents from many different host sites. A decision was made in conjunction with the Copyright Librarian to break from prior policy to contact each one individually with an opt-out notice email, because this would have required excessive time. This decision was only possible with changes to the Canadian Copyright Act in 2012 that makes broader concessions for research and academic uses. Currently, this policy applies to all collections. Comparing different crawls from the longitudinal CPP collection therefore requires considering these changes to the crawler's design, the organizational policies, and the wider legal context, and how they interact to influence what is included or excluded from the crawls over time.

Other organizational directions such as changes to staffing and expansion of the web archiving program to new kinds of collections also prompted changes. For example, the Pan Am collection involved summer students on temporary placements in Quality Assurance. This led to the development of a QA checklist form (based on work by NYARC https://sites.google.com/site/nyarc3/web-archiving/quality-assurance and NCSU http://toddstoffer.github.io/presentation/IIPC-2016/#/). This standardized QA process outlines common issues that can occur and aggregates a list of sites that require further attention from the librarian. Other recent initiatives include collaborative collections with university departments and external groups that bring new challenges because distributed staffing impacts workflows and the resources allocated to certain collections. The Global Summitry collection was the first to be shaped by a new collection development policy for web archiving that required a formal proposal to outline the scope and allocate resources to the project, as well as establish responsibilities. In this case, faculty create a seed list, a research staff member conducts production crawls, and RAs complete quality assurance work with limited Library support. The proposal also had to cover training and management of the collection's scope and data budget. It was found that test crawls to help manage data budgets are crucial when working with researchers beyond the library. Although prompted by the development of specific collections, the new policy and QA checklist were subsequently applied to others.

Ongoing changes to policy, new types of collections, and differences in staff resources and data budgets all impact how web archiving activities are carried out. We have found that the timeline in Figure 2 and its categories are a useful tool to capture the different elements that guide, shape, and constrain web archiving activities and their outcomes over time. Collections are thus not only shaped by individual decisions at discrete points in time, but by the "decision space" that sets boundaries and limitations on the available choices. We begin mapping that space below.

DOI: 10.1002/asi

TABLE 2. Scoping elements.

Element	Key questions and information to document
Motivation	What is the purpose of the collection? Has its mandate changed over time?
Focus	Which geographic, temporal, technical, political, topical and/or social boundaries are defined to scope the collection?
Access & discovery	Who is the intended audience? Do they have known characteristics or needs? Which contractual, organizational, legal, or other agreements restrict access? What metadata fields and indexes support discovery? At what degree of granularity (by collection, site, or individual resource)? Which data formats or derivative datasets are available?
Seed list	What seeds were used in the scoping of the collection? What was the process of discovering and selecting seeds?
Crawl timing	What is the frequency of crawls? How long do crawls run or what time limit is set?
Crawl configuration	What settings control the depth of a crawl? For example, settings for capture by distance from original seed. Is the goal to have a more comprehensive or a breadth-focused collection?
Inclusions and exclusions	Are certain sites or media types included or excluded? For example, are regular expressions used to target certain files or directories in a URL structure.
Permissions from site admins	How were restrictions such as robots.txt and blocks handled?

Documenting Elements of Web Archives Provenance

At UTL, different modes of accessing web archives are available, through the default *Archive-It* web interface or, in special cases like our coauthor's work with CPP, through analysis of the underlying WARC files (Milligan, 2016). Still, some questions that arise for a researcher working with either interface can only be answered through discussions with the librarian, who may also need to consult reports from *Archive-It* or their own documentation. We highlight here the need for a framework to consider provenance separate from the implementation of a given technical system, and this approach goes beyond efforts to improve or develop a particular access interface for provenance needs. Based on the findings noted above, we identify here an initial set of elements necessary to address web archives provenance, proposed to facilitate the elicitation and documentation of decisions in future cases. The elements are described in terms of scoping, process, and context, as listed in Tables 2 to 4 and in the following subsections.

Scoping Elements

Curatorial decisions for a collection's scope determine which web resources will be captured, and over what timeframe. These scoping decisions range from higher-level strategic positions to lower-level operational choices. We've attempted to order the scoping elements in Table 2 from higher-level to lower-level, and from decisions shaping a "collection" at large to those made for an individual crawl. Key decisions relate to the overarching motivation and intentions for starting a collection, and may involve collaboration or coordination among different actors or institutions. In some cases, this motivation may be clearly linked to a specific ongoing institutional mandate, like maintaining a record of government websites. Others may be formed on an ad hoc basis around certain themes or events that may be selected based at the discretion of the curator. Other decisions involve specific configurations of the crawler. These broad categories overlap: Crawl frequencies are often set dependent on target sites to ensure the capture of important content, and seed lists may be generated from other sources.

Process Elements

A number of scheduled and unexpected events occur during the process of running an individual crawl. These are often unpredictable, discovered only as the crawler interacts with individual servers to access resources on the live web. The process can be monitored to varying degrees, and decisions to take action in response to different events of a crawl can influence the presence or absence of web resources in the resulting collection. Instead of a fixed set of process elements, we distinguish between scheduled and unscheduled events (Table 3).

Scheduled events include system notifications upon completed stages that may trigger a review or Quality Assurance process. Unscheduled events include HTTP errors; crawling anomalies such as calendar traps or other crawler traps; resource limits reached due to unexpected volumes of video content or issues with Content Management Systems; and site restrictions that surface through delays that impact crawler performance, or through the detection of a block triggered through an automated filter or manually by a site administrator. Both scheduled and unscheduled events can trigger a renegotiation of collection scope throughout the crawling process.

Context Elements

The context in which curatorial decisions are made includes a range of organizational and environmental factors that guide, shape, and constrain the web archiving activities and their outcome (Table 4). Assuming that web archiving takes place within an organization, these activities must

DOI: 10.1002/asi

TABLE 3. Process elements.

Element	Key questions and information to document
Scheduled events	How are scheduled events handled? For example, does the completion of a crawl and generation of crawl reports prompt new decisions like reappraisal of captured content?
Unscheduled events or process anomalies	Which actions or decisions are taken in response to unscheduled or unpredicted events? For example, when are site administrators contacted directly? What is the process for identifying and capturing resources that are missing from a collection?

TABLE 4. Context elements.

Element	Key questions and information to document
Legal context	What laws and regulations apply to the institution and its web archiving activities? For example copyright laws, legal deposit mandates, user agreements and contracts.
Institutional setting and mandate	What is the organizational commitment to web archiving? What is the role of web archiving within the organization?
Policies and Guidelines	Which organizational policies or regulations affect web archiving activities? Do policies exist that guide and constrain web archiving activities specifically? For example, outlining approaches to permissions, access restrictions, or the division of responsibilities across departments. Have these policies changed over the time period of a collection?
Resources available for web archiving	What dedicated staff resources support web archiving? Which software or platform for crawling is in use? Which storage limits or data budgets limit collection? When are resources significantly increased or decreased over the time period of a collection?

align with that organization's mission or legal mandate for collecting, as well as comply with any regulations or legal, organizational, technical, and cultural constraints. Some of these decisions may be discussed and documented when institutions develop web archiving programs, including determining the goals and objectives of web archiving, the allocation of resources, the principles determining the provisioning of access, and the risk management approach taken. Other aspects of context are given by the environment in which the organization operates and may influence these decisions at a higher level; for example, considering access limitations based on requirements of specific copyright law.

The Need for New Perspectives on Web Archives Provenance

The elements above outline the kinds of questions that users must start to ask of web archive collections, and provide a guideline for what information web archivists might provide. However, as highlighted in Finding 3, these individual questions and answers are only a start, and web archives provenance must also consider the different sociotechnical interactions and entanglements that influence an individual web archivist's decisions as they set up a crawl, considering the evolving interface and options available through systems like *Archive-It*, and various organizational policies. The framework proposed above is limited to a single institutional context, where web archiving is carried out by an information professional in the setting of an academic research library. More empirical work is necessary to better understand how different settings and contexts are grounded in epistemological assumptions underlying their decisions. We highlight several tensions and the need for stronger frameworks and theory to support this work.

Toward New Models and Frameworks

These findings begin to reveal how a web resource is captured in a collection, and how this results from several interrelated decisions that interact in complex ways, changing over time. Beyond the individual case findings, this work also highlights the need for a system of provenance documentation, and suggests that existing models for web archiving practice are inadequate to understand the decision space that influences provenance. This has led us to develop the elements presented above as one potential documentation structure that cuts across separate phases of the lifecycle.

We further observe that, although many web archive initiatives reside within libraries and use "collection development" policies as frameworks to guide this work, collection development is not the ideal framework for considering questions of transparency and evidence in the process. It is increasingly unclear how the notion of a "collection" of resources is applied to different possible units of analysis: individual WARC files, WARCs from a single crawl, or aggregated WARCs over time, or even more broadly applying to aggregated data from multiple collections. In future work, we hope the elements proposed above can be considered, tested, and expanded in different contexts and scales beyond thematic collections, such as national web archives or wide web crawls. Studying these different contexts will also provide contrasting audiences and uses

DOI: 10.1002/asi

compared with the setting of a university research library. Although we have highlighted a few concerns around Access & Discovery as an element of scoping, other concerns might be revealed in terms of audience and unanticipated uses in the setting of national libraries or national archives. More empirical work might also reveal a larger question whether or not a single model of provenance documentation can accommodate these different needs and communities.

Broader Perspectives

Additional perspectives and discussions of provenance from archival theory and data curation can provide a foundation for this type of documentation. For example, Peterson (2015) has called for an "archival description for web archives" in the sense of a systematic approach to capturing provenance information used in archival description standards like the International Council on Archives' *General International Standard Archival Description* (ISAD(G)), Canadian Council of Archives' *Rules for Archival Description* (RAD), or Society of American Archivists' *Describing Archives: A Content Standard* (DACS). From another perspective, computational approaches to data curation on the web have been developed with W3C's Open Provenance and PROV ontology (W3C, 2013). We seek to understand how both of these approaches might apply to web archives systems and interfaces, with two caveats. First, all cannot be achieved with automation, but recognizing and designing for human elements in the system is also important. Second, we should acknowledge the preceding work studying description in archival practice, and be critical of the standards we emulate, to avoid carrying over existing tensions to this new domain. This means, for example, recognizing the power of naming and categorizing, and how description influences whose voices are heard, and which stories are told from the records (Duff & Harris, 2002). Reflecting critically on the human elements in web archiving does not necessarily mean recognizing biases in a negative sense, but also recognizing the knowledge, experience, and judgment of the web archivist. In our case the librarian employs years of experience to review source code for potential crawler traps before running test crawls. The tacit knowledge used in these judgments shapes a collection long before the web crawler system is involved in the process.

Although we have chosen to focus on perspectives on provenance, future work might also explore models beyond libraries and archives, such as those from fields and literature in human and organizational behavior. Further, by aligning with approaches like critical data studies, future work in web archives provenance aim for a deeper understanding of the sociotechnical system involved in production, the limitations and constraints imposed, and the workarounds and invisible labor required to sustain web archives systems. These concerns are central to the recent work of Summers and Punzalan (2017) and Ogden et al. (2017).

Summary and Outlook

The study described here aims to provide a conceptual framework that structures the decision space of web archiving and enables comparative studies of specific cases. In exploring and presenting here the choices made in creating three different web archives collections, we identify a structured set of elements that provide key aspects for understanding this space, including: the mandate and motivations for collecting; the technical settings and affordances of the particular crawler software; the contextual factors of the collecting institution and its wider environment; and the unique parts of the process that happens when a crawler interacts with the live web. A discussion of different perspectives on provenance highlights how these can inform the emergent concept of "web archives provenance" and contribute to an understanding of both the individual decisions that shape a collection through the web archiving process and their situated nature within specific and evolving organizational and technical contexts. The article thus contributes a systematic approach to empirical work that captures choices in context and facilitates the study of the impact of specific technical system on curatorial choices.

Acknowledgments

Part of this work was supported by the National Science and Engineering Research Council (NSERC) through RGPIN-2016–06640, and the Social Sciences and Humanities Research Council (SSHRC) through Insight Grant 435–2015-0011 and Canada Graduate Scholarship 767–2015-2217. Ian Milligan was also supported by the Marshall McLuhan Centenary Fellowship in Digital Sustainability at the University of Toronto iSchool Digital Curation Institute. We also thank our anonymous reviewers for their thoughtful comments that helped strengthen this work.

References

Ainsworth, S.G., Alsum, A., SalahEldeen, H., Weigle, M.C., & Nelson, M.L. (2011). How much of the web is archived? In Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (pp. 133–136). Ottawa, Canada: ACM Press. https://doi.org/10.1145/1998076.1998100

Bailey, J., Grotke, A., McCain, E., Moffatt, C., & Taylor, N. (2017). Web Archiving in the United States: A 2016 Survey. National Digital Stewardship Alliance. Retrieved from http://ndsa.org/documents/WebArchivingintheUnitedStates_A2016Survey.pdf

Banos, V., & Manolopoulos, Y. (2016). A quantitative approach to evaluate Website Archivability using the CLEAR+ method. IJDL, 17(2), 119–141.

Bastian, J.A. (2006). Reading colonial records through an archival lens: The provenance of place, space and creation. Archival Science, 6(3–4), 267–284.

Bechhofer, S., et al. (2013). Why linked data is not enough for scientists. Future Generation Computer Systems, 29(2), 599–611.

Ben-David, A., & Huurdeman, H. (2014). Web archive search as research: Methodological and theoretical implications. Alexandria, 25(1), 93–111.

Bode, K. (2017). The equivalence of "close" and "distant" reading; or, toward a new object for data-rich literary history. Modern Language Quarterly, 78(1), 77–106.

Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world. Cambridge, MA: MIT Press.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662–679.

Bragg, M., & Hanna, K. (2013). The web archiving life cycle model. Internet Archive. Retrieved from http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf

Brown, A. (2006). Archiving websites: a practical guide for information management professionals. London: Facet.

Brügger, N. (Ed.). (2017). Web 25: Histories from the first 25 years of the World Wide Web. New York: Peter Lang.

Brügger, N. (2016). Digital humanities in the 21st century: digital material as a driving force. Digital Humanities Quarterly, 10(2).

Brügger, N. (2005). Archiving websites: General considerations and strategies. Århus, Denmark: Centre for Internet Research. Retrieved from http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf

Brunelle, J.F., Kelly, M., Weigle, M.C., & Nelson, M.L. (2016). The impact of JavaScript on archivability. IJDL, 17(2), 95–117.

Cook, T. (2001). Archival science and postmodernism: New formulations for old concepts. Archival Science, 1(1), 3–24.

Costa, M., Gomes, D., & Silva, M. J. (2017). The evolution of web archiving. IJDL, 18(3), 191–205.

Dougherty, M., & Meyer, E.T. (2014). Community, tools, and practices in web archiving: The state-of-the-art in relation to social science and humanities research needs. JASIST, 65(11), 2195–2209.

Douglas, J. (2017). Origins and beyond: the ongoing evolution of archival ideas about provenance. In H. MacNeil & T. Eastwood (Eds.), Currents of archival thinking (2nd ed.). Santa Barbara, CA: Libraries Unlimited.

Duff, W.M., & Harris, V. (2002). Stories and names: Archival description as narrating records and constructing meanings. Archival Science, 2(3), 263–285.

Edwards, P.N., Mayernik, M.S., Batcheller, A.L., Bowker, G.C., & Borgman, C.L. (2011). Science friction: Data, metadata, and collaboration. Social Studies of Science, 41(5), 667–690.

Farag, M.M.G., Lee, S., & Fox, E.A. (2017). Focused crawler for events. IJDL. Advance online publication. https://doi.org/10.1007/s00799-016-0207-1

Fernando, Z. T., Marenzi, I., & Nejdl, W. (2017). ArchiveWeb: Collaboratively extending and exploring web archive collections—How would you like to work with your collections? IJDL Epub ahead of print. https://doi.org/10.1007/s00799-016-0206-2

Flanders, J., & Muñoz, T. (n.d.). An introduction to humanities data curation. Retrieved from http://guide.dhcuration.org/contents/intro/

Gitelman, L. (Ed.). (2013). "Raw data" is an oxymoron. Cambridge, MA: MIT Press.

Gomes, D., & Costa, M. (2014). The importance of web archives for humanities. International Journal of Humanities and Arts Computing, 8(1), 106–123.

Helles, R., & Jensen, K.B. (2013). Introduction to the special issue Making data – Big data and beyond. First Monday, 18(10).

Hockx-Yu, H. (2014). Access and scholarly use of web archives. Alexandria, 25(1), 113–127.

Hockx-Yu, H. (2015). The unknown aspects of web archives. Retrieved from https://hhockx.wordpress.com/2015/08/11/7/

Jackson, A. (2015, November 20). The provenance of web archives. Retrieved from http://britishlibrary.typepad.co.uk/webarchive/2015/11/the-provenance-of-web-archives.html

Maemura, E., Becker, C., & Milligan, I. (2016). Understanding computational web archives research methods using research objects (pp. 3250–3259). IEEE Big Data *2016*. https://doi.org/10.1109/BigData.2016.7840982

Masanès, J. (2006). Web archiving. Berlin: Springer.

Meehan, J. (2009). Making the leap from parts to whole: Evidence and inference in archival arrangement and description. The American Archivist, 72(1), 72–90.

Meyer, E.T., Thomas, A., & Schroeder, R. (2011). Web archives: The future(s). SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1830025

Millar, L. (2002). The death of the fonds and the resurrection of provenance: Archival context in space and time. Archivaria, (53), 2–15.

Milligan, I. (2016). Lost in the infinite archive: The promise and pitfalls of web archives. International Journal of Humanities and Arts Computing, 10(1), 78–94.

Milligan, I., Ruest, N., & Lin, J. (2016). Content selection and curation for web archiving: The gatekeepers vs. the masses. In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (pp. 107–110). Newark, NJ: ACM Press. https://doi.org/10.1145/2910896.2910913

Mohr, G., Stack, M., Ranitovic, I., Avery, D., & Kimpton, M. (2004). An introduction to Heritrix: An open source archival quality web crawler. In J. Masanès & A. Rauber (Eds.), 4th International Web Archiving Workshop (IWAW'04). Bath, UK. Retrieved from http://iwaw.europarchive.org/04/Mohr.pdf

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

Nesmith, T. (2005). Reopening archives: Bringing new contextualities into archival theory and practice. Archivaria, (60), 259–274.

Nielsen, J. (2016). Using web archives in research: An introduction. Aarhus, Germany: NetLab.

Niu, J. (2012). An overview of web archiving. D-Lib, 18(3/4).

Ogden, J., Halford, S., & Carr, L. (2017). Observing web archives: The case for an ethnographic study of web archiving. In WebSci'17 Proceedings (pp. 299–308). New York: ACM Press. https://doi.org/10.1145/3091478.3091506

Padilla, T. (2016). Humanities data in the library: Integrity, form, access. D-Lib, 22(3/4).

Pennock, M. (2013). Web-Archiving. Digital Preservation Coalition. Retrieved from http://www.dpconline.org/component/docman/doc_download/865-dpctw13-01pdf

Peterson, C. (2015). Archival description for web archives. Retrieved from https://medium.com/on-archivy/archival-description-for-web-archives-1d9dce8dcef0

Posner, M., & Klein, L.F. (2017). Editor's introduction: Data as media. Feminist Media Histories, 3(3), 1–8.

Ribes, D., & Jackson, S.J. (2013). Data bite man: The work of sustaining a long-term study. In L. Gitelman (Ed.), "Raw data" is an oxymoron (pp. 147–166). Cambridge, MA: MIT Press.

Rogers, R. (2013). Digital methods. Cambridge, MA: MIT Press.

Ruppert, E., Law, J., & Savage, M. (2013). Reassembling social science methods: The challenge of digital devices. Theory, Culture & Society, 30(4), 22–46.

Stirling, P., Chevallier, P., & Illien, G. (2012). Web archives for researchers: Representations, expectations and potential uses. D-Lib, 18(3/4).

Summers, E., & Punzalan, R. (2017). Bots, seeds and people: Web archives as infrastructure. In CSCW'17 Proceedings (pp. 821–834). New York: ACM Press. https://doi.org/10.1145/2998181.2998345

Trettien, W.A. (2013). A deep history of electronic textuality: The case of English reprints. John Milton Areopagitica. Digital Humanities Quarterly, 7(1).

Truman, G. (2016). WebArchiving Environmental Scan (Harvard Library Report). Harvard Library. Retrieved from http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314

W3C. (2013). PROV-O: The PROV Ontology. Retrieved from https://www.w3.org/TR/prov-o/

Wallis, J.C., et al. (2007). Know thy sensor: Trust, data quality, and data integrity in scientific digital libraries. In L. Kovács, N. Fuhr, & C. Meghini (Eds.), ECDL 2007. Lecture Notes in Computer Science, vol. 4675. Berlin, Heidelberg: Springer. Retrieved from http://link.springer.com/10.1007/978-3-540-74851-9_32

Winters, J. (2017). Web archives for humanities research – some reflections. In N. Brügger & R. Schroeder, The web as history: using web archives to understand the past and the present. (pp. 238–248). London: UCL Press.

Yakel, E. (2003). Archival representation. Archival Science, 3(1), 1–25.