What’s Cached is Prologue: Reviewing Recent Web Archives Research Towards Supporting Scholarly Use

Emily Maemura
University of Toronto, Canada.
e.maemura@mail.utoronto.ca

ABSTRACT
Web archives are essential to support historical scholarship in the online age. Research on web archives spans many disciplines, often requiring domain-specific expertise. The wide-ranging nature of the literature makes it difficult to obtain a current overview of the field, but this view is needed to identify which core challenges define the field, and assess the different approaches taken to address them. This paper provides such a review of the current landscape of web archives research, focusing on addressing the common challenges faced to support scholarly use of archived web materials. The analysis describes three challenges and identifies key concepts and current approaches for each: (1) how to organize and select from web archives collections; (2) how to critically examine these sources; and (3) how to approach ethics and consent for using archived web materials. The discussion addresses open questions and tensions, highlighting the sociotechnical nature of these challenges and revealing opportunities to apply existing work from the body of knowledge of information studies. It concludes with several recommendations for future research directions to support scholarly use of web archives.

INTRODUCTION
Web archives capture versions of web resources at fixed points in time, in order to prevent the loss of ephemeral web material and preserve digital cultural heritage. The growing number of resources contained in these collections means they are becoming valuable primary sources for researchers from across the humanities and social sciences, who can use these records to study the past 25 years of cultural production and social interaction found on the web. As the amount of materials in web archives collections is increasing, the web archives community must also attend to how these materials are being collected and how they can be used in order to ensure a sustainable future for this work.

‘Web Archives Research’ is used broadly here to encompass the study of all activities involving web archives. This includes the creation of web archives, and the study of activities like how collections are created with technical tools and systems like web crawlers, as well as the organizational and curatorial aspects of collection development. This also includes the study of activities to support use of web archives, through developing access interfaces or specific research methods and techniques.

It can be difficult to compose a clear overview of the current state of Web Archives Research because of the wide-range of topics it includes. This paper develops such an overview by assembling and analyzing the body of literature in Web Archives Research for the past five years. Several common trends are identified from this recent work, and the research related to three specific challenges is analyzed in depth. The analysis reviews and compares key concepts and examples of the approaches to: (1) exploring, organizing, and delimiting a corpus; (2) critically examining collected materials; and, (3) considering ethics, consent and responsibility when using this material for scholarly work. The discussion further addresses gaps and open questions to be addressed by future work, and several recommendations for web archives research and development of theory going forward, particularly highlighting potential points of connection to other areas of information studies. The contribution of this paper is threefold: to assemble the web archives research literature, to identify common challenges through a synthesis of organizational, curatorial and technical perspectives, and to propose an agenda for Web Archives Research that aligns with the sociotechnical nature of these challenges.

BACKGROUND
The community of practice for creating web archives has centered on institutional programs, in particular the work of the International Internet Preservation Consortium (IIPC) which was formed in 2003, with twelve founding members including the non-profit California-based Internet Archive, the British Library, the National Library of Australia, and the U.S. Library of Congress. The IIPC has now grown to a community in over 45 countries who have collaborated to develop common tools, methods and standards for web archiving (http://netpreserve.org/about-us/). In addition to these large-scale collections by national institutions, many other organizations have begun collecting web materials of varying size and scope, such as websites focused on a single topic or event. The different types of collections, tools, and resources used by these different web archiving initiatives are characterized in several surveys that describe and compare web archiving programs at different institutions (Bailey et al., 2014; Bailey et al., 2017; Costa et al., 2017; Internet Memory Foundation, 2010). Truman (2016) surveys a wide range of practices through interviews, providing a series of ‘institutional profiles’ and a summary of tools available for different stages of the web archiving process. Webster (2017) provides a contrasting view to these comparative approaches with a cultural history that highlights the influence of Internet Archive from early on, as well as the many national
libraries working within legal deposit frameworks. Further, recent developments from the Internet Archive and their subscription service Archive-It have focused on ‘Community Webs’ as local history collections, and it is expected the number of smaller collections from local libraries will increase in the coming years (Rollason-Cass, 2017).

The community of web archives researchers is more challenging to locate and characterize due to the interdisciplinary nature of this work. Research publications in recent years have centred on several events and publication venues. These include several conferences and workshops: Web Archiving and Digital Libraries (WADL) workshops at the Joint Conference on Digital Libraries (JCDL), Working with Internet Archives for Research (WIRE) conference in 2014 (http://wp.comminfo.rutgers.edu/nsfia/), the Web Archives 2015: Capture Curate, Analyze conference (https://www.lib.umich.edu/webarchivesconference), and the Research Infrastructure for the Study of Archived Web materials (RESAW) conferences in Aarhus, Denmark in 2015, and London, UK in 2017. Annual conferences include the Web Archiving Conference held in conjunction with the IIPC General Assembly since 2015, and the annual meetings of the Web Archiving Section of the Society of American Archivists since 2013. Research is also found in several special issues of journals, as well as recent books focused on web history (Brügger, 2016, 2017b; Brügger & Schroeder, 2017; Fox et al., 2018; Gibby & Hockx-Yu, 2014;)

Since there is no single venue or central community of web archives researchers, several organizations and individuals have created bibliographies of key articles on web archives (Archive-It, n.d.; International Internet Preservation Consortium, n.d.; Webster & Milligan, 2014). Ayala (2013) assembles an extensive bibliography, with over 100 publications on web archiving, organized by several categories like challenges, legal issues, practices, standards and tools, and research on web archives. Select sources are also highlighted by Brügger (forthcoming) who describes key anthologies and journals, existing web collections, and select literature on web archiving and web history. While these various listings and bibliographies provide initial points of entry to the field, none of the recent work aggregates the wide range of publications on web archives. Such an overview is needed to address how web archives research has developed and understand the direction it currently takes.

**REVIEWING AND ANALYZING THE RECENT WEB ARCHIVES RESEARCH LITERATURE**

I begin by assembling a list of research papers representing the current web archives literature, largely located through the bibliographies and sources listed above. In contrast to a systematic literature search, this approach was chosen because of the wide-ranging venues of publication, and the ambiguous definitions of “web archives” returning non-relevant items when searching article databases. The full bibliography of recent web archives literature - 185 items published between 2013 and early 2018 - is available as supplemental material (Maemura, 2018). These sources serve as the starting point for the analysis below that identifies and develops common themes, trends and challenges.

*From Optimizing Collections to an Agenda for Supporting Scholarly Use*

A great deal of research continues to develop new ways to optimize the process of collecting, and improve the coverage or completeness of web archives materials. This includes, for example, the development of metrics for archivality and methods to capture dynamic Javascript content (Banos & Manolopoulos, 2016; Brunelle et al., 2016; Kelly et al., 2013). There have also been developments to automate discovery of new content for genres like blogs or websites of public broadcasters (Baltussen et al., 2014; Kasiomis et al., 2013). This work focuses on developing solutions to create better, more complete, or more timely web archives; it remains a major challenge for archives to keep up with the changing technologies of the web.

The need for the web archives community to support research users is noted in several earlier publications and reports (Dougherty et al., 2010; Meyer et al., 2011; Stirling et al., 2012; Thomas et al., 2010). This research direction is extended in papers from Hockx-Yu (2014) and Dougherty & Meyer (2014) who ask how web archives can engage users and suggest specific approaches. Hockx-Yu (2014) provides examples of how researchers have worked with the British Library’s UK Web Archive, and recommendations towards a goal of ‘ready-to-download’ datasets for scholars. Dougherty & Meyer (2014) identify several common themes and challenges of web archiving through interviews with librarians and archivists, researchers, and technologists from the U.S., U.K., and Europe. They highlight the lack of research and development to support access, and the need for improved navigation interfaces, as well as indexing and search to support detailed analysis. Interviewees suggest a number of specific solutions, including building community towards development of standards and tools, as well as funding to support both infrastructure and individual archiving activities. Together, these papers set an agenda for the development of search tools, and also emphasize that the importance of community engagement in development efforts.

This agenda has been advanced by institutional initiatives like the development of an Archive Research Services (ARS) workshop by the Internet Archive (https://github.com/vinaygoel/ars-workshop), and a competition by the Portuguese Web Archives to promote use of their collection (http://sobre.arquivo.pt/en/arquivo-pt-prizes-2018-call-for-submissions). Additionally, several ‘researcher-centric’ projects and initiatives have made existing web archives collections available to scholars while also studying their approaches and the outcomes. This includes; projects from NetLab in Denmark (netlab.dk); the WebART project with the National Library of the Netherlands (www.webarchiving.nl); the British Library’s BUDDAH pro-
jekt (https://buddah.projects.history.ac.uk/), and predecessor AADDA (http://www.history.ac.uk/projects/digital/AADDA); and development of Warbase with the Archives Unleashed project in Canada and the US (archivesunleashed.org). It is notable that while these initiatives often serve to further technical developments like indexing tools and search interfaces, they are not framed as ‘user testing’ to evaluate performance of specific tasks; instead, they are framed in a way that ensures researcher perspectives and needs are primary concerns. The insights emerging from these projects are of particular interest here in that they address the goals set out in the agenda to build community.

Through this work, as well as through methodological and theoretical reflections by individual researchers, three common challenges arise (Weber, 2017; Winters, 2017). First, there is a need to further develop specific methods and approaches for researchers to explore and organize a collection. Second, working with web archives reveals the need to identify and address which resources are absent, or outside the scope of a given collection. Third, many of the researchers encounter questions of ethics, asking who has responsibility for how data is collected, maintained, and used, and who ultimately has power to determine what is or is not archived. Approaches to each of these challenges are discussed in detail below, presenting key theoretical concepts that have been developed by past work, as well as how these concepts have been applied in specific studies and projects.

**Challenge 1: Exploring and Organizing a Collection**

Scholars working with web archives need to find and select the content relevant for their research questions. Technical and methodological challenges arise from the scale of data processing. Methodologically, determining the different steps necessary to define a subset (e.g. a corpus or a dataset) can also be a challenge. Conceptual challenges also arise in searching and exploring web archives, as researchers must familiarize themselves with the form of archived material, and how this relates to the definitions and formats of their own discipline. For example Ankerson (2015) uses web archives to study new media objects, and Shumate and Weber (2015) take a social science perspective on collecting web data. Milligan (2016) takes the perspective of a historian, addressing the range of technological methods needed to access and use these new types of sources.

Previous work has developed several central concepts that provide a point of entry to delimit materials for analysis. Five conceptual ‘web strata’ define typical objects of analysis: (1) the web element or individual resource such as an image defined by a URI; (2) the web page, the assembled resources that are rendered in a browser to present a single HTML document; (3) the website, a number of connected pages within a web domain; (4) the web sphere, or the inter-connected network of sites covering a topic or event; and, (5) the web as a whole (Brügger, 2013a). These strata build upon each other, and ultimately all reduce to individual resources, and the hyperlinks between those resources. Archived resources are associated with a standardized time-stamp, and the Memento extension of the HTTP protocol provides a common framework and infrastructure for retrieving different time-based versions across web archival holdings (Van de Sompel et al., 2009). Hyperlinks can also be considered at different scales, and link direction (in-links, out-links), and internal (intra-domain) or external (inter-domain) links may also be relevant for analysis.

Multiple procedures and techniques are often combined to explore a web archive collections. In addition to selection of particular URIs, pages, or sites, other qualities or characteristics of the content may be drawn upon in the analysis. This includes drawing on qualities of text formats like sorting by mark-up structures or performing analyses like named entity recognition or topic modelling. Similarly, approaches to the study of images might perform visual analysis or use embedded metadata like geographic location.

To illustrate how these concepts are used, some select example approaches are discussed here. One approach used by several projects is to focus on a single website as the unit of analysis. Materials are selected by specifying a single URL (or set of URLs) and comparing or analyzing the changes to form and content over time. Rogers (2017) describes the process of creating ‘screencast documentaries’ that assemble images of a website homepage over time to enable a visual analysis of changes to design elements. Nanni (2017) similarly reconstructs the University of Bologna’s website from a variety of sources.

A second approach is to combine the concepts above to develop a dataset with multiple dimensions for analysis. Huurdeman et al. (2013) provide examples for specific research questions studying a collection from the National Library of the Netherlands’ web archives focused on news aggregator site ‘nu.nl’. One example takes articles on different topics (e.g. Hurricane Sandy, Syria, and US Elections) and studies the out-links on those pages, comparing which domains are most frequently linked to as references in these new stories. This revealed which topics had references predominantly to other news sites in Dutch, or in English, and which link to Wikipedia, Google Maps, or YouTube. In another example, individual images are extracted based on a keyword search for “Mubarak” and are then plotted on a timeline for visual analysis.

Further comparing the findings from past work, there are notable differences between using these concepts to organize a collection at different points in the research process. Lin et al. (2017) present examples for how a scholar can begin to explore and orient herself to a collection, using the case of a collection capturing websites from Canadian Political Parties. They de-
scribe methods of creating link graphs as a navigational tool, suggesting that locating central domains or studying out-links can provide entry points for further questions. They propose that a researcher can follow a series of steps to Filter, Analyze, Aggregate, and Visualize (FAAV), and suggest that several iterations over these steps may be needed. This work highlights that users have different information needs in early stages using exploratory methods, compared with moving towards more targeted analysis based on domain-specific content to address a particular research question.

Several open questions and tensions on the processes of exploring and organizing web archives remain to be addressed by future work, in particular: what kinds of analyses align with what forms of corpora or datasets? Many of these recent studies and reflections have surfaced the challenge of connecting available data with appropriate research methodologies. One determining factor is the scope and scale: at a certain size, a corpus requires a shift from methods like close-reading to computational analysis. Both Hockx-Yu (2014) and Ben-David & Huurdeman (2014) note a shift in methodology is required at a certain scale, and distinguish between ‘document-centric’ and ‘whole-collection’ approaches. Further, developing approaches to whole-collection analysis relates to the open question of how standardized or customized data and methods should be. The approach by the Internet Archive in their ARS workshops has been towards standardized corpus formats for research: Web Archive Transformation (WAT), Web Archive Named Entities (WANE), Longitudinal Graph Analysis (LGA) that enable analysis of item metadata, named entities (people, places, organizations), and link graphs, respectively – all derivatives of the standard WebARChive (WARC) file format (Bailey & Goel, 2015). Others like the WebART project note that they have developed custom formats for researchers in the context of ‘data sprints’ but this would not be feasible or sustainable as an ongoing practice (Huurdeman et al., 2013).

Further, it is observed that approaches can fall along a spectrum: some approaches sort and organize by taking advantage of standard data structures in archived materials, for example, taking metadata fields recorded by a crawler like the timestamp, file format, or URL of a resource; other approaches rely upon domain-specific classification or ‘coding’ and interpretation of materials relative to the context a particular research question. Other examples challenge the separation between ‘content’ and ‘container’ by taking the infrastructural elements of the web like tracker cookies as the object of study, or even interrogating the archival search interface itself (Ben-David & Huurdeman, 2014; Helmond, 2017). In addition, one can consider the labour involved in making selections, and how it is spread across human and non-human actors (Maemura et al., 2018; Ogden et al., 2017; Summers & Punzalan, 2017).

**Challenge 2: Critically Examining Collected Materials**

In addition to selecting materials from web archives, scholars also need to evaluate this material in order to understand what is and is not included in web archives. An essential starting point is understanding that archived materials differ significantly from the live web. Past work has discussed the nature of web archives as new versions (not copies) of web materials; the archived version is understood as both a ‘subjective reconstruction’ and ‘always deficient’ when compared to the live web (Brügger 2013b; Brügger & Finnemann, 2013). Brügger (2013a) further describes a web archive as being ‘both incomplete and too complete’, meaning that web archives will never capture all materials from the web, but they may also capture some materials with greater frequency or at scale than others. A scholar working with web archives must both surface the materials absent from a collection, and surface the biases in a collection’s construction that may skew the interpretation of results.

Several authors have identified the need for ‘source criticism’ when working with web archives, and the challenge of determining reliability or veracity of materials (Rogers, 2017; Brügger 2013b; Nanni, 2017). These discuss three main approaches:

- examining the source itself for evidence of absences or inconsistencies;
- triangulating with other sources (online or offline); and,
- tracing the process of creation by drawing on documentation.

It should also be acknowledged that these concepts are not unique to web archives, and connect to source criticism in the traditions of analytical bibliography, archival diplomatics and provenance. However, when working with archived web materials, new methods may be required for performing source criticism computationally and at scale, and some of these emerging approaches are discussed here.

One approach to source criticism is through analyzing inconsistencies within the material, and how this might impact its reliability as a source. For example, absences may be encountered when viewing an archived web page, as Winters notes “A missing image confronts us with a blank square on the web page; a broken link produces an error” (Winters, 2017, p. 245). Analyzing a whole collection, Huurdeman et al. (2015) describe approaches to address ‘the unarchived’ by identifying ‘the aura’ of a collection, defined as the resources just beyond the boundary of a collection i.e. links are present within the archive, but the resources are not captured. Other inconsistencies may be harder to identify or resolve. At the level of the page, there may be temporal inconsistencies between individual resources like text and images captured at different times. At a larger scale, questions of inconsistency arise with visualizations produced from extracted data. Lin et al., (2017) use a bar
chart visualization to observe some ‘suspicious’ statistics of a collection whose distribution of pages from different websites varies widely by crawl date. Similarly, Laursen & Moldrup-Dalum (2017) visualize data from the Danish Netarkivet collection over ten years and the resulting chart reveals uncharacteristic spikes or dips. These visualizations can identify outliers in the overall trend, but it is difficult to determine what this means or if this is correct, since, “nobody knows what the data is supposed to look like.” (Laursen & Moldrup-Dalum, 2017, p. 209).

A second approach to source criticism is to compare multiple sources. At a small scale, a researcher can compare the archived version of a webpage to screenshots of that page from contemporary newspapers and magazines, or descriptions of a page’s content from other archival sources (see e.g. Nanni, 2017; Ankerson, 2015). At a larger scale, several studies have tested methods and approaches for comparing coverage between different web archives, or between archives and ‘the live web’. I focus here on studies that compare coverage between collections for a particular research topic or domain, as opposed to general quantitative evaluation of an archive’s coverage compared to what exists on the live web (e.g. Ainsworth et al., 2011; Ainsworth et al., 2015; Brunelle et al., 2015). For example, Brügger (2013a) considers the coverage of material relating to Danish parliamentary elections by comparing historical network graphs available from the Danish Netarkivet collection and the Internet Archive. Samar et al. (2016) analyze coverage of trending topics for the Netherlands in 2014 by comparing the National Library of the Netherlands’ web archive to the Common Crawl dataset. Milligan et al. (2016) use a case study of Canadian federal elections, comparing collections resulting from three different crawling strategies: starting with curated seed lists, collecting URLs from public tweets for a given hashtag, and broad crawls by the Internet Archive. Hale et al. (2017) similarly compare the archive and the live web for a population of web pages from TripAdvisor.

A third approach to source criticism is to review documentation and trace the process leading to particular datasets or artifacts. This can be difficult since documentation may not always exist, or may not be readily available to the scholar. Certain questions about the composition of the archived data can only be answered with details about how a crawl was conducted. For example, Lin et al. (2017) find that the bar graph visualization provokes a number of questions that may be explained by documentation of the crawl settings, or a change in the crawl policy. Alternatively, when Laursen and Moldrup-Dalum (2017) analyzed their visualization of data collected each year it led them to further examine documentation on the harvests to explain a large dip in 2010; ultimately they traced the anomaly to an error in the data extraction process to generate the visualization which requires long processing times and many steps for data preparation.

While the need for documentation and transparency is becoming more widely recognized, it remains unclear which type of documentation is necessary at what stages of the research process, and how this might be presented as part of an access interface. Recent initiatives like the OCLC’s study of descriptive metadata provide recommendations for metadata to describe actors like collectors and creators, as well as elements about scope or extent of a collection (Dooley & Bowers, 2018). They further acknowledge that future work must address ‘machine-actionable description’ both to extract metadata, and to enable discovery and access at scale. Others have taken up an ethnographic approach that richly describes the web archiving process but is not easily translated to standard forms (Ogden et al., 2017).

**Challenge 3: Approaches to Ethics and Consent**

Underlying all of the methodological challenges described above, a scholar working with web archives must consider the ethical implications of how materials are used, questions of consent, and her responsibility to the people represented in this data. There are (at least) two sets of ethical frameworks that can influence this work: the ethics of collecting (which may be informed by institutional policies as well as professional codes of ethics) and the ethics of performing research (which may be informed or limited by requirements of Institutional Review Boards). Much of the professional and institutional discussions focus on the impact of legislation for copyright and privacy. However, it should be noted that the legal authority for collecting is not enough to address ethical practice, and additionally, assumptions and interpretations of these concepts vary across contexts. Past work by Rauber et al. (2008) focuses on ethical questions to identify and challenge the core assumptions inherent in archiving the web as a publication medium. Dougherty (2013) further describes how use of the web has evolved and reconfigured as an extension of self, resulting in complicated, conflicting expectations and interpretations of the private and the personal in online expression. Additional complications arise when studies span multiple legal jurisdictions with differing stances on issues like the right to be forgotten.

A common approach to manage use of web archives to protect privacy is by placing restrictions or limits on access. Many of the examples above use the Internet Archive whose collection is openly accessible through the public Wayback Machine and additionally via APIs (https://archive.org/help/wayback_api.php). Other collections restrict access to varying degrees – for example, the Danish Netarkivet requires researchers to apply for access, and the UK Web Archive has some publicly accessible content, but access to the full collection is only on-site. Taking a more active stance to manage or monitor use, Rauber et al. (2008) outline technological approaches to identify malicious users, and to classify personally identifiable content.
Another approach for addressing ethical questions and privacy is to consider when and how permission is granted to ingest materials into the archive. In creating the archive, an organization may use an opt-in model that requires the archive to contact individual website owners to inform them their material will be archived which is difficult to manage at scale. The alternative is an opt-out model where archiving proceeds unless a site owner explicitly denies permission. Some organizations infer permission is granted (or denied) by a website’s robots.txt file that can specify which parts of a website may or may not be crawled, and by which specific crawlers. Other archives may have the legal authority to capture material whether or not granted explicit permission by a website owner. Further, much recent work focuses on social media collections, but these raise new questions of ethics and consent, and come with a different set of concerns around privacy and user agreements made directly with social media platforms (and subject to change without notice).

Different approaches may be considered by researchers using web archives, who are guided by frameworks like informed consent of participants. Weber (2017) notes that research ethics boards are still adapting to address web materials, and “a central challenge with regards to ethics of web based research is that many of the existing mechanisms for oversight are working to catch up with technology” (p. 132). Milligan (2016) discusses some of the considerations for privacy when working with GeoCities data, and how this depends on whether or not data is viewed in aggregate, and whether the original author of a website considered its content available to a wide or narrow audience.

**Limitations**

The analysis above is developed from examples drawn from the literature, and is limited to English-language publications. It is important to reflect here on how this research literature is situated, its limitations, and what is absent from this analysis. Many of the projects highlighted here centre on collections by IIPC members like the Internet Archive and national libraries – and would not encounter materials that fall outside these organizations’ mandates, are not in their interest to collect, or are complicated by legal barriers, or technical issues to ingest at scale. While new tools are also becoming available for personal archiving and community archives of web materials – for example Documenting the Now (docnow.io), Web Archiving Integration Layer (WAIL https://machawk1.github.io/wail/), and Webrecorder (webrecorder.io) – the study of personal and community web archives is not well represented in the current literature. If and how more personal and community collections can be used in research remains an open question since the archives may have missions at cross-purposes to scholarly communications, or work outside the legal and ethical requirements of academic institutions.

**GAPS AND RECOMMENDATIONS**

The above examples have addressed the different methods, conceptual frameworks and approaches that are being developed for supporting scholarly use of web archives. First, while the original agenda to engage users had set the goal to develop search and navigation interfaces, these recent studies reveal that scholars explore these collections through a series of complex processes for selection and delimitation of archived web materials. Even before starting a search, a scholar must gain an understanding of both a collection’s content, scope and focus, as well as the unique material conditions and constraints of archived web artifacts. Jackson et al. (2016) describe that many scholars ‘don’t know where to start’ when approaching web archives and there is little available to guide scholars in this exploratory work. Second, the need for documentation and transparency in all steps of the process of creating and using web archives is a paramount concern. They also illustrate that source criticism may require an iterative process involving many different views of a collection. Statistical analyses may identify gaps or inconsistencies in a collection, leading to close inspection of individual elements. But it is difficult to address exactly why these differences occur and how that impacts interpretation or analysis, and some questions can only be answered by consulting with the individuals involved. There is generally a pressing need for documentation of the process of creating web archives (Brügger, 2013a; Winters, 2017). Finally, there is simply not enough discussion of ethics in the field (with the exception of the recent National Forum on Ethics and Archiving the Web https://eaw.rhizome.org/). New approaches are needed that go beyond the binary of ‘opt-in’ or ‘opt-out’ models, and that are possible to scale with current resources.

Across all these challenges, there is also a need to balance between human and non-human labour, especially for large-scale collections. While the entire World Wide Web of 1997 was estimated at a size of 650,000 sites, today a single national web archives collection can rival or exceed that size - the Danish Netarkivet Collection contained over 1.2 Million unique domains as of 2015 (Brügger et al., 2017; Gray, 1996). Managing these collections of millions of individual web resources, is further complicated by the need to manage multiple versions of each resource captured at different times, over decades. Providing access to data at this large scale will require methods from computer science to address technical challenges, as is reflected in much of the recent literature that focuses on addressing user needs through technical development. However, I argue that the analysis presented here reveals that these challenges to support research use of web archives are sociotechnical in nature.

The body of knowledge from Information Studies (IS) therefore offers opportunities and an existing body of research that web archives might draw upon. Several possible points of connection between core web archives concepts and existing ap-
proaches from IS are identified and described here. First, there is an affinity between the various approaches to delineate a corpus or dataset, and the long history of classification, representation and knowledge organization. The existing literature on information seeking behaviour and user studies in information retrieval would provide a useful approach for studying this question of ‘where to start’, as well as concepts like conducting reference interviews to identify information goals and needs.

Approaches to critically examining collected materials can draw from precedents in source criticism, and archival diplomacies (as also noted in Winters, 2017). The concepts and theory of archival provenance, as well as media-specific analysis, can provide approaches to understand how materials are embedded in a specific, historicized sociotechnical context. These approaches reframe the problem – it is not simply a matter of providing efficient or effective access, but of carefully considering the impact of information organization and representation, and how different information structures can emphasize or elide the ways the archived materials are shaped by the constraints of web medium, and web infrastructure – and the infrastructure of the archive itself.

Addressing ethics and consent is more difficult, and I highlight here the need to bridge between the legal frameworks that influence collecting and the ethics frameworks of conducting research. Many existing collections that originate from legal deposit consider websites to be ‘publications’ and the primary framework influencing research use with this material is copyright law. Other large-scale collections like government web archives treat materials as government records. However, as the web has evolved, ‘publication’ and ‘record’ no longer seem adequate as metaphors for the activities on social media sites and other platforms. It can be argued that these materials should be treated less like individually authored publications, and more as traces of communications or activities of daily life – for research purposes these should be treated as data of human subjects. From a researcher perspective, working with this data requires mechanisms to ensure informed consent of web creators, as well as opportunities for creators to specify how their data may or may not be used in the future. Discussions in critical library and information studies may be relevant for web archives, and similar work that addresses power and politics of creating collections. A more practical approach might be found in traditional mechanisms like donor agreements for archives. Ultimately this may also require re-evaluating the purpose of web archives, and redefining the role of web archivists as more active mediators between researchers who use web archives and the individuals who are the subjects of the data collected in those archives. For the ethics and responsibilities of using web archives as scholarly sources, many questions lead back to the overarching purpose of archives, such as who these archives serve, and to whom they are accountable. Web archives might also benefit from reviewing how archives have been framed in the past, particularly the archival paradigms described by Cook (2013) – archives as evidence, memory, identity, community – and which of these paradigms fits with the goals of different web archiving initiatives.

Finally, like other work in the digital humanities, a core question for web archiving is how to use computational methods without necessarily having the work driven by computational thinking; there is also a need for perspectives that critically reflect on the construction of data ontologies used in the development of tools and quantitative methods. For example, computational approaches that can be used to seek out absences in a collection may be grounded in a particular position of truth and veracity that presumes completeness can be measured quantitatively against a perfectly complete past version (i.e. a Platonic ideal). In contrast, other approaches adopt a humanistic perspective to create a transparently subjective reconstruction of a past web artifact, consciously acknowledging the limits of reproducing the past state of an object (questioning the limits of identity, permanence, and fixity). As web archiving develops and engages researchers from across humanities, social sciences and computational sciences, more work is needed to address and explicate the fundamentally different world views represented in these different areas of research. The tensions amongst different methodological and epistemological viewpoints in web archiving were noted by Dougherty and Meyer (2014) and in many ways this remains a central challenge. While it is necessary to develop shared practices and standards for this large-scale distributed work, it is also important to address how web archives can support a range of scholarly work with diverse models, methods and approaches.

In summary, I highlight here five areas for potential future work, to address the current challenges and gaps in the literature:

1. **Conduct studies to characterize user communities**: allowing for web archives to more clearly define and characterize their community of users (existing or aspirational) – and acknowledge when these users are not from a cohesive singular community. Recent work by Costea (2018) addresses users and non-users, but there is a need for more studies aimed at addressing the epistemological and ontological orientations of user communities towards web materials and how that impacts their needs – whether these are practical needs like access to data via APIs, or conceptual framing of the collected materials like an interest in preserving particular interface behaviours, or paratextual element like targeted ads.

2. **Explore new approaches to consent**: developing models and approaches to engage with web content creators in order to discuss and agree upon future uses of data. One option explored in work and project like Documenting the Now (docnow.io) is to collaborate directly with particular online communities to develop archives of their materials. More work is needed that explores researcher perspectives, and the ethics concerns arising from use of smaller, curated, community-driven collections, in contrast to larger-scale collections with access and privacy restrictions.
3. **Test methods across collections:** contributing more cases of testing how methods perform across different collections. Existing work tends to focus on single collections, or the context of a national archive, and these can be extended to develop more common statistical evaluations, and heuristics for selecting a particular version of a webpage (for a start, some ‘rules’ are presented in Brügger, 2011 and ‘probes’ discussed in Brügger (2017a)). In order to test methods across a wide range of web archiving infrastructures, funding opportunities are needed that can support work internationally.

4. **Develop standard concepts and models:** particularly, developing precise and consistent core terminology to describe key concepts, that can bridge between abstract concepts of researchers and connect to the highly specific language of technical work. Potentially this could develop into a formal model for the kinds of problems being tackled by different studies of web archives research.

5. **Ensure continuous availability of research outputs:** As a more practical recommendation, the web archives research community would benefit from a common repository or database of literature. This paper and the accompanying bibliography are intended as a start to building such a resource to aid those entering the field.

I hope these points can serve as an agenda for web archives research and support further development of web archives theory.

**CONCLUSION**

The review of recent web archives literature presented here reveals the core challenges to support and engage scholarly users, and argues that the sociotechnical nature of these challenges can benefit from the theories and approaches of Information Studies. The discussion section further identifies gaps and recommends specific areas that future web archives research work might explore. This overview of literature and research also begins to highlight the need to critically reflect upon collecting practices, evaluating how they are meeting the needs of users in order to ensure the overall utility and sustainability of future of web archiving work.

**ACKNOWLEDGMENTS**

This work is supported by the Social Sciences and Humanities Research Council (SSHRC) Canada Graduate Scholarship 767-2015-2217 and Michael Smith Foreign Study Supplement.

**REFERENCES**


COPYRIGHT

The standard copyright permission is included. This may be to be modified should you wish copyright to be retained by someone other than the authors.

81st Annual Meeting of the Association for Information Science & Technology | Vancouver, Canada | Nov. 9 - 14, 2018

Author(s) Retain Copyright