## 12. Web Archives, Critical Digital Literacy, and the Growing Primacy of Born Digital Objects. Cal Murgu

Cal (cmurgu@brocku.ca) is Instructional Design Librarian at Brock University, Canada.

### Introduction

*It is hard to imagine how one might study the history of the developed world in the late twentieth and early twenty-first century without recourse to the archived Web.*

Jane Winters (2017, p. 238)

This chapter focuses on two closely related fields in critical digital pedagogy: archival studies and technology studies. It at once draws attention to the increasing primacy of born digital files for research and knowledge creation, while questioning the positivism that is often associated with technology. It supports the claims of Churchill and Van House, who write that "what is remembered individually and collectively depends in part on technologies of memory and the associated socio-technical practices, which are changing radically" (Churchill and Van House, 2008, p. 296). It does so by looking at one specific form of digital memory capture and creation: web archives. This chapter does gesture at theoretical discourse, including the work of Jacque Derrida and Paulo Freire; however, I focus on concrete applications of Web archives in library-led pedagogy.

I make the case for librarians involved in critical pedagogy, especially those interested in information and digital literacy, to introduce students and researchers to Web archives and the Web archiving process. Why? Web archives preserve the contents of the World Wide Web for posterity. This may seem to be a relatively straightforward process absent of choice and subjectivity. However, a critical approach to archival practices challenges notions that regard archiving processes as neutral and objective. Scholars have shown how the decisions involved in archiving, such as what to archive and how archives are made available, challenge our understanding of archives and archivists as neutral custodians of records (Schwartz and Cook, 2002). From time to time, the true subjectivity of these processes is laid bare. The Windrush scandal in the UK — a 2018 scandal wherein the Home Office willfully destroyed landing cards used to verify immigration status — is an example of the impact of negligent records policies on

individuals' lives. The very same is true in the context of the Web. Take, for instance, the Trump administration's decision to remove peer-reviewed scientific material on matters of climate change from the websites of the Environmental Protection Agency and the departments of the Interior, Energy, Agriculture, and State. However, the process of archiving the Web — what Niels Brügger refers to as creating 'reborn digital' artefacts — is significantly different from that of archiving analog documents (Brügger, 2016). The process of using Web archival collections is different, as well.

One challenge for Web archive use is the sheer abundance of content within Web archival collections (for example, as of 2018 the size of the Internet Archive reached 40 petabytes of data — approximately 40 million gigabytes). Another challenge is that websites, unlike most archival material, tend to change over time. The size of Web archives, as well as the technical skills required to explore this data in a meaningful and efficient way, presents an unprecedented challenge for researchers, students, and librarians. While the technical novelty of Web archives may deter some students and researchers, I argue that Web archives provide an opportunity for intellectual growth and skill development for students and faculty in a completely familiar context: primary source analysis. After all, the archived Web will be a crucial source for humanities and social science researchers interested in the recent human condition (consider, for instance, how strange it would seem if a scholar of the 2000s chooses not to include Web resources, instead focusing solely on analog documents as evidence).

It would be appropriate to consider Web archives and the content that they hold as one element of the massive discourse surrounding big data. As has been argued elsewhere, there is profound inequality between those who have access and know how to apply data-oriented methods and those who do not (D'Ignazio, 2017). As D'Ignazio writes, in the context of big data "knowing how to collect, find, analyze, and communicate with data is of increasing importance in society. Yet, ownership of data is largely centralized, mostly collected and stored by corporations and governments. Critically, the technical knowledge of how to work effectively with data is in the hands of a small class of specialists" (D'Ignazio, 2017, p. 6). Mark Andrejevic has termed this the "Big Data Divide" (Andrejevic, 2014), and boyd and Crawford in their seminal piece have referred to a distinction between "data-haves and have-nots" (boyd & Crawford, 2012). Web archives represent another field of play on which these disparities are evident.

Data literacy is important for all students, irrespective of their level of study or disciplinary focus. As MacMillan (2014) asserts, instruction in the use of data resources is relevant both at the undergraduate and graduate levels, and Shorish (2015) argues that data literacy skills are relevant even if students do not continue their studies to attain more advanced degrees. D'Ignazio makes the point, with which I agree, that we require "creative" solutions to empower students with data literacy skills. Creative solutions, D'Ignazio argues, will best accommodate "non-technical learners who may need an alternative to the traditional quantitative approach to working with data" (2017). Given the technical and instructional abilities of librarians, strategies for incorporating creative data literacy into library instruction has been a popular topic of research and pedagogy in academic libraries (Prado, J. C., & Marzal, M. Á., 2013; Koltay, T., 2017). Examples of creative approaches to data literacy include data murals (Bhargava et al., 2016) and mapping social inequality data with ArcGIS (Hoskins, 2019). Albeit in a slightly different context, Paulo Friere foreshadowed the emancipatory potential of data literacy in 1987, when he characterized the process of literacy education both as a technical act and an act of learning the tools of expressing oneself: "to learn how to read is to learn how to say the own word. And the own human word imitates the divine word: it creates" (Freire, 1987, p. 11).

While Web archives present a new technical challenge, we must continue to interrogate them using critical approaches by considering their authority, value, labor, and ethical use. As Alan Tygel writes of big data in general, focusing on the technical use of computers, specialized software, and computational methods creates a "tendency of the [students] to leave behind the critical reflection about the social meanings of data in the world, and therefore the emancipatory perspective may be put in background" (Tygel and Kirsch, 2016, p. 113). More recently, Stine Lomborg has made the more explicit case of emphasizing ethical considerations of Web archive use and creation, arguing that Web archives prompt questions about "not just what kinds of data can be collected and archives, but also what can be used in research" (Lomborg, 2019, p. 100). In particular, Lomborg complicates the notion of Web archives as simply repositories of digital textual material; for Lomborg, Web archives are also repositories of the actions of human subjects on the Web. Lomborg's distinction rightly emphasizes the need for risk assessment and human subject protection, among other ethical considerations.

The critical approach I put forward here mirrors that of Juliet Hinrichsen and Antony Coombs, who argue in favour of an educational shift away

from a simple "skills agenda" and towards "the idea of situated practices" in an effort to "to rebalance the emphasis on … operational skill with a focus on the practices and intellectual traditions of the disciplines both as meaningful sites of learning and as a reflection of a shift towards critical academic literacies" (Hinrichsen and Coombs, 2013, p. 4). For Hinrichsen and Coombs, the critical dimension refers to an internal and external criticality: internal refers to the "faculties of analysis and judgement as applied to the content, usage and artefacts of the technology," while the external meaning "relates to a position regarding the development, effects and social relations bound in technology" (Hinrichsen and Coombs, 2013, p. 4). In other words, equal attention needs to be given to the use of technology, as well as the impact of that technology in larger socio-cultural contexts. Librarians, given their ability to critically consider both the technical (infrastructure, resourcing) and the social (critical and digital literacy, and the ethics of preservation and use) are perfectly positioned to teach students and researchers about the internal and external implications of Web archives. This chapter means to extend the efforts of those committed to critical digital literacy—defined here as a set of abilities, competencies, and analytical viewpoints that enable a user to operate, understand and create digital media, data, and tools—to include the growing primacy of born-digital documents in Web archives and their ephemerality as an important talking point (Kotlay, 2015).

## Archiving the Web

Nearly 30 years ago Steve Lubar wrote of archives, "[they] reflect and reinforce the power relationships of the institutions that organize them; they represent not just a technological solution, but also an organizational solution. They document and carry out not only knowledge and technique, but also culture and power" (Lubar, 1999, p. 16). While this continues to be true, as Brügger states, "the politics of the Web archive lie not only in the selection policies, but also in the more intangible politics of the archiving process itself" (Brügger, 2018, 73). Therefore, to teach about Web archives requires some knowledge about how they are created, accessed, and used. Some may find the use of the term "archive" peculiar in the context of the Web. However, consider that we use the verb "to archive" in digital settings: for instance, your email clients enable you to archive emails for storage and preservation. Indeed, in the context of cloud computing where there is a seemingly infinite amount of space the notion of compression and archiving is quaint; nevertheless, the Web is more ephemeral than we like to admit.

Web archiving is the process of collecting the constituent parts of the Web into files that are amenable to preservation and future use. Web archivists typically employ Web crawlers for automated capture. The largest Web archiving organization based on a bulk crawling approach is the Internet Archive, which strives to maintain an archive of the entire Web (https://archive.org/Web/). Institutions that archive the Web typically use specialized software to download copies of websites and preserve them in a standardized format. The crawling tools begin with a "seed URL" (the homepage of a particular website, for instance) and the crawler follows the links it finds, preserving content as it navigates from link to link. This process is not entirely automatic and is often managed by technical staff and crawler instructions. For example, individuals include scoping instructions for the crawler to follow particular links based on permissions policies. The frequency of crawls varies from institution to institution. However, since the goal of Web archiving is to track changes of websites over time, most sites are archived more than once over a given period. The frequency of collection varies depending on the site and decisions made when the site is selected for inclusion in a particular collection.

The result of crawls is stored in a standardized format: the WARC format (WebARChive). A WARC file (ISO 28500) is a container for archived websites and rich metadata that enables preservation, indexing, and access. Consider how intricate modern websites are today: hundreds of lines of HTML markup, CSS for styling, JavaScript for dynamic Web design and user interactivity, videos, links to other pages, and so on. WARC files contain all of this content, in addition to important metadata headers that describe a file's provenance, among other things (ISO, 2009). There are some limitations to what can be crawled, however. For instance, streaming media, database content, and user inputs are not included in Web crawls. Nevertheless, the WARC file (and the ARC file before it) enable tools such as the Internet Archive's WayBackMachine (https://archive.org/Web/) to arrange, index, and render snapshots of websites from the past while maintaining certain interactive aspects, such as links and images.

## The Value of Archiving the Web

I have briefly established how we archive the Web; now I will turn my attention to why it is worth the trouble. The value of preserving born digital documents emerged in the late 1990s as various institutions, including non-profits like the Internet Archive and public organizations like national libraries, recognized the relative ephemerality of Web content

(see, for example, the Library of Congress' policy on Web archiving) (Library of Congress Collections Policy Statements Supplementary Guidelines, 2017). The question of why we should invest such resources into archiving the Web becomes clear when considering the fleeting nature of a website. While estimates vary, a 2014 study on the longevity of Web pages found that the average lifespan of a website was around 1,132 days (Agata et al., 2014, p. 464). An additional point gestures to a problem unique to Web archives: beyond domains simply expiring, websites change over time and are sometimes unrecognizable after only a few updates. The ability to manipulate digital objects to such a degree in such a short time is a testament to the power of digital production; however, it poses a serious challenge to our traditional methods of archiving objects, even digitized objects, because these methods assume a certain level of fixity.

Web archiving is a largely decentralized practice. Given the familiarity institutions such as museums, libraries, and archives have with archiving methods and best practices, they have been leaders in Web archiving (both in terms of developing software for Web archiving as well as implementing it in appropriate workflows and policy). Web archiving initiatives are often underpinned by collection development policies that circumscribe the specific theme or subject that the collection focuses on. For instance, a Web archiving initiative could be as specific as a discrete movement or event (see, for example, the Arab Spring Web archive) (858 Archive, 2011). The Library of Congress (LoC) uses a "named subject, event, or theme-based collection" strategy that sees staff members recommend collections for archiving. Alternatively, a mandate to archive official government websites or top level domains of specific countries is common among national libraries and archives, such as Library and Archives Canada (http://Webarchive.bac-lac.gc.ca/?lang=en). In the UK, the UK Web Archive's mandate has expanded since 2005, from only archiving the websites of leading UK institutions based on selection criteria such as historical, social, and cultural significance, to archiving the whole of the UK Web domain (.co.uk) (Brügger and Schroader, 2017).

With this technical and contextual information, Brügger's argument, that the politics of a Web archive lie "not only in the selection policies, but also in the more intangible politics of the archiving process itself", begins to come into focus (Brügger, 2018, 73). Web archiving is, in fact, a highly subjective practice. Web archiving is obscured by the black box of technology, such as automated Web crawlers and Internet standards. However, Web archiving, like the process of curating traditional archives,

is influenced by the technical and human resourcing available as well as the collection strategy of the institution doing the archiving. As Dougherty and Meyer suggest, the specific epistemological assumptions made during the collection and curation of Web archives are either not made explicit to potential users or are seen as an impediment to their use and/or re-use (Dougherty and Meyer, 2014, p. 2198). Decisions about which websites are deemed valuable enough to archive depend on a relatively small group of individuals occupying positions with the power to enact this type of process. In focusing our attention on the human acts driving Web archives, Ogden et al. propose the concept of 'Web archival labour' to represent the "ways in which Web archivists (as both networked human and nonhuman agents) shape and maintain the preserved Web through work that is often embedded in and obscured by the complex technical arrangements of collection and access" (Ogden et al., 2016, p. 1). Furthermore, while the process of automated Web crawling — opaque as it can be to some — gestures at a level of objectivity, the crawling strategies for automated capture are programmed by people and are therefore far from neutral. As Ogden et al argue, there have been recent calls for studies into the "performative nature of crawlers and other Web archiving technologies," where technologies such as "crawlers (non-human, automated agents, bots, algorithms, code) are conceived as not merely passive or objective participants in the collection of Web resources, but are intricately implicated in the active shaping of the 'doing' of Web archiving" (Ogden et al., 2016, p.2). Ultimately, as we affix a critical eye on Web archiving actors, the documented lack of diversity among professional librarians (Vinopal, 2016) and archivists (Banks, 2006) — compounded by the severe homogeneity among information technology professions in general — should prompt us to consider the representativeness of Web archive collections.

With all of this in mind, Jacques Derrida's writing in *Archive Fever* seems equally fitting in the context of the digital: "There is no political power without control of the archive, if not of memory... Effective democratization can always be measured by this essential criterion: the participation in and access to the archive" (Derrida, 1995, p. 11). This statement is particularly striking in this context due to the increasing amount of information (official or otherwise) that exists solely in digital formats. Ultimately, it is essential that we consider Web archives and Web archiving practices in the same critical light as we do traditional archives and digital collections in the context of library practices. In many ways this chapter builds on Joan Shwartz and Terry Cook's seminal rallying call,

which argues that "[we] need to look anew at the archive in the light of changes in the production and preservation of documents, in the abundance of documents, in the changing media of record, and in the nature of what is documented or who is doing the documenting, as well as the need to examine the impact of these changes, in turn, on records management and its practices and on archives and its practices" (Schwartz and Cook, 2002, p. 5). To do so in the context of growing born digital documents is predicated on bolstering critical digital literacies in the technical and social processes that make Web archives possible.

## Why it's important to teach students critical digital literacy

As mentioned previously, scholars have identified widening gaps between data "haves and have-nots," and have argued that in the context of big data "knowing how to collect, find, analyze, and communicate with data is of increasing importance in society" (D'Ignazio, 2017, p. 6). However, it is simply not enough to introduce students to notions of the digital. A critical approach to digital literacy would have us critique the many ways we utilize, apply, and build digital tools and methods. In the context of Web archives, what may appear to be a neutral act of setting up Web crawlers to automate the process of capture, may actually involve a series of complex, sensitive decisions on the part of both the Web archivist and the institutional context in which they work. However, it is not enough to simply state this prescriptively and move on, especially given that students and researchers often do not understand the process of archiving the Web and are, in many ways, alienated from learning. By this I mean that they are alienated by the opaque technical process of Web archiving, as well as by the skills required to make sure of Web archival collections.

Holger Pötzsch's recent contribution to the field of critical digital literacy clearly delineates the important challenge and opportunity that we are facing as educators in a period of serious technological upheaval. According to Pötzsch, to sidestep these problems we must "combine critique, practice, and self-reflection … [this] brings to the forefront the various affordances and contexts of our most salient technologies and connects these to situated practices of use and appropriation" (Pötzsch, 2019, p. 224). Indeed, as Pötzsch writes, one key asset of digital technologies is the fact that "they undermined the gatekeeping function of traditional media channels and enabled audiences… to combine the roles of receiver and producer of media messages" (Pötzsch, 2019, p. 235) Ideally, "this key characteristic can be harnessed for educational purposes

by opening opportunities for student-driven active creation of expressions documenting, for instance, their own practices with and attitudes towards digital technologies" (Pötzsch, 2019, p. 235). The challenge, then, is to find meaningful and creative ways to combine critique, praxis, and self-reflection.

This should also be the case when specifically teaching about Web archives, and in general when discussing the archives-as-data paradigm. As Van House and Churchill write, "the facility with which material can be digitized, replicated and distributed … has resulted in profound shifts in how we conceptualize memory, our personal and collective archive practices, and even our views of persistence and permanence" (Van House & Churchill, 2008, p. 296). In this new facile environment exists a tacit "techno-centric belief in an infallible memory machine, in contrast to a notionally capricious, context-dependent and therefore fallible human memory" (Van House & Churchill, 2008, p. 296). As consumers and information users, "we may be seduced by the promise that we can accumulate and store everything with minimal cognitive effort and within the confines of a limited (physical) space" (Van House & Churchill, 2008, p. 296). Despite the feeling of technological infallibility our digital assets may be more fragile than we think — Web archives make that explicitly clear. Web archives take up space (lots of it, in fact), require labor and resourcing, and are embedded in complex socio-technical practices. It is in this space that librarians can affect change through critical pedagogy in practice.

## Teaching Web Archives and Digital Literacy

In what follows I offer not prescriptive answers to the problem of integrating Web archives in library-led curriculum, but a possible springboard into what could become fruitful areas of inquiry for both students and faculty. Web archives are approachable for users with experience in the technical elements of archival work, in that they provide those users with a referential entry point (i.e., I've used archives in the past, therefore this must be similar). But as Brügger writes, "[Web archives are] fundamentally different from those associated with other types of collections, because the process of collecting, preserving, and making the online Web available is more complex and opaque than is the case with other source types" (Brügger, 2017, p. 71). One's feeling of (or hope for) familiarity can be dashed by the inherent complexity found in the archives-as-data paradigm that must be adopted by scholars interested in using born

digital documents as primary sources. Therefore, knowledge about why and how the online Web needs to be archived is an important piece of information for any would-be user of Web archives. To overcome what Emily Maemura has identified as a Web archives bottleneck (2018), librarians can develop a set of pedagogical strategies that introduce new learners to the inherent and arguably unavoidable digitality of the future of research.

Web archives provide several opportunities for digital literacy instruction for librarians. One option is to concentrate on how such archives are created and why they are necessary in the context of today's research workflow. This perspective overlays with elements of the ACRL Information Literacy Framework, in particular the themes of constructed authority and the value of information, as well as other approaches to digital literacy that focus on the technical considerations of software and application development. This approach focuses on larger themes of archiving practices, including community engagement, preservation, and the importance of collective memory. A critical studies of technology approach asks us to consider the technical processes driving Web archives and ask questions about the choices and biases that are impacting how these archives are created, curated, and made available. Indeed, we must attempt to consider what we view as not valuable now that may be invaluable in the future, and "how we distinguish signal from noise in the grand bazaar of internet goods" (Van House & Churchill, 2008, p. 304).

An additional problem with digital memory is the issue of retrieval. The information may exist and be retrievable, but it will only be useful to us if we know that it exists, where it exists, and if we can get access to it. Therefore, an alternative pedagogical focus for librarians is to concentrate specifically on Web archive use in teaching and research. This approach favors utility and applicability in the context of primary source research.

Ideally, librarians would introduce both elements in a way that makes sense to them given their technical knowledge and resources. The following section provides strategies for both. Although the lesson plans I introduce below are neither exhaustive nor appropriate for all circumstances, they can assist librarians looking to introduce a novel perspective as they tackle teaching digital literacy. The value of these strategies is that they can be introduced in different settings and with different constituencies. For instance, they could be adopted for a specialized library workshop, or an instruction session in a range of disciplines, including history, sociology, and computer science. While these approaches can be remixed and

adopted for a variety of settings, I see the following approaches as best suited for one 80-minute class session.

## Critical Approaches for Web Archive Instruction

## Lesson 1: An Introduction to Web archives in an Age of Abundance[9]

### *Introduction*

The question of why Web archives are required for future research is an important methodological question that can be introduced in the greater context of information and digital literacy. In my experience, students and faculty are unfamiliar with the technical process of Web archiving as well as the need for Web archiving for the future of socio-cultural and historical research. However, students and researchers alike would benefit from considering the implications of the archives-as-data paradigm on their own research practices. To reconcile this, librarians should consider including Web archives as discussion and activity options during lessons or workshops on primary sources. A LibGuide featuring different Web archiving initiatives is an appropriate place to start, especially in the context of local or national history.[10]

The following lessons use critical pedagogical methods to enable students and instructors to learn about Web archives and question the potential impact of this new form of archiving in research and scholarship. I designed these lessons for undergraduate history classrooms, where digitized primary sources are increasingly popular. Having said that, I hope that other instruction librarians become inspired and apply these lessons in the context of other humanities and social science disciplines.

---

[9] I created these lesson plans while working as a Research, Instruction and Digital Humanities Librarian at the New College of Florida, a small liberal arts college in southwest Florida, USA.

[10] For an excellent example of a LibGuide dedicated to Web archives in the context of Government Data, see: https://libguides.uvic.ca/c.php?g=256600&p=2905190#s-lg-box-9377888

I designed the subsequent lessons with Juliet Hinrichsen and Antony Coombs' notion of "situated practice" in mind, such that emphasis is not only placed on skill acquisition, but also on source criticism and critical reflection. In summary, the following lessons attempt to explain *how* we preserve the Web, *why* that is an important practice, *who* decides which Web pages are worthy of preservation, and *what* opportunities exist for using Web archives in research and scholarship.

### *Setting*

Each lesson in this plan can be carried out in the context of one instruction session or broken up into consecutive sessions. This lesson is ideal for courses with a strong research methods component. I recommend at least 80 minutes for the entire plan.

### *Readings*

There are several articles that offer representative case studies for how Web archives can be used to influence our creation of cultural knowledge and shared memory. The article I chose as an example here, however, is especially useful in that it offers a readable and comprehensive treatment of Web archive creation and use, which also includes an interesting narrative. Additionally, this co-authored piece (written by a librarian/programmer, an archivist, and a historian) provides a perspective not only on the technical components of Web archives, but their use in contemporary historical research:

> Milligan, I., Ruest, N., & Onge, A. S. (2016). The great WARC adventure: Using SIPS, AIPS, and DIPS to document SLAPPs. *Digital Studies/Le champ numérique*.

The (open access) article details the circumstances surrounding a libel case that was filed against academic librarian Dale Askey by publisher Herbert Richardson, the resulting online debate and advocacy, and the effort by the authors to capture, preserve, and make available preserved websites related to the event. The article presents the technical aspects of capturing and preserving WARC files in a technical but accessible way. Furthermore, it reflects on some of the challenges of creating a traditional finding aid to contextualize and provide access to the collected electronic content. Finally, the article discusses some preliminary findings based on analysis of the data set by a professional historian.

### *Before class*

Students should come prepared having read the article and having watched a short 3-minute video from the Library of Congress on Web archiving.[11] Given that this is a somewhat technical article, I provide the discussion questions ahead of time (uploaded to the LMS along with the article) so that students can use them to guide their reading. In the past, I have also used Hypothesis[12], a Web annotation tool integrated in our LMS (Canvas), to encourage students to read the article and annotate it collaboratively. For students who do not read the article, a brief summary discussion at the beginning of the lesson would help situate the reading.

### *Discussion questions*

1. What do you think about the motive behind the #freedaleaskey collection given the perspective of the authors?
2. Could you explain the process of creating a surrogate of a Webpage after reading this article's technical description of Web archiving?
3. This was not an entirely straightforward process. What challenges did the authors face and how did they overcome these challenges, both technologically and socio-technically?
4. Beyond a description of how and why this Web archive was created, how did the addition of a historian's use of the Web archive influence or inform your understanding of the article?
5. If you could set up a Web archiving workflow, what website(s) would you want to preserve, and for what reason?

---

[11] Library of Congress Digital Preservation Video Series: Web Archiving: https://www.youtube.com/watch?v=T0943YkhLWU

Transcript is available: http://www.digitalpreservation.gov:8081/videos/docs/Webarchiving_video_transcript.pdf

[12] https://Web.hypothes.is/

***Group Activity***

The small irony of this example, which makes it even more powerful as an example for this lesson, is that the "#freedaleaskey collection" no longer exists in the Web accessible form it once did. The repository that once held its contents[13] is no longer available. The only way to get a sense of what the #freedaleaskey collection looked like is to use the Internet Archive's WayBack machine.

1. Break up the participants into groups and ask them to attempt to locate the #freedaleaskey collection on the Web. A search for the #freedaleaskey collection on a search engine like Google only retrieves secondary literature, or information about Dale Askey and the authors of the article cited previously. After several minutes, reveal to the students that what they have been searching for no longer exists, and discuss possible reasons behind it (human and technical resourcing being chief among them).

2. Direct students to the Internet Archive WayBack Machine (http://Web.archive.org) and have them search for the following url: http://freedaleaskey.plggta.org/. The WayBack machine is an open source software that renders WARCs in a Web browser. The resulting calendar view maps the number of times http://freedaleaskey.plggta.org/ was crawled.

3. Students will see that twenty-four snapshots of the website exist between August 13, 2013 and March 23, 2018. Ask students to navigate through the snapshots to gain an understanding of how the WayBack Machine works, as well as how the repository operated when it was live. Under the "Letters of Support" tab, for instance, point out that thumbnails (as described in detail in the article) are included; however, upon clicking on the links the user is redirected to an error. This means that while the links were crawled, the endpoint was not. As Winters writes, "A missing image confronts us with a blank square on the Web page; a broken link produces an error" (Winters, 2017, p. 245). This inconsistency

[13] http://freedaleaskey.plggta.org/

reflects not only the thoroughness of a crawl, but also represents a seemingly lost digital object.

4. Consider discussing how the website has changed sometime between March 14th, 2014 and August 17th, 2014, from a traditional repository view with document navigation to a simple rendering of a WordCloud.

5. Finally, emphasize the fact that Web archiving is a unique form of archiving that shifts the focus on capturing versions of webpages over time, rather than archiving and preserving one document through time.

## Lesson 2: Web archives and Source Criticism

### *Introduction*

In the first lesson, students are introduced to Web archives through a specific example, the #freedaleaskey collection. The article ends with a short description about how Web archives like this one can be used for historical research. Yet, this type of primary source analysis necessitates additional strategies that may not be familiar to students or scholars. Others have identified the need for 'source criticism' when working with Web archives, and the challenge of determining reliability or veracity of materials (Nanni, 2017; Vlassenroot et al., 2019). As Maemura makes clear, critically examining Web archives leans on many of the same skills required in other fields, such as Bibliography and History (Maemura, 2018). However, opportunities for unique forms of analysis present themselves in this new context. For instance, one approach to source criticism is through analyzing inconsistencies within the material, and how this might impact its reliability as a source (the "#freedaleaskey collection" Web archive mentioned above, for instance, would not fare well under this scrutiny). Another approach includes comparing multiple sources to each other, such as an archived version of a website with the current or "live" page. Finally, a third approach focuses on the process of Web archiving. Certain questions about the composition of the archived data, such as determining the provenance of the archived data, can only be answered by examining the metadata included in the WARC file, and analyzing the details—through specific documentation or general guidelines—of how the crawl

was conducted (Maemura, 2018). The second lesson focuses on applying critical source analysis methods to born digital Web archives.

***Lesson***

This lesson focuses on introducing students to Library of Congress collections and identifying potential instances in which Web archives can prove useful for research. To teach this lesson, I chose the Library of Congress' collections because they provide an interface for searching through Web archive records[14] (as well as a search interface for their curated collections[15]). These interfaces should seem familiar to students that have used library resources in the past. LoC collections vary in size, from small collections featuring a few dozen websites to collections that include thousands of individual pages. The variation in size speaks to the strategy that guides the collections' development.

1. Introduce the Library of Congress Web archive collections, point out the variety in size and subject matter, and briefly describe their collection development policies.

2. Explore two of the largest collections in the catalog: the September 11, 2001 Web Archive, a collection that "preserves the Web expressions of individuals, groups, the press and institutions in the United States and from around the world in the aftermath of the attacks in the United States on September 11, 2001," and the US Elections Web Archive, which includes the "campaign sites archived weekly during the election seasons since 2000, documenting sites associated with presidential, congressional, and gubernatorial elections." Beyond providing a metaphorical portal into the past, which allows us to explore these websites almost as if we had stepped back in time, both of these collections are ripe for applying the source criticism strategies discussed previously.

[14] https://www.loc.gov/websites/
[15] https://www.loc.gov/websites/colletions/?st=gallery

### *Group Activity*

1. Split the class into four groups (A-D) and assign each group one of the two collections (alternatively, ask each group to pick a unique collection).
2. Ask that group A compare how the Web pages (content) of federal agencies changed on September 11, 2001 to reflect what had happened in New York and Washington, D.C.
3. Ask that group B analyze the 'depth' of the Library of Congress' crawl, noting, for example, how many Web pages included missing images, broken links, and other errors.
4. Ask that group C study the US Elections Web Archive to examine the websites of Republican or Democratic candidates in a specific district chosen by students, paying close attention, for instance, to how graphics and imagery were utilized to denote certain elements of their political platform.
5. Finally, ask that group D perform a similar study to group B, by analyzing the 'depth' of the crawl for Republican or Democratic candidates of their choice.

### *Final Discussion*

The final discussion can be a flexible group discussion on the issue that students encountered in their work. Guide students to make connections between the importance of preserving the Web (citing the #freedaleaskey collection), the technical components of Web archives (*versions* instead of *fixed* analog documents), as well as the subjective decisions that go into Web archive collections (variability in content, size, and quality in the Library of Congress collections).

### *Reflection*

As an instruction librarian who is also responsible for introducing information technologies in research and in the classroom, I often struggle with finding the right balance between skill acquisition and critical reflection. I've found that this lesson strikes that balance. In addition to being relatively demanding for students with little knowledge of this

technology, it also encourages critical reflection by challenging students' perceptions of what it means to be a historian today.

I recognize that this lesson is slightly off the beaten path, and that there are many factors in a one-shot session that makes this type of instruction difficult. At the time of developing this lesson plan, I was fortunate to work at a small liberal arts college that strived to keep class sizes small, making this type of instruction more manageable. But even so, there was limited time for class activities and limited time to focus on students to the degree necessary for a really deep discussion. Balancing technical and critical elements is also always a challenge. In fact, the first iteration of this lesson felt like a complete failure, with students disinterested as a result of poor pacing and unclear directions on my behalf. Subsequent iterations have been positive, and in some cases, students have followed up asking for additional information about Web archives, including how to create personal Web archives.

## Conclusion

Canadian historian Ian Milligan offers perhaps the most convincing arguments for why additional attention needs to be directed to increasing Web archive competencies. "Imagine a history of 2019 that draws primarily on print newspapers," he writes, "approaching this period as 'business as usual,' ignoring the communication technologies that fundamentally affected how people share, interact, and leave historical traces behind." No websites, no blogs, no Twitter. He continues,

> we need to be knowledgeable of [Web archives] functionalities, strengths, and weaknesses: we need to begin to theorize and educate ourselves about them, just as historians have been cognizant of analog archives since the cultural turn. The challenge is considerable, but the potential is even greater (Milligan, 2019, p. 28).

It is worth noting that Milligan is not simply calling for increased technical competencies: his is a call for a comprehensive engagement with the changing nature of research in a digital world, one that considers the technical as well as ethical elements involved in Web preservation. In short, Milligan is pointing to "a situated practice" that emphasizes technical skills as well as an understanding of the social and cultural implications of this technology.

In this chapter I've argued that librarians have an opportunity to reconcile this knowledge and skills gap by introducing Web archives to students and researchers. The approaches described here serve as an introduction to the field and provides opportunities for critical pedagogy practices. As educators, our objective should not be to transform researchers used to analog research into "Web researchers." Rather, as Winters writes, "For most humanities scholars it will be a very long time before they transition to using solely digital sources, let alone solely born-digital sources, and for many, this will never be the case… Their research, however, will be impoverished if they are unaware of what Web archives may contain – even if it is only to discount that information as unhelpful or unreliable" (Winters 2017, p. 239). Our objective, then, should be to "to equip them to use Web archives, and to encourage others to do the same" in hopes of combining new and old approaches to solving historical and socio-cultural problems. This is one element of a concerted effort to develop creative, meaningful strategies for critical digital literacies. The lessons offered in this chapter offer skill development as well as contextualizing the impact of these technologies on the way that we create and preserve human records; in that way, these lessons satisfy Juliet Hinrichsen and Antony Coombs' notion of "situated practice" in order to "rebalance the emphasis on … operational skill with a focus on the practices and intellectual traditions of the disciplines both as meaningful sites of learning and as a reflection of a shift towards critical academic literacies" (Hinrichsen and Coombs, 2013, p. 4). Ultimately, alongside these new digital competencies exist rather old lines of questioning. Questions about provenance, authority, bias, and subjectivity parallel and sometimes intersect questions regarding technical skills and resourcing. We should take heart knowing that these intersections, as difficult as they are to reconcile at times, are intersections that librarians are faced with on a daily basis. Librarians are therefore in a good position to help students and faculty researchers with this transition.

## References

Agata, T., Miyata, Y., Ishita, E., Ikeuchi, A., & Ueda, S. (2014, September). Life span of Web pages: A survey of 10 million pages collected in 2001. In *IEEE/ACM Joint Conference on Digital Libraries* (pp. 463-464).

Banks, B. (2006). Part 6: A* CENSUS: Report on Diversity. *American Archivist*, *69*(2), 396-406.

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.

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

Brügger, N., & Schroeder, R. (Eds.). (2017). *The Web as History: Using Web Archives to Understand the Past and the Present.* UCL Press.

D'Ignazio, C. (2017). Creative data literacy. *Information Design Journal*, 23(1), 6-18.

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. *Journal of the Association for Information Science and Technology*, *65(11),* 2195-2209.

Hinrichsen, J., & Coombs, A. (2014). The five resources of critical digital literacy: a framework for curriculum integration. *Research in Learning Technology, 21.*

ISO. Information and documentation - WARC file format, 2009.

Koltay, T. (2017). Data literacy for researchers and data librarians. *Journal of Librarianship and Information Science, 49(1),* 3-14.

Lomborg, S. (2019). Ethical Considerations for Web Archives and Web History Research. In Neils Brugger and Ian Milligan (Eds)., *The Sage Hangbook of Web History,* 99-111, Sage.

Maemura, E. (2018). What's cached is prologue: Reviewing recent Web archives research towards supporting scholarly use. *Proceedings of the Association for Information Science and Technology*, *55(1),* 327-336.

Maemura, E., Worby, N., Milligan, I., & Becker, C. (2018). If these crawls could talk: Studying and documenting Web archives provenance. *Journal of the Association for Information Science and Technology*, *69(10),* 1223-1233.

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

Mordell, D. (2019). Critical Questions for Archives as (Big) Data. *Archivaria*, *87(87),* 140-161.

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

Ogden, J., Halford, S., & Carr, L. (2017, June). Observing Web archives: The case for an ethnographic study of Web archiving. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 299-308).

Pötzsch, H. (2019). Critical Digital Literacy: Technology in Education Beyond Issues of User Competence and Labour-Market Qualifications. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society, 17(2),* 221-240.

Prado, J. C., & Marzal, M. Á. (2013). Incorporating data literacy into information literacy programs: Core competencies and contents. *Libri, 63(2),* 123-134.

Schwartz, J. M., & Cook, T. (2002). Archives, records, and power: The making of modern memory. *Archival Science, 2(1-2),* 1-19.

Shorish, Y. (2015). Data Information Literacy and Undergraduates: A Critical Competency. *College & Undergraduate Libraries, 22(1),* 97-106.

Tygel, A. F., & Kirsch, R. (2016). Contributions of Paulo Freire for a Critical Data Literacy: A Popular Education Approach. *The Journal of Community Informatics, 12(3).*

Van House, N., & Churchill, E. F. (2008). Technologies of memory: Key issues and critical perspectives. *Memory Studies, 1(3),* 295-310.

Vinopal, J. (2016). The quest for diversity in library staffing: from awareness to action. *In the Library with a Lead Pipe.*

Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web Archives as a Data Resource for Digital Scholars. *International Journal of Digital Humanities, 1(1),* 1-27.

Winters, J. (2017). Coda: Web archives for humanities research–some reflections. In Niels Brügger and Ian Milligan (Eds.), *The Web as History: Using Web Archives to Understand the Past and the Present*, 238-248, UCL Press.