Since the Web's conception, it has been a technology that connects people across borders. Ed Summers contributes a review essay on two volumes about web archiving: *The Web as History: Using Web Archives to Understand the Past and the Present*, edited by Niels Brügger and Ralph Schroeder; and *History in the Age of Abundance?*: *How the Web Is Transforming Historical Research,* by Ian Milligan. Both publications interrogate the Web as a site for historical research, though Summers takes this opportunity to ask what it means to archive the Web and how archivists can attune their practices to meet emerging research needs that capture the Web as not only WARC files, but also as a sociotechnical construction.

American Archivist is the Society of American Archivists's premier journal, but it is—and should be—a journal for all archivists, regardless of country of origin. And to be an inclusive journal, it requires voices from beyond the Global North. I hope to continue including more reviews that represent the breadth of the international archival literature and connect us all together across boundaries, divides, and borders.

Note

[1] Herman Kahn and Harold Larson, "Reviews of Books," *American Archivist* 2, no. 1 (1939), 46–68, https://doi.org/10.17723/aarc.2.1.c203372787715898.

Review Essay: What We Talk about When We Talk about Archiving the Web

The Web as History: Using Web Archives to Understand the Past and the Present.
Edited by Niels Brügger and Ralph Schroeder. London: UCL Press, 2017. 296 pp.
Softcover, EPUB, and Open Access PDF. Softcover £22.99UK.
Softcover ISBN 978-1-9113-0755-6; EPUB ISBN 978-1-9113-0759-4;
PDF ISBN 978-1-9113-0756-3. Freely available at
https://ucldigitalpress.co.uk/Book/Article/45/70/0/ and
https://discovery.ucl.ac.uk/id/eprint/1542998/.

History in the Age of Abundance?: How the Web Is Transforming Historical Research.
By Ian Milligan. Montreal: McGill-Queen's University Press, 2019. 328 pp.
Softcover and EPUB. $32.95CAD. Softcover ISBN 978-0-7735-5697-3;
EPUB ISBN 978-0-7735-5822-9.

What is the Web? Is it the collection of standards, such as hypertext markup language (HTML), hypertext transfer protocol (HTTP), and uniform resource identifiers (URI) that have evolved for the past three decades?

Is it various pieces of software such as servers in the cloud, the browser on a laptop, and the apps on a smartphone? Is it all the types of content (text, video, audio) that get linked together and made interactive with JavaScript? Is it the businesses, governments, organizations, and collectives that render behavioral norms, economies of scale, and laws that shape the global distribution of information? Of course, the answer to all these questions (and more) is yes. So, what could it possibly mean to archive the Web?

Understanding what *archiving the Web* means in all these contexts requires archivists to not only to ask what the Web *is* in terms of records, but also to consider how a web archives functions and (perhaps most important) how it is used. Put more abstractly, understanding web archives is as much a question of sociotechnical practice as it is a question of what constitutes the records that comprise web archives.

Two recently published books—one by Ian Milligan (2019) and one edited by Niels Brügger and Ralph Schroeder (2017)—provide essential guides to help answer the question of what web archives are by describing concrete, nonhypothetical examples of how social science and humanities researchers are using web archives today. For those who have participated in web archiving activity and pondered how the records would get used, and for those who are looking to get involved in web archiving but are not sure what it takes, these two books are essential reading.

Even though one volume is a collection of essays and the other a monograph, considering these titles together is useful because they share much in common. Both books are largely targeted at the academic research community, with the goal of broadening awareness of the research potential of web archives, while also providing methodological examples of how to conduct research with them. It is no accident that the word "history" figures prominently in each of the titles, *The Web as History* and *History in the Age of Abundance?*, as both books have a pronounced interest in the historical use of the Web. Milligan and Brügger both serve as founding editors of the journal *Internet Histories* that publishes "social, political and technological histories of the internet."[1] It is also worth noting that *The Web as History* is coedited by Ralph Schroeder who, as a social scientist at the Oxford Internet Institute, brings a social science flavor to this collection of essays.

The Web Historical Shunt

Given their historical bent, it is instructive to recall the debate within archival studies about the role of the historian-archivist. Put simply, the concerns of history and archives often, but not always, completely align. An archives presents historians with evidence of the past that is crucial for their work. But archives are not assembled solely to provide primary sources for historical

research. Archives are a set of information practices that get deployed in particular settings to achieve specific instrumental goals. This deployment, and the evidentiary traces archives leave behind, confers historical value on the records.

The professionalization of archives in the United States was achieved in no small part by Margaret Cross Norton, who distinguished the archivist as an expert in the processes of documentation, rather than being only a caretaker of history.[2] Hugh Taylor memorably warned that archivists needed to avoid the "historical shunt" to remain relevant as a profession, especially as archives increasingly became sites for automation during the mid- to late twentieth century:

> . . . . we must be prepared to abandon the concept of archives as bodies of "historical" records over against so-called active records which are put to sleep during their dormant years prior to salvation or extinction. Records are active in direct proportion to the relevant information that can be retrieved from them, and dormancy is closely related to the inability to retrieve information.[3]

I mention all this here not to disparage the historical treatment of web archives that these two books offer, but rather to draw attention to how the two books actually do something more than simply describe how web archives can be used in research. While both volumes provide excellent examples of the types of historical and social science scholarship that is possible with web archives today, significant strands in each book speak to what we conceive web archives to be. These themes concern the *ontology* of web archives, or how web archives are themselves social and technical constructions that have historical specificity. Both books contain latent (and explicit) arguments about what web archives are and are not. These arguments amply describe the current state of web archives, and archives more generally, and suggest some promising areas of future research for web archives in archival studies.

Web Archives as Data

One recurring theme that these books illustrate is the prevailing idea that web archives are collections of records extracted from the Web and then placed into spaces as data to be used by researchers. Indeed, this conception of web archives flows naturally from traditional ideas of archives as custodial spaces where inactive records go for long-term preservation and use.

Take, for example, the JISC UK Web Domain Dataset,[4] which is used as the basis for several chapters in *The Web as History*. The JISC data set is a collection of web content crawled by the Internet Archive from web domains ending in .uk between 1996 and 2013. The data were transferred to the British Library in

two separate tranches totaling 28,554 files using the WARC (Web ARChive) file format[5] and its predecessor, ARC (ARChive). Several studies in *The Web as History* put the JISC data to use: to analyze the growth of UK academic websites (Eric T. Meyer et al., "Analysing the UK Web Domain and Exploring 15 years of UK Universities on the Web"); to measure the geographic coverage of the BBC's content using the external links from its website (Josh Cowls and Jonathan Bright, "International Hyperlinks in Online News Media"); to examine the coverage of the Internet Archive's own web crawlers (Scott A. Hale et al., "Live versus Archive: Comparing a Web Archive to a Population of Web Pages"); and to explore the use of web archives data in arts and humanities research (Josh Cowls, "Cultures of the UK Web").

One interesting aspect of the JISC data set is its provenance. It was initially collected by the Internet Archive using a variety of sources that are now somewhat obscured:

> The Internet Archive (IA) web collection comes from crawls run by the IA Web Group for different archiving partners, the Web Wide crawls and other miscellaneous crawls run by IA, as well as data donations from Alexa and other companies or institutions. IA is not able to share the names of these companies, but can state that they include a few vertical search engines, and some other Google-like companies.[6]

The Joint Information Systems Committee (JISC), now Jisc, is a UK nonprofit that "commissioned" the Internet Archive to donate the .uk web crawl data, which was then housed at the British Library. The data complemented the UK Web Archive with historical data, which helped it bootstrap the infrastructure needed to support the UK's legal deposit web archiving program. Interestingly, not many of the studies in *The Web as History* draw on the actual WARC data; Scott A. Hale's "Live versus Archive" is one notable exception. Instead, the studies use derived data, such as the separately available "host link graph data," which details the source and target of hyperlinks in the WARC data and can be accessed via the Web.[7] This chapter is also a notable example of how analyzing the representativeness of coverage of a web archives is essential for social science research where validity, reliability, and generalizability are a central concern.

The size of the full JISC data set is approximately twenty-seven terabytes, which means it is difficult to make available on the Web. But the data set is further encumbered by legal restrictions (2017, p. 28) that prevent it from being used outside the British Library without permission.[8] This issue of access to WARC data is in fact quite a complex one. For example, the Internet Archive, which aims to provide "universal access to all knowledge," does not make its underlying WARC data available to the public. But the Internet Archive has been known to grant access to individuals for research.

Reading Web Archives

One of the most significant contributions of Milligan's *History in the Age of Abundance?* is that it provides a highly accessible history of how web archives have come to be in their present shapes. His description is just as relevant for the archivist as it is for the historian or social scientist. For example, he devotes an entire chapter to debate around the term "web archive," which centers on the difference between an archives and a collection, and the importance of provenance and original order to understanding what an archives is. Milligan cites none other than Brügger to make the case that web archives are the "deliberate and purposive preserving of web material,"[9] but concedes that "Web archives are not traditional archives—not in content, form, or conception" (2019, p. 72). He describes the contested terrain around the term "web archives" by situating it in historical context and essentially makes a pragmatic case for the term "web archives," which is not entirely consistent with archival theory, but does describe the practice of "web archiving" that has emerged over the last twenty years.

The description of web archiving practice in *History in the Age of Abundance?* details the work of the Internet Archive, the national libraries that make up the International Internet Preservation Consortium (IIPC), the libraries and archives that subscribe to the Archive-It service, and even volunteer organizations like Archive Team. One common thread running through these chapters is the central importance of WARC data: understanding how the data are collected using crawlers like Heritrix; how they are made accessible or viewable using tools like the Wayback Machine; and how they are analyzed as data using digital methods such as network analysis and topic modeling.

As the primary investigator on the Archives Unleashed Project, Milligan has spent a significant amount of effort over the past five years "developing web archive search and data analysis tools to enable scholars, librarians and archivists to access, share, and investigate recent history since the early days of the World Wide Web."[10] I attended two of the Archives Unleashed workshops and was struck by how the they focused on working with web archives as data, specifically WARC data.

History in the Age of Abundance? can be read like a missing textbook for the Archives Unleashed workshops, providing background material for what the Web is, why it is significant for historians, how archivists create web archives, and the research methods available for analyzing (or reading) web archives. However, unlike the documentation provided during the workshops, *History in the Age of Abundance?* contains very few examples of actual code to use for analysis. This was done for practical reasons because the tools themselves are bound to change: "Historians will not all become programmers. Rather, they must be able to implement—with understanding—algorithms designed by others" (2019,

p. 155). Coupled with the workshops, Milligan's volume provides a comprehensive picture of the current state of web archives.

Access to WARC data is central to the analyses provided in both of these books. To apply the distant reading[11] or statistical techniques the books describe, a researcher will need to have access to the WARC data that are the result of "archiving" some portion of the Web. Consequently, it is curious to note that institutional archives that perform web archiving do not typically have procedures for making WARC data available, either remotely through the Web, or locally for researchers who are able to travel to the repository. Instead, they use an instance of the Wayback Machine (either their own, or the one running at the Internet Archive) to access item-level views of web documents at a particular URL at a particular time. Web archives also lack the type of description needed for researchers to fully contextualize what was (and was not) collected, and how.[12] Both *The Web as History* and *History in the Age of Abundance?* make an implicit argument that archives need to move beyond simply allowing researchers to view what a webpage looked like, to providing services that make the underlying WARC data available for analysis. Perhaps efforts such as the recently funded project at the Library of Congress to explore infrastructure for digital research (Milligan is on its advisory board) will establish some guidance for how a digital equivalent to the reading room can work in practice.[13]

Web Archives as Infrastructure

But the focus on using WARC data and tools really tells only one particular story of web archives, one that is suitable for historians using some of the web archives that are currently available. As noted earlier, archives are not only the historical records left behind, they are the sociotechnical systems used to create and manage what Hugh Taylor called "active records."[14] Indeed, in more recent work, archival theories such as the records continuum model[15] recognize the value of understanding the full scope of human interactions and relationships that records participate in—that includes, but is not limited to, their use in history.

Both of these books contain latent hints of this larger perspective, particularly when they discuss the pivotal role that the Domain Name System (DNS) plays in research with web archives. For example, the management of a country code top-level-domain (ccTLD) is delegated by the global domain name registrar ICANN (Internet Corporation for Assigned Names and Numbers), to a regional registrar such as Nominet in the UK, DK Hostmaster in Denmark, and AFNIC in France. These registrars handle the Internet's address system within each of the two-letter suffixes for countries and territories, such as .uk, .dk, or .fr. Because the lists of ccTLD domain names provided by these organizations constitute a

comprehensive inventory of all the web domains within the national domain, it is relevant to include them in any study of the development of a national Web because they delineate the outer limits of the national domain name space and they attest to the development of the national web domain over time. The domain name list itself can help to answer research questions regarding, for instance, the number of domain names per year, the number (and names) of domain names that have disappeared or been added since last year, and the number of domain names per domain name owner (Niels Brügger et al., "Exploring the Domain Names of the Danish Web," p. 65).

The significance of these DNS registrars to archives cannot be overstated. DNS provides a juridical view of what constitutes a nation's Web, which (as highlighted in both books) is essential to the functioning of web archiving programs in countries that have legal deposit programs that include web content. But DNS also provides critical infrastructure for recording the transactions of domain ownership (e.g., google.com or bl.uk), without which the day-to-day operation of the Web would be impossible.

When we consider the Web as an archival information system, DNS functions much like the registries, lists, and indexes that have supported more traditional, paper-based forms of archives. As archival studies practitioners and scholars, we must recognize that the administrative and maintenance work that supports a service like DNS is itself a form of records management. This archival view of DNS is in fact just one of many ways to look at and study the Web as an archival system. For example, we could also study the ways in which websites are maintained over time using content management systems that must relay records forward through time. Or, we could examine the algorithms used to both collect content from the Web and make it available. While some may consider these topics outside the scope of web archives, it is important that the scope of studies related to archives and the Web not be artificially limited to today's particular stack of technologies and standards. It is also important to see the Web as a branch in a genealogy of media systems—not as an aberrant break with the past that requires all theory to be thrown out the window.

Conclusion

Of course, the topic of web archiving has been no stranger to the pages of *American Archivist*. Examples abound, such as Timothy Arnold and Walker Sampson's collection development practices for topical social media archives;[16] Brewster Kahle's call for "universal access to all knowledge" in the creation of the Internet Archive;[17] Steven Lubar's analysis of the benefits of hypermedia for archival context;[18] and Margaret Hedstrom's framework for research in electronic records that foreshadowed much of the research to come, right at

the dawn of the Web.[19] I highlight these here simply to note the diversity and duration of interest that has come from the journal you are reading now and to invite more to come. Archival studies researchers must recognize the full scope of archival functions that exist on the Web, rather than being artificially limited to their current infrastructural form. For a broader perspective on the topic of web archives from the field of archival studies, I recommend Emily Maemura's bibliography,[20] as well as the resources made available by the Web Archiving Section of the Society of American Archivists.[21] However, it bears repeating that these two books are essential reading both for understanding how historians would like to use the web archives we have been assembling and for hinting at how archival theory and practice can engage with a much richer conception of what archiving the Web means.

© Ed Summers
University of Maryland

Notes

1 "Aims and Scope," *Internet Histories*, https://www.tandfonline.com/action/journalInformation?show=aimsScope&journalCode=rint20, captured at https://perma.cc/SGP6-5DWK.

2 Randall Jimerson, "Margaret C. Norton Reconsidered," *Archival Issues* 26, no. 1 (2001): 41–62, http://digital.library.wisc.edu/1793/45982.

3 Hugh Taylor, "Information Ecology and the Archives of the 1980s," *Archivaria* 18 (1984): 25–37, https://archivaria.ca/index.php/archivaria/article/view/11075/12011, captured at https://perma.cc/UG2V-BA7H.

4 JISC and the Internet Archive, "JISC UK Web Domain Dataset (1996–2013)" (2013), https://doi.org/10.5259/ukwa.ds.2/1.

5 WARC Specifications, "The WARC Format 1.1" (International Organization for Standardization, 2017), https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1.

6 JISC and the Internet Archive.

7 Andrew Jackson, "JISC UK Web Domain Dataset (1996–2010) Host Link Graph" British Library Research Repository (2013), https://doi.org/10.5259/ukwa.ds.2/host.linkage/1.

8 Andrew Jackson, personal communication, January 6, 2020.

9 Niels Brügger, "Website History and the Website as an Object of Study," *New Media & Society* 11, nos. 1–2 (2009): 115–32, https://doi.org/10.1177/1461444808099574.

10 The Archives Unleashed Project, "Welcome to the Archives Unleashed Project," https://archivesunleashed.org, captured at https://perma.cc/65FL-BYEW.

11 Ted Underwood, "A Genealogy of Distant Reading," *Digital Humanities Quarterly* 11, no. 2 (2017), http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html, captured at https://perma.cc/46AZ-X352.

12 Emily Maemura et al., "If These Crawls Could Talk: Studying and Documenting Web Archives Provenance," *Journal of the Association for Information Science and Technology* 69, no. 10 (2018): 1223–33, http://hdl.handle.net/1807/82840.

13 Library of Congress, "Library Receives $1M Mellon Grant to Experiment with Digital Collections as Big Data," press release, October 4, 2019, https://www.loc.gov/item/prn-19-098, captured at https://perma.cc/RMZ8-WVFJ.

14 Taylor, "Information Ecology and the Archives of the 1980s," 30.

15 Sue McKemmish, Frank Upward, and Barbara Reed, "Records Continuum Model," in *Encyclopedia of Library and Information Sciences*, ed. Marcia Bates and Mary Niles Maack (Taylor & Francis, 2010).

[16] Timothy Arnold and Walker Sampson, "Preserving the Voices of Revolution: Examining the Creation and Preservation of a Subject-Centered Collection of Tweets from the Eighteen Days in Egypt," *American Archivist* 77, no. 2 (2014): 510–33, https://doi.org/10.17723/aarc.77.2.794404552m67024n.

[17] Brewster Kahle, "Universal Access to All Knowledge," *American Archivist* 70, no. 1 (2007): 23–31, https://doi.org/10.17723/aarc.70.1.u114006770252845.

[18] Steven Lubar, "Information Culture and the Archival Record," *American Archivist* 62, no. 1 (1999): 10–22, https://doi.org/10.17723/aarc.62.1.30x5657gu1w44630.

[19] Margaret Hedstrom, "Understanding Electronic Incunabula: A Framework for Research on Electronic Records," *American Archivist* 54, no. 3 (1991): 334–54, https://doi.org/10.17723/aarc.54.3.125253r60389r011.

[20] Emily Maemura, "Web Archives Bibliography" (2019), https://github.com/emilymae/web-archives-bib#readme.

[21] Society of American Archivists, "Web Archiving Section" (2019), https://www2.archivists.org/groups/web-archiving-section, captured at https://perma.cc/75YS-KNEA.

Teoria i praktyka archiwistyki USA

By Bartosz Nowożycki. Warszawa: Naczelna Dyrekcja Archiwów Państwowych, 2017. 285 pp. Softcover, Open Access PDF, EPUB, and Mobi. 20 zł PLN.
Softcover ISBN 978-83-65681-15-7; PDF ISBN 978-83-65681-16-4;
EPUB ISBN 978-83-65681-17-1; Mobi ISBN 978-83-65681-18-8.
Freely available at https://perma.cc/LW28-PG6L.

The archives of Europe have a centuries-long history that serves as a starting point for understanding and developing manuscript collections and records repositories in North America. Contemporary archival theory and practice across the United States owe their roots to these traditions, which are well known and intertwined with US archives history. Across the Atlantic, and south of the Baltic Sea more specifically, however, the story of American archives is not as well known. In an effort to resolve this, Bartosz Nowożycki explores the issue in great detail in his Polish-language monograph, *Teoria i praktyka archiwistyki USA*, or *Archival Theory and Practice in the USA*.

A historian and archivist, and a senior specialist at the State Archives of Poland, Nowożycki had the opportunity to explore the history, laws, theory, and practices of American archives while serving as a visiting archivist at two prominent Polonian institutions in New York City. As a result of a comprehensive review of a variety of published sources, his publication focuses almost exclusively on the National Archives and Records Administration (NARA). Nowożycki has written a thoroughly researched and heavily cited chronological review of NARA's history from its inception in 1934 to 2009 when David S. Ferriero was appointed archivist of the United States, the laws governing its decisions and