You shouldn't Need to be a Web Historian to Use Web Archives

Ian Milligan

WARCNET PAPERS

WARCnet
web archive studies

You shouldn't Need to be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure

Ian Milligan

Associate professor of history
The University of Waterloo
i2milligan@uwaterloo.ca

WARCnet Papers
Aarhus, Denmark 2020

WARCnet Papers ISSN 2597-0615.

Ian Milligan: You shouldn't Need to be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure

Published by the research network WARCnet, Aarhus, 2020.
Editors of WARCnet Papers: Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Sofie Flensburg, Peter Webster, Michael Kurzmeier.

This publication has gone through single blind review.

Cover design: Julie Brøndum
ISBN: 978-87-972198-1-2

WARCnet
Department of Media and Journlism Studies
School of Communication and Culture
Aarhus University
Helsingforsgade 14
8200 Aarhus N
Denmark
warcnet.eu

The WARCnet network is funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

DANMARKS FRIE FORSKNINGSFOND
INDEPENDENT RESEARCH FUND DENMARK

WARCnet Papers

Niels Brügger: *Welcome to WARCnet* (2020)

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

You shouldn't Need to be a Web Historian to Use Web Archives: Lowering Barriers to Access Through Community and Infrastructure

Ian Milligan

If a researcher currently wants to use web archives at scale, they need to be a significant focus of their research activity – in other words, a web historian really can't just dabble with these sources. Yet historians need to be able to draw on wide varieties of sources for their projects (from archival records to newspaper records to oral histories and beyond). In this paper, I explore how we are developing tools to lower these barriers so that a historian could use web history data without a significant investment of time. The paper begins by discussing the current situation of working with web archives, before highlighting the Archives Unleashed project and the approaches we have taken to tackle these questions.

Keywords: Web archives, Historical Research, Historical Methods, Digital Humanities, Research Interfaces

What will we call a historian studying an event like the COVID-19 pandemic in ten, twenty, or thirty years? Will we call them a "web historian" – or just a historian? In our recent SAGE Handbook of Web History, Niels Brügger and I advanced an expansive definition of "web history." We used that term to refer not only to histories of webpages or the Internet, but also to any historian who might happen to use web archives in their work. As we explained, "the Web is both a historical source and an object of study in its own right" (Brügger & Milligan, 2018). I have argued elsewhere that web archives lie at the future of the historical profession, as any historian wishing to do justice to topics in the time period following the mid-1990s advent of web archives will almost inevitably need to access and use web-based primary sources (Milligan, 2019).

There are considerable barriers to using web archives at scale. Right now, if a historian wants to use web archives for anything beyond source replay, they need to invest in a considerable amount of skills training, secure computational resources beyond the field norm, and crucially, will need to look towards communities of practice like WARCnet or the Archives Unleashed project. In this essay, I argue that a major focus of the community

needs to be on reducing these barriers to use, so that more – perhaps not all – historians are able to avail themselves of web archive analysis at scale. We can do so by lowering barriers to access primarily through the development of accessible infrastructure which can in turn make web archived content more accessible to researchers. I begin the article by first introducing the problem with web archives, before discussing how we can rise to this challenge through skills and new infrastructure. I then turn to the Archives Unleashed project and how three personas have informed our work, before making brief conclusions around community development.

THE PROBLEM WITH WEB ARCHIVES

To be provocative and rely on somewhat of a strawman argument: the problem with web archives is not that we do not collect enough information. This can seem a bit foreign to a historian, as we tend to want our archives to be always more comprehensive and complete, as we stumble upon frustrating gaps and omissions in records. As of writing in May 2020, the Internet Archive has over 900 billion URLs and 60 petabytes (one petabyte is a thousand terabytes) of unique data; other institutions (primarily national libraries around the world) probably have about the same over again in their own holdings. While there are gaps in collection – especially when it comes to user-generated content that can be overlooked by algorithms – this is by any measure a dramatic amount of information.[1]

This is not to be overly dismissive of the challenges that lie on the capture side of the equation. There remain serious problems with harvesting, and we can understand web archiving as a "cat-and-mouse game": web developers invent something new, so web archivists have to catch up. For example, the "infinite scroll" of websites requires a user to scroll down to actually load the content, which has necessitated the development of simulated user interaction by the web crawler to capture content. However, we can understand the broad strokes of collection and capture as more or less solved. We generally understand how a researcher, curator, or librarian could scope out a collection (for example, "websites germane to COVID-19 in the Canadian province of Ontario") and use tools such as Archive-It, Heritrix, or beyond to collect it.

The problem that we as a web archiving community face is the problem of analysis. It is not that we do not collect enough data, but rather that our users and collectors do not always know both what exact data we have (it is difficult to know exactly what various web archives have collected) as well as what to do with the collected data. We have collected over a hundred petabytes of unique information, but what happens when it comes to analyze it? What should we do with the data that we have expensively curated and, even more expensive, committed to preserving in perpetuity? What happens if a researcher has a question? How can they answer it?

1. We discuss some collection models in Ian Milligan, Nick Ruest, and Jimmy Lin, "Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses," in Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL '16 (Newark, New Jersey, USA: Association for Computing Machinery, 2016), 107–110, https://doi.org/10.1145/2910896.2910913.

For some research questions, using a Wayback Machine would be enough. A Wayback Machine refers to an open-source software project that provides replay for web archival content. For example, there is the Wayback Machine hosted by the Internet Archive at https://archive.org/web/, providing access to their collections. Other institutions also run Wayback Machines, including national and university libraries. Using a Wayback Machine, a user can enter a URL or in some implementations a keyword to search on metadata or home page content, and can find websites to replay. If a researcher already knows what website they are looking for, this can be very fruitful – i.e. if their research question is "What did the World Health Organization homepage say about COVID-19 on January 15th, 2020," this can be readily explored using the Wayback Machine. Advanced functionality found at the Internet Archive's version is also impressive, allowing you to see a "difference" between two pages; highlighting what might have changed on a webpage between two snapshots.

The Wayback Machine, however, suffers when it comes time to ask questions at scale. A few potential research questions can illustrate. For example, a researcher might be interested in websites that contain a certain keyword (i.e. "COVID-19") and link to a particular domain (i.e. the World Health Organization). Or perhaps, a scholar might want to do exploratory text mining or even keyword searching deep into a website. Finally, we can imagine many questions arising out of emerging techniques in the digital humanities, for example if a scholar wanted to work with images or videos en masse.[2]

Instead of replay via Wayback Machine, scholars in these cases need to work with the web archives as data: the underlying WebARChive files that make the Wayback Machine possible. These are standardized files, defined by an ISO standard (28500:2009) (ISO, 2009). The rigid standardization made possible through a published standard has enabled international cooperation as well as the formation of a robust analytical ecosystem. This means that a Danish web archive, or a Canadian web archive, or a Chinese web archive, or a web archive created by an individual researcher can all be analyzed by the same tools based around the WARC specification. However, working with WARCs at scale requires two things: the ability to work with data at scale (small collections are often dozens of gigabytes in size, and medium-to-large ones can quickly get into the terabytes) and the ability to use specialized software.

If replay is not enough, but working with web archives is so difficult, it is worth briefly reflecting on why we should want to rise to this challenge. As I have argued elsewhere, this boils down to the significance of two key factors: scope and scale. In terms of scale, the Internet Archive and other institutions have data on a previously unimaginable scale. But even more promising to me is the scope: data that never before would have been collected is now being collected by people who are not traditionally in the historical record (Milligan, 2019). These two factors combine to make web archives indispensable. I do not believe you could do justice to a history of the 1990s and beyond – in domains as varied as political, social, economic, cultural history – without making fruitful use of web archives. It is hard to imagine studying almost any topic, whether it is a cultural history of pocket pets, a history

2. For example, see this Taylor Arnold and Lauren Tilton, "Distant Viewing: Analyzing Large Visual Corpora," Digital Scholarship in the Humanities, March 16, 2019, https://doi.org/10.1093/digitalsh/fqz013.

of childhood and youth in the late 1990s, or electoral histories, without drawing on webpages. Yet this leaves us with a serious problem. On the one hand, researchers need to use web archives. On the other hand, the tools and supports that currently exist do not easily support research beyond replay.

What can we do? We can attempt to solve this impasse through three discrete approaches: first, new skillsets for researchers; second, the development of new services to deliver data; thirdly, and finally, by introducing new ways to equip scholars to understand the data.

New Skills to Work with Web Archives at Scale

What new skillsets does a scholar need? First, they need to be able to work with data at scale. Many of the principles that support the use of WARCs are akin to working with any other large dataset. Here, I draw on an understanding of "Big Data" that defines it simply as "more data than you could conceivably read yourself in a reasonable amount of time, or that requires computational intervention to make new sense of it" (Graham, Milligan & Weingart, 2015). To do so, a scholar requires an understanding of:

- Natural Language Processing: This is using computers to analyze a body of unstructured text – the text of webpages, for example, or the text of hundreds of books extracted from a library. Popular packages for this include the Python Natural Language Toolkit, which also has an accessible and free textbook (Bird, Klein & Loper, 2010).
- Basic Statistical Knowledge: Even qualitative questions often need some understanding of their quantitative context. This requires an understanding of how to normalize numbers, calculate averages and medians, and beyond.
- Flexible Data Science Skills: Data science is an ever growing field, and those working with data often need basic computational fluency. This includes an understanding of how to work with comma-separated value files, move data between formats, draw on data science libraries and code examples, and – crucially – understand how to troubleshoot when things go wrong. Resources like the Programming Historian (https://programminghistorian.org/) can help here.

In other words, a scholar needs to be equipped with the skills and capacity to analyze data at scale generally. All of the above are applicable to web archives as well as most other forms of computational data.

As well as being able to work with data, a researcher needs to be able to understand the data's context. These skills are a little bit different but are also widely applicable to humanists and social scientists working with data:

- How and Why was Data Collected: What were the selection criteria? Why was a given website selected and another one not? We have run into this problem in the past with web archival collections, where we discover that the selection criteria were never documented – leading to trouble with making historical claims. In other words, we needed to know whether we were measuring the underlying 'reality' or the collection itself (Maemura, 2018).

- How the Collecting Software has Changed: The same website crawled in 2010 and 2020 will have different results, even if the actual website is the same. We need to have a basic understanding of crawling – and the state of the field – to know what might be getting collected and what might not. For example, user comments that are provided by a third-party platform (i.e. Disqus) may or may not be collected depending on the crawl mechanism (Ruest & Milligan, 2014).
- How to Clean or Normalize Data: This entails understanding the basics of data cleaning; an excellent overview can be found at the Programming Historian (Hooland, Verborgh & De Wilde, 2013). For example, imagine a dataset where the URL is collected at different points as either http://www.ndp.ca, https://www.ndp.ca, and http://ndp.ca. These are all the same underlying resource, but in analysis, we need to make sure to aggregate them (or not, depending on the research question).

On top of all of this, of course, is one final skillset that looms over everything: the skillset of a historian/social scientist/humanist. All of the discipline-specific knowledge, acquired through the painstaking process of graduate-level education, is still required as one parses these new sources. As scholars seldom learn much of the above during their doctoral education, adopting new skillsets can seem like a tall order indeed.

To take stock, then: web archives exist, but analyzing them is hard. Existing tools to analyze web archives take quite a bit of new skills if they are to be effectively employed. The learning curve might simply be too high. So should researchers give up? It would not be much of a WARCnet paper, of course, to argue this! Rather, we need to rise to these challenges in order to meet the problem of web archive analysis. Fortunately, our research team – discussed below –has been working on this very problem.

THE ARCHIVES UNLEASHED PROJECT: DEVELOPING INFRASTRUCTURE FOR RESEARCHERS

Since 2015, our Web Archives for Historical Research Group at the University of Waterloo and York University began to tackle the problem of making web archives accessible at scale for researchers. All of the above was at the front of our minds as we began to develop tools and infrastructure, primarily since 2017 as part of our Andrew W. Mellon Foundation-supported Archives Unleashed project. Our team is interdisciplinary and has three investigators: myself, Nick Ruest (a librarian/archivist at York University), and Jimmy Lin (a computer scientist at the University of Waterloo). We are joined by a full-time project manager, Samantha Fritz, as well as a wonderful team of graduate students and between 2017 and 2019, Ryan Deschamps as a postdoctoral fellow. Throughout this paper, you'll see me using the words "we": it really is a group effort.

We informed our work by wanting to understand the approach that scholars would take to work with web archives. Our tools needed to support a wide range of research activities by historians and other scholars, and through first-hand observation and a series of

datathons, we found that scholars tend to carry out their work in four discrete stages.[3] These stages, discussed in depth in a forthcoming paper (Ruest et al., 2020), and which were adapted and extended from a previously-introduced model (Lin et al., 2017), are:

- Filter: As we note, a "scholar usually begins by focusing on a particular subset of the web archive, which we characterize as filtering. This can be accomplished by content, metadata, or some extracted information" (Ruest et al., 2020) For example, content might be keywords (only pages that contain the string "climate change"); metadata (only pages crawled in 2018); or other extracted information (only pages detected using Apache Tika to be in French language). This helps move a scholar from needing to work with terabytes to perhaps gigabytes or even less.
- Extract: The scholar then "extracts" the information of interest: the raw HTML, entities (i.e. people, places, or things), images, PDFs, hyperlinks, etc. This stage needs to be extendable through user-defined functions to facilitate ever-expanding research questions
- Aggregate: We then want to find information from this extracted information – this might range from summarizing information in tabular format, to finding outlier pages (i.e. the pages that receive the most inbound hyperlinks) or average pages (i.e. pages that receive the median amount of inbound hyperlinks).
- Visualize: Finally, results are presented in some sort of output for the scholar to consume: whether this is a table, a dynamic graph visualization to be loaded into Gephi, or beyond.

We understand this as the "FEAV" cycle. Our use of cycle refers to the iterative process, where a scholar might continue in several processes to filter, extract, aggregate, visualize; find more research questions, filter more, and so forth.

The first place that this work saw fruition was in the development of the Archives Unleashed Toolkit. The Toolkit evolved out of the earlier Warcbase project (Lin et al., 2017). It is "an open-source platform for analyzing web archives built on Apache Spark, which provides powerful tools for analytics and data processing" ("The Archives Unleashed Toolkit", 2020) In short, it allows a user to take WARC (or the earlier ARC format) files and carry out the operations described in the FEAV cycle on it.[4] Apache Spark, a modern platform for big data processing, allows it to scale: the same script can be used to process terabytes of data on a laptop, a server, or a multi-node cluster. Indeed, I have slowly processed a 4TB GeoCities collection on a 2015 MacBook Pro using the Toolkit. Users can code in Scala or Python to work with the WARC files directly. You can see the Toolkit in Figure 1.

3. Datathons described in Ian Milligan et al., "Building Community and Tools for Analyzing Web Archives through Datathons," in Proceedings of the 18th Joint Conference on Digital Libraries, JCDL '19 (Champaign, Illinois: IEEE Press, 2019), 265–268, https://doi.org/10.1109/JCDL.2019.00044.

4. WARC files can be difficult to find. Some web archiving collection tools, such as WebRecorder.io, allow for the native export of WARC files. At other times, you may need to talk to a librarian or forge a partnership to be able to work with these files. When in doubt, ask!

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.3.0
      /_/

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_161)
Type in expressions to have them evaluated.
Type :help for more information.

scala> :paste
// Entering paste mode (ctrl-D to finish)

import io.archivesunleashed._
import io.archivesunleashed.matchbox._

val r = RecordLoader.loadArchives("example.arc.gz", sc)
.keepValidPages()
.map(r => ExtractDomain(r.getUrl))
.countItems()
.take(10)
```

Figure 1: The Archives Unleashed Toolkit in Action

Our initial expectations were that humanists and historians would be able to use the Toolkit by taking examples in our robust documentation "cookbook" and changing variables in order to work with it on their own data (i.e. changing the input path, the output path, and the specific filters or extractors being used). Unfortunately, we discovered that this required more technical competency than we could expect of the average scholar in light of all the skills they already needed: users needed not only to know how to tweak code, but how to use a command line, follow the flow of programming scripts, and then have patience for open-source documentation and projects. In order to engage historians, it became clear that we would need to do something more than just provide command-line-based tools. With this in mind, we sought to bridge the gap between tools and users.

Bridging the Gap between Tools and Users

Thinking about infrastructure, we realized that we needed to move away from WARC files and towards derivative file formats. While WARCs are fantastic, standards-based files, the research user community is too small to use them effectively. For example, many digital humanities centers or university libraries did not seem to be able to provide effective support to researchers who wanted to work with WARCs; however, if a scholar came with a question to do with generic text analysis or network analysis, they would be able to find supportive resources. Relatedly, the nature of WARC files makes them large in size. Derivative files, properly filtered as discussed above, could make them easier to work with even on a scholar's personal laptop.

Through discussions with researchers and datathons, we settled on a series of standard derivatives that the Toolkit would support. These included plain text, the hyperlink network of a collection, the images in a collection, the domains present, a variety of standard binary files (from Word documents to PDFs to PowerPoint presentations to software programs), and other such types.

The problem with the Toolkit was that we had left too big a gap between researchers and web archives tools. I often like to visualize this as a bridge that takes mutual effort. Tools developers need to try to come towards their users and craft usable, intuitive, yet powerful tools. Conversely, researchers need to make a good-faith effort to attempt to use tools and develop their skills accordingly. Figure 2 sketches this out.

Figure 2: The Gap between Researchers and Web Archives

So how can we bridge this gap? Our team identified three main approaches. First, researchers could become programmers/developers and move towards the web archives. Second, tools could become completely intuitive and meet the lower technical skills level of researchers – think of something like the JSTOR approach (you need to be smart to use JSTOR, but you do not have to be a "JSTOR historian"). Or, third, everybody could meet in the middle. We decided to assign three personas to these three rough outcomes: a computational humanist, a digital humanist, or a conventional historian.

Ultimately, we developed three different tools and approaches to line up with three of these main approaches/personas: the Archives Unleashed Toolkit to serve a scholar familiar with advanced research computing; the Archives Unleashed Cloud to serve a researcher familiar with working with broad derivative formats and data, but not at the scale or depth needed to work with WARC data; and the Archives Unleashed Notebooks to serve a more general audience of researchers familiar with web browsers but perhaps not much more. Let us discuss each of these audiences in turn.

Our first audience was the computational humanist, or the model where we could reasonably expect users to come towards "us" and thrive in a computational environment.[5]

5. My use of the "computational humanist" label is not meant to be unduly provocative, and is shorthand for a scholar who is using advanced research computing to work with data. There have been some discussions around the computational humanities as distinct from the digital humanities, see Leah Henrickson, "Humanities Computing, Digital Humanities, and Computational Humanities: What's In a Name," 3:AM Magazine

The basic criterion for this was an individual who was comfortable installing software packages, understanding documented dependencies, fluent on the command line to the extent that they could follow intermediate-level instructions and locate files/directories/and beyond, and – perhaps most crucially – that they could parse error dumps and find the right question to ask the popular Stack Overflow question-and-answer site. This did not mean that we would simply throw this individual into the thick of it without support: rather we would provide technical documentation and sample scripts, and would thus let the user take WARC files and generate derivatives themselves. The Archives Unleashed Toolkit does all of this. As you can see from above, our project's original mistake between 2015 and 2017 was to assume that all of our users fell into this category. While there are some users in this category, there are not enough. Crucially, as a project we had made the mistake of not coming far enough towards our users.

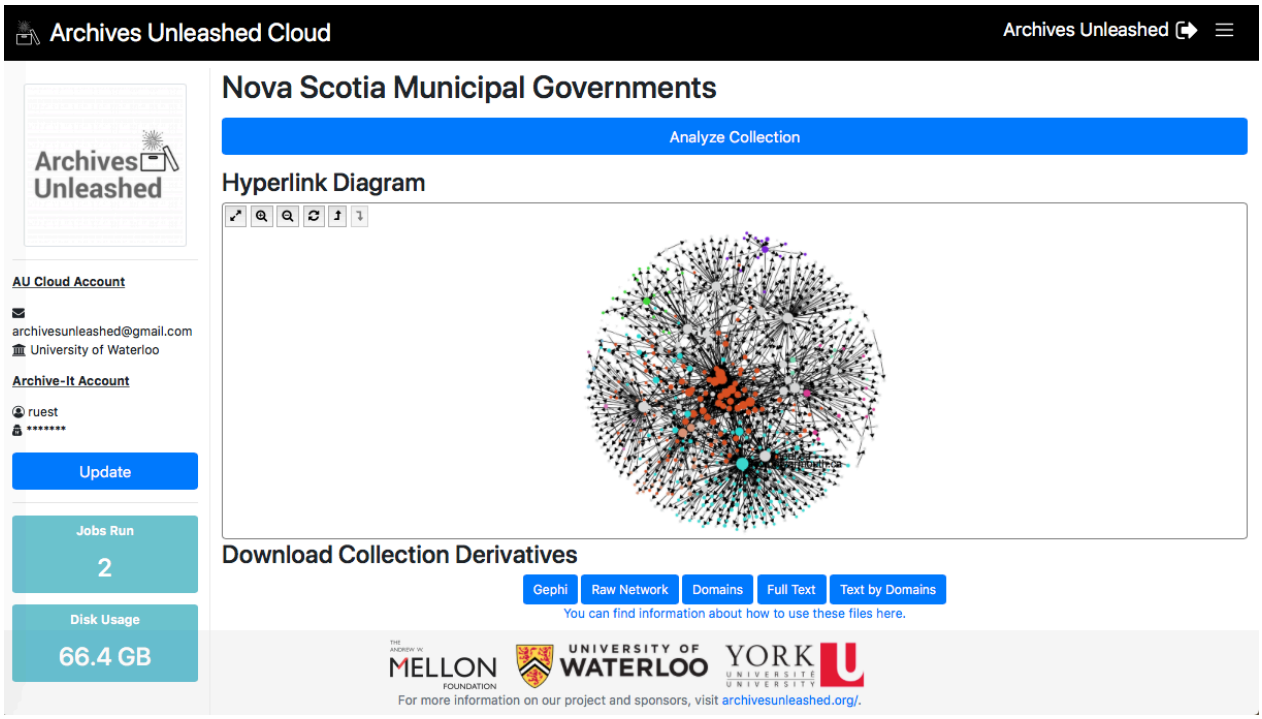


Figure 3: The Archives Unleashed Cloud in Action

The second audience, then, was digital humanists more generally. We identified these scholars as comfortable with computers, having a critical understanding of data, and having the ability to draw on larger networks to allow them to use tools such as Voyant, Gephi, or – with some work – drawing on the Programming Historian for resources for basic Python or R. Yet while full of research questions, digital humanists could not easily take advantage of the Archives Unleashed Toolkit in our experience. We felt that these users needed to have some sort of mediating body to translate WARC files into something they could work with, and while they could work with the command line, once we started getting into fairly

(blog), October 24, 2019, https://www.3ammagazine.com/3am/humanities-computing-digital-humanities-and-computational-humanities-whats-in-a-name/.

tricky dependency clashes we would begin to lose our users too quickly.[6] The Archives Unleashed Cloud was designed around this persona. The Cloud allowed a user to take WARCs and use a modern UI to sync collections, run basic analyses in the browser, and download derivative file formats that can integrate with standard workflows. You can see the Cloud in Figure 3. In other words, it gets the WARC out of the equation and translates it into a standard file format. We were moving towards our users, but specialized digital humanities skills were still required.

Finally, what about the "conventional" audience? To us, this was the sort of scholar who might use computers – for Word, some light Excel – but otherwise generally just wants to do historical research without learning new technical skills. Our first internal question was the degree to which we should attempt to serve this audience. We did not take this question lightly. This is the toughest group of users to reach. If with the Archives Unleashed Toolkit our team was arguably not making enough of an effort to serve our users, in this case, users may not be making the effort that they need in order to reasonably work with web archives. While we want to keep barriers to entry low, we do want to have tools that are not complete black boxes that researchers use and have no clue what is happening. The skills that a 21st-century historian needs to have are, of course, still subject to debate.[7]

We decided to try to serve this group using the increasingly-popular Jupyter notebook approach. Notebooks allow users to write programming code, coupled with text descriptions and embedded visualizations, directly in their web browser; while they still require knowledge to use, they are easier than needing specialized software.[8] As notebooks are increasingly able to be run in the cloud via platforms such as Google Colab or Binder, they raise the possibility of being able to carry out advanced research computation both in the cloud and without the barriers to entry of installation and navigating software dependencies. In order to introduce researchers to web archive analysis, our team has begun to host derivatives in the cloud.

6. Lest anybody think I am being cavalier; I still struggle with Rails dependencies on the software project that I am the PI of.

7. As I discussed in Milligan, History in the Age of Abundance?

8. The International Internet Preservation Consortium is also funding a notebook approach to web archive analysis. See http://netpreserve.org/projects/jupyter-notebooks-for-historians/.

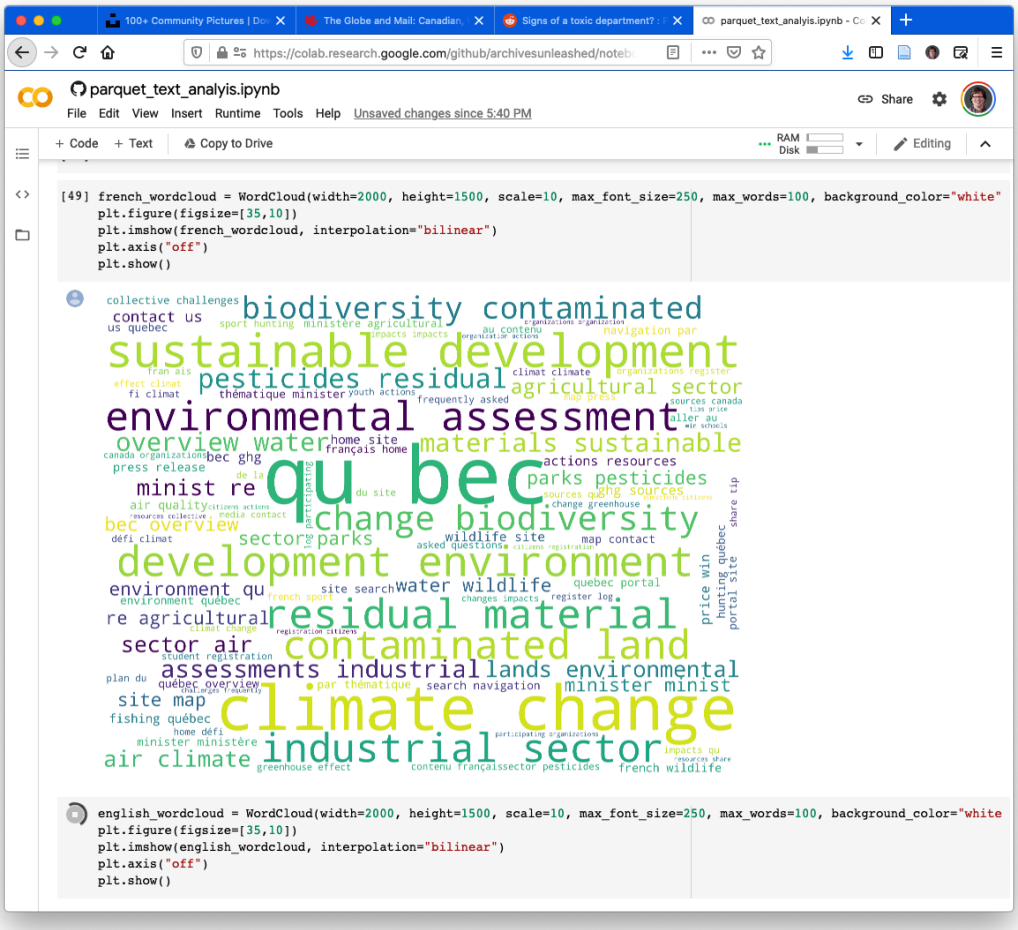


Figure 4: Archives Unleashed Notebook

Our team has written about the notebooks elsewhere, but our approach can be essentially summed up as "fill in the blanks." This is what we mean:

> The notebook guides scholars through sample analyses, each corresponding to a potential research question, using a fill-in-the-blanks 'madlibs' approach. For example, to specify the collection to analyze, the scholar only needs to enter the collection id (see Figure 1), and upon reexecution all analyses and visualizations will update appropriately. Each analysis supports one or more parameterizations (e.g., time period of study) that scholars can adjust, supporting customization in specific ways. Once again, the scholar only needs to fill in the blank, and all analytical results will be updated (Deschamps et al., 2019).

An example of the notebook can be seen in Figure 4. While significant work remains to be done, the goal here is to let researchers begin to get a sense of what is possible with web archive derivatives; and perhaps, with smaller collections, even carry out their primary research within them. Thanks to cloud-hosted notebooks, one click from our project website and suddenly researchers are able to explore the derivatives firsthand. In other words, the gap is narrowed as far as our team can make it. Will users come?

The final step to help make much of this possible, then, was developing community through in-person – and more recently, virtual – events that bring researchers, curators,

and tools developers together to think about what a vibrant web archiving community would look like. Community is essential both to the development of communal research agendas and platforms, but also for the survival of any open-source project such as our Archives Unleashed one. Accordingly, cognizant of funded projects that run into sustainability issues following a successful launch, we regard regional datathons staged under the Archives Unleashed banner as vital to ensuring broad community buy-in and continued involvement. The datathon model brings together researchers, programmers, visualization experts, graphical designers, and others into one room in order to facilitate their intensive collaboration on a shared project. In our case, programmers, academics, memory institution professionals, and other librarians gather to work on accessing web archives. Our goal with these events is to lower barriers, bring people together to help network, and crucially, to establish a community of practice to work with web archives (Milligan et al., 2019). More recently, during COVID-19, we moved our New York datathon online and are considering using our experience to perhaps run further online events in order to better internationalize our audience.

CONCLUSION

The goal of all this is to lower barriers. As noted in my introduction, right now if you are a historian who wants to research with web archives, you almost need to become a full-time web historian. With Archives Unleashed, and other similar initiatives, the barriers to entry can decline. While this will still require hard work and skills development, the goal is that researchers will be able to focus on their core research questions. One limitation, hinted at above, is still access to data – to use the Archives Unleashed Cloud, one needs to have an institutional account with the Internet Archive's non-profit subscription service Archive-It. While a librarian can give a researcher a read-only account to use the Cloud, it still requires relationships and connections to libraries and other archiving institutions. This remains a barrier to access.

Historians in the future will need to understand the Web. They're not ready to do so and we need to make sure that they can, not by devoting their lives to web history, but rather, by using the expertise and talent of web historians and other scholars in networks like WARCnet. This will involve several considerations acting in concert: the development of new, usable tools (i.e. toolmakers coming towards researchers) and new cultures in the humanities and social sciences (i.e. this might involve researchers coming towards toolmakers), but ultimately, it will require a shared vision that web archives are important and that we need to find common ground to support the next generation of scholarship. If we think of the gap outlined in Figure 2, we need to make sure to shrink it. If we can do this, it will be worth it.

REFERENCES

Bird, S., Klein, E. & Loper, E. (2010). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* https://www.nltk.org/book/

Brügger, N. & Milligan, I. (2018). *Introduction in SAGE Handbook of Web History*. London, SAGE Publications

Deschamps, R., et al. (2019). The Archives Unleashed Notebook: Madlibs for Jumpstarting Scholarly Exploration of Web Archives. *ACM/IEEE Joint Conference on Digital Libraries (JCDL*). https://ieeexplore.ieee.org/document/8791210/

Graham, S., Milligan, I. & Weingart, S. (2015). *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press. https://www.worldscientific.com/worldscibooks/10.1142/p981

Hooland, S. V., Verbourgh, R. & De Wilde, Max. (August 2013). Cleaning Data With OpenRefine. Programming Historian https://programminghistorian.org/en/lessons/cleaning-data-with-openrefine

ISO. (2009) ISO 28500:2009. https://www.iso.org/standard/44717.html

Lin, J., et al. (July 2017). Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. *ACM Journal of Computing and Cultural Heritage* 10, no. 4

Maemura, E., et al. (2018). If these Crawls Could Talk: Studying and Documenting Web Archives Provenance. *Journal of the Association for Information Science and Technology 69*, no. 10. https://doi.org/10.1002/asi.24048

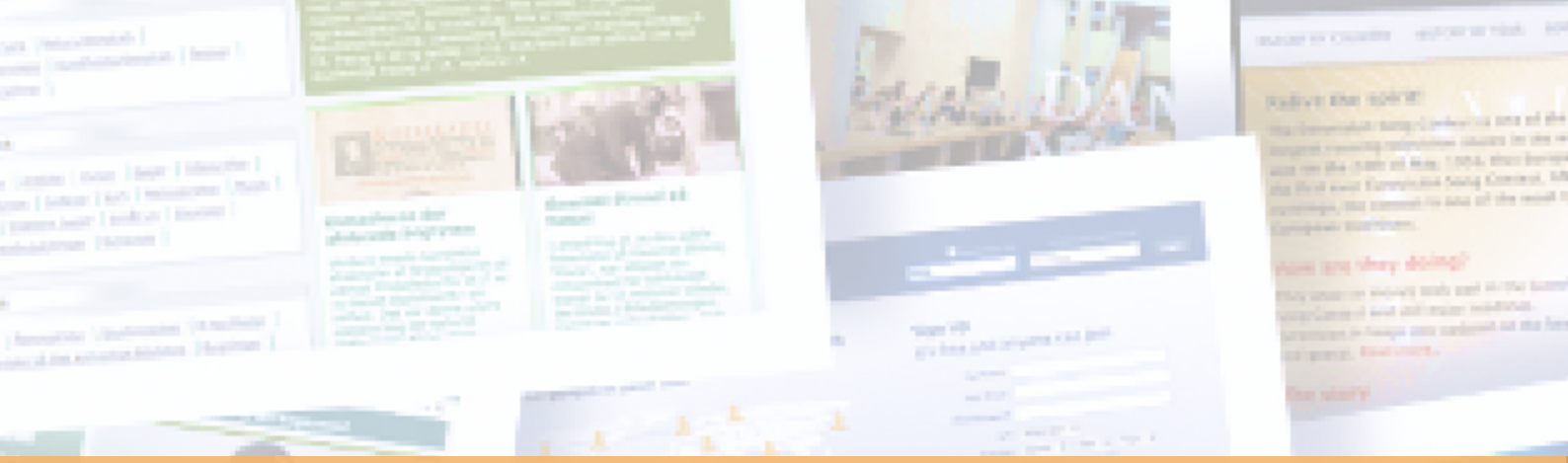
Milligan, I. (2019). *History in the Age of Abundance? How the Web Is Transforming Historical Research*. Kingston and Montreal: McGill-Queen's University Press.

Milligan, I., et al. (2019). Building Community and Tools for Analysing Web Archives through Datathons. *Proceedings of the 18th Joint Conference on Digital Libraries, JCDL*. Champaign, Illinois: IEEE Press, p. 265-268. https://ieeexplore.ieee.org/document/8791131/

Ruest, N. & Milligan I. (2014). The Great WARC Adventure: WARCs from Creation to Use. *Association of Canadian Archivists*, Victoria, BC.

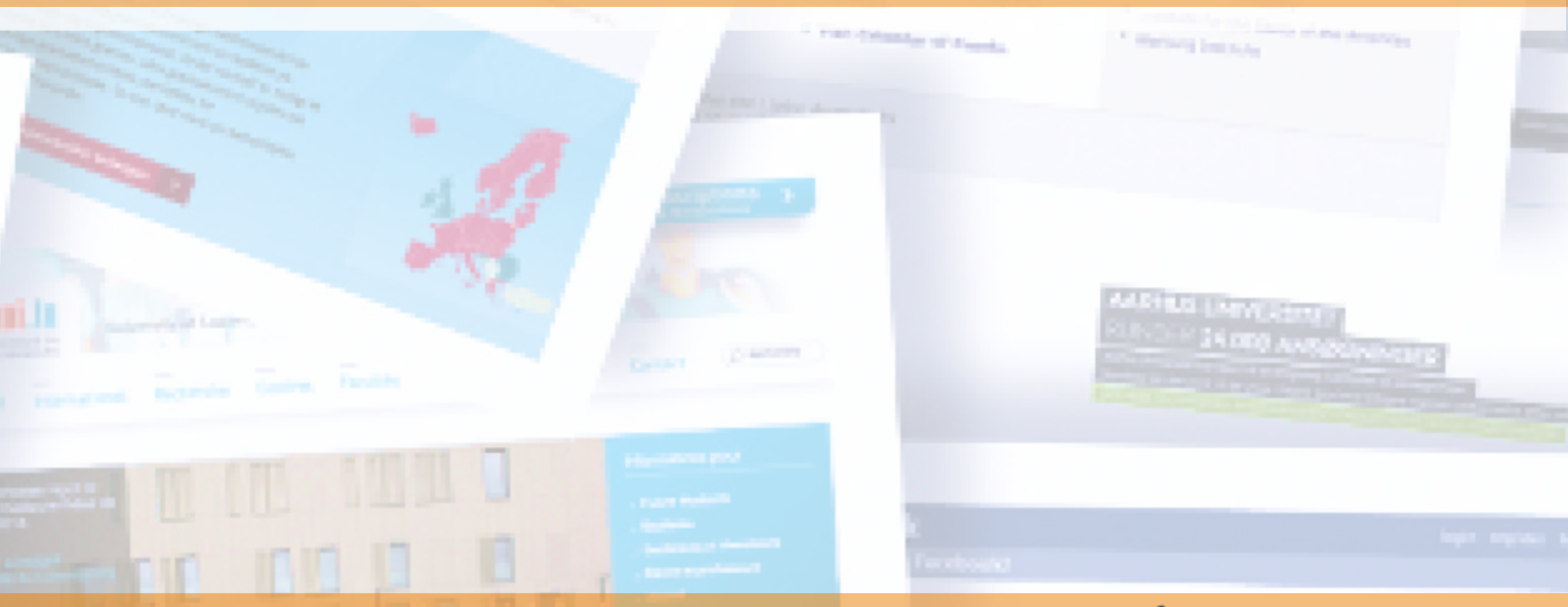
Ruest, N., et al. (2020). The Archives Unleashed Project: Technology, Process and Community to Improve Scholarly Access to Web Archives. *Proceedings of the Joint Conference on Digital Libraries, JCDL*. http://arxiv.org/abs/2001.05399

The Archives Unleashed Toolkit. (2020). Archives Unleashed Project. https://archivesunleashed.org/aut/



WARCnet Papers is a series of papers related to the activities of the WARCnet network. WARCnet Papers publishes keynotes, interviews, round table discussions, presentations, extended minutes, reports, white papers, status reports, and similar. To ensure the relevance of the publications, WARCnet Papers strives to publish with a rapid turnover. The WARCnet Papers series is edited by Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Sofie Flensburg, Peter Webster and Michael Kurzmeier. WARCnet Papers have gone through a process of single blind review.

The aim of the WARCnet network is to promote high-quality national and transnational research that will help us to understand the history of (trans)national web domains and of transnational events on the web, drawing on the increasingly important digital cultural heritage held in national web archives. The network activities run in 2020-21, hosted by the School of Communication and Culture at Aarhus University, and are funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).



WARCNET PAPERS

INDEPENDENT RESEARCH FUND DENMARK

warcnet.eu warcnet@cc.au.dk twitter: @WARC_net facebook: WARCnet
youtube: WARCnet Web Archive Studies slideshare: WARCnetWebArchiveStu