Building Community and Tools for Analyzing Web Archives through Datathons

Ian Milligan,[1] Nathalie Casemajor,[2] Samantha Fritz,[1] Jimmy Lin,[1]
Nick Ruest,[3] Matthew S. Weber,[4] and Nicholas Worby[5]

[1] University of Waterloo [2] INRS [3] York University Libraries
[4] University of Minnesota [5] University of Toronto Libraries

ABSTRACT

Starting in March 2016, the Archives Unleashed team and our collaborators have brought together social scientists, humanists, archivists, librarians, computer scientists, and other stakeholders to explore web archives as research objects. Three objectives motivated our team to develop and organize these events: facilitating scholarly access, community building, and skills training. We believe that we have been successful on all three fronts. For each event, over the course of two to three days, participants formed interdisciplinary teams and explored web archives using a variety of methods and tools. This paper details our experiences in designing these "datathons", with an intent to share lessons learned, highlight interdisciplinary approaches to research and education on web archives, and describe future opportunities.

1 INTRODUCTION

Starting in March 2016, we have brought together disparate groups that include those who create web archives, those who create tools and platforms, and those who use them for research. Each of these "datathons" has brought together twenty to fifty individuals, and over the course of two to three days, participants formed interdisciplinary teams and were given access to data and computing infrastructure to develop a project around a web archive collection. These events have resulted in expanding participants' knowledge of methods, tools, and approaches to tackling web archive data at scale. Our datathons have three primary objectives: facilitating scholarly access, community building, and skills training. Most notably, the community-building element has seen us connect, build relationships, and develop the social infrastructure for a burgeoning network of individuals and groups held together by the common goal of exploring web archives as research objects.

By reflecting on these datathons, we present an approach to tools development and community building that can accomplish several goals. We begin with a descriptive characterization of these events and then articulate the contributions that they have made to the study of web archives, in three main ways. These include:

- The tangible development of tools and platforms that meet demonstrated needs (i.e., better support for scholarly inquiry, as identified in our original proposals for these events);
- A better understanding of the processes by which scholars, curators, and others work with these materials, providing a reference workflow with which to evaluate future research tools;
- The building of a community, in part supported by the continued use of datathon communication channels and standing infrastructure, as well as encouragement to attend follow-up events.

Finally, through feedback and an iterative process of designing events, surveying participants, and starting anew, we provide generalizable lessons around running datathons in the digital library and cultural heritage environment.

2 CONTEXT AND BACKGROUND

Big data has the potential to reshape humanistic and social science research. The sheer amount of cultural information that is generated and, crucially, preserved every day in electronic form, present exciting new opportunities for historians [6]. Much of this information is captured within web archives, which contain hundreds of billions of web pages, ranging from individual homepages and social media posts to institutional websites.

Web archives provide an opportunity for researchers and scholars in the humanities and social sciences, as they become an access point to reconstruct large-scale traces of the relatively recent past [8, 9]. Simply put, web data enhances research topics that date back to the mid-1990s; this is not for those studying the web per se, but for those examining social and cultural activities taking place in an era of born-digital web sources.

Yet the opportunity to explore information and artifacts presented in web archives is hindered by several challenges, most notable of which stems from the need to process and analyze the sheer amount of data currently available. We have more accumulated data than ever before, and the rate at which we are capturing potentially valuable historical data is accelerating. But the scale is overwhelming. For serious research, we need to develop new tools and methods to make sense of this digital deluge, a point which has been explored in several papers and workshops at JCDL [2–5, 10].

The size of these archives eludes traditional finding aids and requires more than the ability to examine individual source documents. Today, scholars are mostly limited to viewing one page at a time in a web archive. In addition, web archives are currently underused because of the high barriers to entry. When a scholar approaches a web archive for the first time, she often has little idea where to start. Yet the need for access is very real. We have participated in numerous web archiving conferences and events (International Internet Preservation Consortium meetings; the Research Infrastructure for the Study of Archived Web Materials conferences; and many others), which have enabled us to become familiar with this community. Conversations with colleagues at these venues have revealed several key, recurring requirements among scholars, which we attempt to tackle with our datathons.

It is clear that current tools for accessing archived web content presents challenges for scholars. While the Internet Archive makes archived web content available to the general public through its

978-1-7281-1547-4/19/$31.00 ©2019 IEEE
DOI 10.1109/JCDL.2019.00044

"Wayback Machine," which allows visitors to enter a URL or to search via keywords to visit archived web versions of a particular page, this system is limited: not only do visitors largely need to know the URL or exact phrase in the first place, they are also limited to individual readings of single webpages. Web archives do, however, offer the potential for more powerful access methods. By directly interrogating web archives in their raw form, primarily WebARChive (WARC) files, our datathons seek to develop new ways to systematically analyze and visualize changes over time.

One final consideration: we find that the discussion around web archives is somewhat segmented. Information professionals and digital archivists lead the ongoing dialogue around digital preservation and web archiving practices, yet, for the most part, humanists and social scientists have not participated in this discussion. This gap must be filled as these researchers, notably political scientists, historians, sociologists, and communications scholars, will be among the primary consumers of large web archives. Despite these challenges, web archives continue to be an important resource for the humanities and social sciences. In recognizing the need for approaches to these challenges, we have (in various combinations) organized several datathons. Our model is presented here.

3 THE DATATHON MODEL

A datathon brings together scholars, curators, subject matter experts, developers, and other interested parties into one room in order to facilitate intensive collaboration over a relatively short time on a shared project. In our case, developers, academics, and memory institution professionals gather to work on analyzing web archives. As social media collections (e.g., Twitter) are also of substantial interest, we have taken a more inclusive view of these as "web archives" also—many Twitter analytics capabilities have been built into web archiving tools (indeed, attendee interest at the datathons have spurred this).

A typical datathon has the schedule seen in Table 1 (which is more reflective of later events that incorporate lessons learned). The event begins with a discussion of current tools, platforms, and related issues to set the stage. Groups subsequently form and then proceed to work on different research projects with the aim of presenting their results at the end of the event. Teams are given access to datasets provided by a variety of institutions, as well as computing resources (virtual machines) from Compute Canada (which aims to offer advanced research computing support to Canadian researchers; in our case, we use their OpenStack platform).

Team formation is one of the most challenging aspects of the datathon, especially on the tight timelines that we have in our events. We use a "sticky notes exercise" to quickly coalesce teams, a technique adapted from participatory design [11]. To begin, each participant is given access to three distinct colours of sticky notes. On one colour they are asked to write research questions, on another colour research methods or tools, and on the final colour datasets that they are interested in. There are no limits on how many (or how few) sticky notes are allowed. Participants are then asked to stick their notes to walls around the room and also examine notes by their colleagues. When most participants have 'stuck' their notes, the organizers (us) begin to cluster similar notes together, with input from the participants—these clusters might contain notes of a single colour, or might reflect emergent themes that cross-cut different categories. One example might be a cluster of participants interested in hyperlink analysis, perhaps applied to examining discourse between political parties. Invariably, coherent clusters emerge, sometimes methodological, other times focused around tools or datasets, and we can quickly bring together people who have never met but share common interests. From this, we ask participants to form initial groups and continue their discussions. Participants are encouraged to physically move around the room, "test out" different groups, and to continue refining their ideas, but with attention to two main rules—that teams should ideally be smaller than six people and they should not contain individuals from the same institution. Periodically, we ask each group to provide a quick summary of their thoughts, and usually in about half an hour working groups are successfully formed.

Time	**Activity**	
	Day 1	*Day 2*
09:00	Breakfast and Welcome	
09:30	Introductory Remarks	Group Work
10:00	Introduction to Tools	Group Work
10:30	Coffee Break	
10:45	Sticky Notes Exercise	Group Work
11:15	Group Work	Group Work
12:30	Lunch and Lightning Talks	
13:15	Group Work	
15:00	Coffee Break	
15:30	Group Work	Awards and Closing
17:30	Evening Social	

Table 1: Sample datathon schedule.

Why did we want to bring people together? While many disparate groups have contributed to the creation of web archive tools and datasets, there have been precious few forums or mechanisms for coordinated, mutually-informing efforts. We identified a much-needed collaborative opportunity to work with cutting-edge research tools and to develop both a consensus on future directions in web archive analysis and a roadmap for getting there.

4 OVERVIEW OF EVENTS

In this section, we discuss the datathons held between March 2016 and November 2018. Another event was held in March 2019 at George Washington University, after the JCDL submission deadline.

4.1 University of Toronto

The first datathon was held in March 2016 at the University of Toronto Library (UTL). Supported by UTL, the Social Sciences and Humanities Research Council (SSHRC) of Canada and the US National Science Foundation (NSF), 45 individuals were selected to participate after a competitive application process. We began with a half day of research talks: research tool developers (Warcbase and ArchiveSpark), Internet Archive staff, and others shared insights through a series of presentations. The actual "hacking" did not begin until the second day. Teams formed into groups, worked

on their projects for a day and a half, before convening for final presentations. Notable projects included:

- The "Interplanetary Wayback", later published in JCDL [1];
- Enhancements to Archive Spark and Warcbase;
- Exploring meme images in Trump's early primary campaign, allowing attendees to explore an unfolding meme event;
- Explorations of Canadian political images and links.

The technical logistics for this event differed from later ones as datasets resided on *physical* media and data transfer times were an issue. Our takeaway from this experience resulted in different processes for data distribution; future events would have cloud-resident data to streamline collection transfer between machines.

4.2 Library of Congress

With support from the National Science Foundation and residual funds from both NSF and SSHRC, the project team ran a follow-up event in June 2016 at the Library of Congress (LC) in Washington DC. With the generous support of both LC and Dame Wendy Hall, 2016 Kluge Chair in Technology and Society, the datathon was held two days immediately preceding the "Saving the Web: Ethics & Challenges of Preserving the Internet" event. This datathon was large, with approximately 45 participants. We followed a similar format to the event in Toronto, with a half day devoted to discussion, followed by a day and a half of hacking. After the event, the datathon was presented at the "Saving the Web" symposium, along with an overview of the datathon [7]. Notable projects included:

- An exploration of news sites within the Cuban web domain;
- An analysis of deleted content in a UK election archive;
- An analysis of web archive data from Supreme Court nominee hearings, examining patterns in linking and content;
- An analysis of tweets from known terrorist accounts.

4.3 Internet Archive

In February 2017, the Internet Archive hosted a Web Archiving Systems API (WASAPI) symposium. Our team partnered with them to host a datathon during the following two days. This was a smaller event, amounting to approximately 15 participants. Similar to previous events, the datathon opened with talks for half of day one, leaving a day and a half for groups to actually do "hands on" work. This was a compressed event, and the decision was made in its aftermath to reduce the amount of talks. We also learned that dataset transfer time in the cloud was still a challenge, as attendees needed to have WARCs transferred from servers to their machines, which led to wasted time. This resulted in the decision to pre-stage datasets on VMs in future events. Projects included:

- Understanding local news flow via headlines and content;
- An analysis of "fake news", specifically how Twitter users interpreted and shared quotes from presidential debates;
- A team which analyzed the "end of term" archives, exploring change in US Congresses between 2001 and 2017.

4.4 British Library

Our final datathon held under the initial set of SSHRC and NSF funding was held in June 2017 in London, UK. It was co-located with another event, in this case a joint conference of the International Internet Preservation Consortium (IIPC) and the European RESAW web archiving research organization. The event was made possible through the generous funding and assistance from IIPC, the British Library, and other partners. This was a two-day event with approximately 40 participants plus organizers. The emphasis was getting people "to the tools" as soon as possible, meaning that teams were formed and working by mid morning. We applied takeaways from previous events, in particular the lesson of pre-staging datasets on VMs. This made for a dramatically quicker ramp up and allowed the organizers to worry less about technical issues and more on addressing content questions. We also polled attendees prior to the event to understand what datasets they would be interested in using. Projects included:

- A project which looked at the overlap between archival collections, using the "Occupy" movement as a case study;
- An exploration of how relative versus absolute URLs have changed over time (important for collections);
- Explorations of the `robots.txt` protocol and how honouring their crawler exclusions influenced collections.

4.5 University of Toronto

From this point on, the datathons changed as they were now being run by a subset of the original organizers: the Archives Unleashed team, with support from a grant by the Andrew W. Mellon Foundation. However, these events continued to benefit from the oversight of an advisory board comprised of all the organizers of the previous events. The focus in these smaller events (~15 participants) also shifted from broader research questions to focus primarily on work using the Archives Unleashed Toolkit,[1] a platform funded by our grant. Based on previous events, we observed that teams spent a lot of time *processing* web archives and not enough time *exploring* them. Accordingly, we looked at ways to improve the technical setup, resulting in the decision to not only *pre-stage* datasets, but also provide pre-processed derivatives (plain text, crawl statistics, domain-level webgraph) on *each* of the VMs. Projects included:

- Exploring British Columbia labour dispute archives;
- Analyzing pipeline politics across several Canadian web archives;
- Extracting quality URL seeds to reduce spam links.

4.6 Simon Fraser University

The next datathon that we ran was hosted by Simon Fraser University Library in Vancouver BC, and was supported by the SFU Library and the Mellon Foundation. It saw 15 participants, and was similar in organization to the previous Toronto event. Due to some last-minute participant changes, we had fewer scholars than before, meaning the group mainly consisted of librarians and archivists. The comparatively smaller number of scholars led to some challenges in terms of articulating projects. Projects included:

- Text mining and analysis of BC provincial politics;
- An exploration of development and indigenous groups in BC;
- Exploring BC wildfire web archives;
- Automating quality in web archives using visual analysis.

[1] https://github.com/archivesunleashed/aut

5 ACCOMPLISHMENTS

The datathons represent significant moments where we learn about new requirements by looking at research questions, allowing for continued improvement at each new event. They also offer a learning opportunity to inform our approach to community building. We believe that we have expanded the community, both through the events themselves, as well as continued use of datathon communication channels and encouragement to attend follow-up events. The datathons have also allowed us to observe the process by which scholars, curators, and others work with these materials, leading to a "reference workflow" with which to evaluate future research tools. Specifically, these events informed the development the filter, analyze, aggregate, and visualize (FAAV) cycle to characterize scholarly interactions with web archives [6].

Finally, the datathons have informed the development of tools and platforms that meet demonstrated needs. With support from the Mellon Foundation, the Archives Unleashed team has built the "Archives Unleashed Cloud", which can be viewed as the canonical deployment of our Archives Unleashed Toolkit. A cloud platform saves scholars from having to procure computing resources and then download, install, and configure the toolkit themselves. Observing projects at the datathon and soliciting feedback from participants has been the primary vehicle for organizing our roadmap. Tools and capabilities developed based on insights from earlier datathons then feed subsequent events, thus creating a virtuous cycle that not only enhances the tools and platforms themselves, but also enriches the community.

6 LESSONS LEARNED FOR DATATHONS

Lessons learned will be useful to others organizing similar events:

- **More Hack, Less Yack**: Although cliché, it remains nevertheless true that a productive datathon needs to be a place for hands-on engagement. We pared down the talks in later events so that teams could work on the projects at hand. This required modifying logistical aspects, such as how we deliver datasets to participants, to ensure efficient use of time.
- **Right Mix of Participants**: Much of the work of the datathon happens *before* the event itself, when applicants are selected. Our experience strongly indicates that we need the right balance of stakeholders. Ideally, we would have roughly one third from technical domains (computer science, tools developers), one third from subject-matter research domains (social scientists, humanists), and one third from collectors and curators (librarians, archivists). Participants also need to demonstrate willingness for hands-on technical work. Through "homework" before the event, technical walkthroughs, and attention to team composition, we can ensure that our tools are accessible to all participants.
- **Organized but Feasible Homework**: Initially, we directed participants to acquaint themselves with WARC files and web archive analysis using tutorials provided by the Internet Archive. As we gained more experience, for later events we created custom Docker-based tutorials so that participants can arrive at the datathons with sufficient preparation.
- **Staging Datasets in the Cloud**: We learned that large datasets should not be on physical media! Pre-staged datasets on VMs that participants can directly access work far better.
- **Pre-Processing Datasets and Providing Derivatives**: Related to above, one of our smartest moves has been to generate standard derivatives for each collection: full text, domain-level web-graph, and URL metadata. This allows teams to focus on research questions without waiting for common processing jobs to finish.

While some of the above may need to be adapted to different contexts, these lessons are generalizable across a wide range of use cases in the digital libraries community. This is an educational model that differs from the traditional classroom mode of lab and lecture. Successful datathons and successful engagement with web archives call for equipping participants with the tools needed for success, but then allowing sufficient time for bricolage—learning through experimentation—in order to achieve success.

7 CONCLUSIONS

Web archives are a critical resource for future research in the humanities and social sciences. The development of infrastructure and community is vital to supporting interdisciplinary collaboration. The Archives Unleashed team and collaborators have seen the positive and far-reaching benefits of bringing together groups of curators, researchers, scholars, and developers to engage with web archival collections. More datathons will be run and we believe that further engagement and co-locations with the broader digital libraries community would be fruitful.

ACKNOWLEDGMENTS

This research was supported by the Andrew W. Mellon Foundation, the Social Sciences and Humanities Research Council of Canada, the US National Science Foundation (Grants #1624067, #1723430), Start Smart Labs, Rutgers University, Compute Canada, University of Waterloo, and York University. Additional support came from University of Toronto Libraries, Library of Congress, Internet Archive, British Library, the International Internet Preservation Consortium, Simon Fraser University Libraries, SFU Key, and Université du Québec en Outaouais.

REFERENCES

[1] S. Alam, M. Kelly, and M. Nelson. 2016. InterPlanetary Wayback: The Permanent Web Archive. In *JCDL*. 273–274.
[2] A. AlSum. 2015. Reconstruction of the US First Website. In *JCDL*. 285–286.
[3] E. Fox, Z. Xie, and M. Klein. 2016. WADL 2016: Third International Workshop on Web Archiving and Digital Libraries. In *JCDL*. 293–294.
[4] H. Holzmann, V. Goel, and A. Anand. 2016. ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation. In *JCDL*. 83–92.
[5] A. Jackson, J. Lin, I. Milligan, and N. Ruest. 2016. Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. In *JCDL*. 103–106.
[6] J. Lin, I. Milligan, J. Wiebe, and A. Zhou. 2017. Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. *J. Comput. Cult. Herit.* 10, 4, Article 22 (July 2017), 30 pages.
[7] J. Mears. 2016. Co-Hosting a Datathon at the Library of Congress. https://blogs.loc.gov/thesignal/2016/07/co-hosting-a-datathon-at-the-library-of-congress/.
[8] I. Milligan. 2016. Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives. *Inter. J. of Humanities and Arts Comput.* 10, 1 (March 2016), 78–94.
[9] I. Milligan. 2019. *History in the Age of Abundance? How the Web is Transforming Historical Research.* McGill-Queen's University Press.
[10] I. Milligan, N. Ruest, and J. Lin. 2016. Content Selection and Curation for Web Archiving: The Gatekeepers vs. The Masses. In *JCDL*. 107–110.
[11] G. Walsh, E. Foss, J. Yip, and A. Druin. 2013. FACIT PD: A Framework for Analysis and Creation of Intergenerational Techniques for Participatory Design. In *CHI*. 2893–2902.