Building the Next Generation of Web Archive Analysis Service

Ian Milligan*[1], Helge Holzmann*[2], Nick Ruest*[3], Samantha Fritz*[1], Kody Willis[2], Karl Blumenthal[2], and Thomas Padilla*[2]

[1]University of Waterloo [Waterloo] – 200 University Avenue West, Waterloo, ON, Canada N2L 3G1, Canada
[2]Internet Archive – United States
[3]York University [Toronto] – 4700 Keele Street, Toronto, ON, Canada M3J 1P3, Canada

Abstract

Panel Abstract and Overview

Panel Chair and Moderator: Ian Milligan

In 2020, the Andrew W. Mellon Foundation funded the second phase of the "Archives Unleashed Project" with the "Integrating the Archives Unleashed Cloud with Archive-It" project. Our vision was to integrate the analytics approaches developed by Archives Unleashed – both our command-line-based WARC analysis platform, the Archives Unleashed Toolkit, as well as our GUI Archives Unleashed Cloud – with the existing services at the Archive-It Research Services program. The goal was to generate a truly end-to-end service that would enable the collection, preservation, access, analysis, and researcher use of web-published materials. This would also help ensure the long-term sustainability of infrastructure.

How time flies! This project is now in its third and final year. This RESAW panel, "Building the Next Generation of Web Archive Analysis Service," provides an overview of the current technical state of the project, the outreach activities being conducted through the Archives Unleashed Cohort program, and concludes with comments on the "next steps/future" of the project. As our goal is to serve researchers well represented by the RESAW community, we see this panel both as an opportunity to discuss our project but also to discuss with the audience future directions and how we can best collaboratively serve web researchers.

The panel will be moderated by Ian Milligan, who will provide brief framing remarks before introducing each of the three presentations. We will then facilitate a conversation with the audience.

Presentation One: Technical Considerations for Building an Arch from Preservation to Research

Authors: Helge Holzmann and Nick Ruest

*Speaker

Over the past five years, both the Internet Archive (IA) and the Archives Unleashed (AU) team have independently and collaboratively developed a suite of tools to address the challenges and barriers of working with web archives at scale.

Powered by the Archives Unleashed Toolkit (AUT), the Archives Unleashed Cloud was the first attempt at providing a self-service web-based platform for conducting analysis on WARC collections. Drawing on the computational methods and power of AUT, the platform provided familiar click-to-results actions that many of us prefer, eliminating the technical burdens of learning and running the AU Toolkit locally.

At the Internet Archive's subscription service Archive-It, pre-defined derivative datasets could be requested for any web archive collection. The Archive-It Research Services (ARS) included WAT (extended web archive metadata format), WANE (named entities), LGA (temporal graphs). ARS datasets used to be manually generated at request by AIT staff, using distributed computing cluster technology, such as Hadoop and PIG.

The next step was to marry the Archives Unleashed approach with that of the Internet Archive.

Enter the Archives Research Compute Hub (ARCH). ARCH is a jointly-developed platform incorporating the previously described approaches. It is backed by its own distributed computing cluster, based on Hadoop, with new computing hardware dedicated to ARCH, and its own distributed storage (HDFS). Using the new infrastructure, the team developed a web interface and advanced job management features to control ARCH jobs.

ARCH supports 16 dataset derivation jobs. This includes all of the jobs from the AU Toolkit and Cloud as well as the 3 ARS jobs, which have all been ported into a modern Spark-compatible framework. The jobs are now much more efficient than either earlier platform, and can be launched by researchers themselves, benefiting from ARCH's advanced features, such as its queuing system, web preview and dataset sharing options.

ARCH is based on the IA Sparkling library, a Spark-based toolkit used at IA for large-scale data processing jobs, with a focus on web archives but also archival collections more generally. As part of this project, both ARCH and Sparkling have been open-sourced to fulfill the open access policy of this project and support interoperability. Further, we pursue open standards, such as a WASAPI-compatible data endpoint to provide derivative datasets in a standard API format, which is particularly useful for dataset types consisting of more than one file.

To account for the distributed infrastructure within IA, which hosts the ARCH service, and in order to provide easy access to its data, ARCH supports full cross-cluster access. This enables direct access to AIT's Hadoop cluster, which is used as a long-term cache for AIT collections and allows for efficient access without loading them from IA main storage system Petabox, if available. Otherwise, collections will be seamlessly loaded from IA's Petabox and cached by ARCH for consecutive job runs in a managed space, keeping data fresh and available. AIT and Petabox API's have been incorporated for data statistics before files have been completely fetched.

These considerations and details have contributed to this novel, deeply-integrated self-service compute hub at IA.

Presentation Two: Archives Unleashed Cohort Program: Opportunities to Access, Explore, Engage with Web Archives

Authors: Samantha Fritz, Kody Willis, Karl Blumenthal

Following a successful series of datathon events (2017-2020), the Archives Unleashed project launched the cohort program (2021-2023) to facilitate opportunities to improve access, exploration and research engagement with web archives.

Borrowing from the hacking genre of events often found within the tech industry, Archives Unleashed datathons were designed to provide an immersive and uninterrupted period of time for participants to work collaboratively on projects and gain hands-on experience working with web archive data. The datathon series cultivated community formation and a sense of belonging among participants, and at the same time, these events empowered scholars to build confidence and the skills needed to work with web archives. However, the short-term nature of datathons ultimately saw focused energy and time to research projects diminish once physical meetings concluded.

Launched in 2021, the Archives Unleashed cohort program was developed as an extension and maturing of the datathon model to support a progressive evolution of research projects. The program ran two iterative cycles of five projects in each of 2021-2022 and 2022-2023. Collectively, the program saw the international participation of 46 researchers from 21 unique institutions. Programmatically, researchers engage in a year-long collaboration project, with web archives featured as a primary data source, and groups received a modest research grant and technical support. A defining feature of the program was the mentorship model, which included direct one-on-one consultation from the Archives Unleashed team, connections to field experts, and opportunities for peer-to-peer support.

The core objective of the Archives Unleashed Project is to lower the barriers of entry for conducting scalable research with web archives. The cohort program advances this goal by continuing to create opportunities for access, exploration, and engagement with web archives research.

As a first step, we recognize that access to data is critical - without access, there is no use. The preservation of web archives over two decades has resulted in an abundance of data, yet, web archives generally remain an untapped resource for research. Our cohort participants further highlight this unrealized potential, as most researchers did not have direct access to web archival collections. The cohort program relied on the ARCH platform's infrastructure and the collaborative spirit of Archive-It institutional subscribers to bridge the gap between researchers and data. As initial pilot users of ARCH, researchers had the unique opportunity to directly access web archive collections for further analysis.

Further, the mentorship model of the cohort program directly impacted and created opportunities for the exploration of web archive collections. While project teams were often interdisciplinary, with a wide variance of expertise and skills, many of these researchers had not previously worked with web archive data. One-on-one research consultations provided an environment to encourage and support the development of skills and knowledge. This ultimately translated into a deeper emersion with web archives, for instance, groups adopted new tools and methods, as well as refined their conceptual understanding of the objects they were investigating.

Finally, several opportunities to engage with web archives contributed to community building with a wide range of audiences. Among the broader cohort group, peer-to-peer connection supported knowledge sharing and expanded the capacity for working with and learning from others in different fields and disciplines. Engaging with web archive collections also afforded opportunities to build relationships between researchers and curators, which ultimately highlights the possibility of how web archives can be investigated and increase the visibility of institutional holdings.

In building the next generation of web archive analysis services, the cohort program demonstrates at a micro level the value of expanding opportunities for accessing, exploring, and engaging with web archives.

Beyond Web Archives: Future Directions

Author: Thomas Padilla

For the past decade cultural heritage organizations the world over have sought to support computational research and pedagogy with cultural heritage collections. Efforts are diverse spanning national libraries, archives, and museums, consortia, and small and medium universities. Specific examples include but are not limited to the Hathitrust Research Center, National Library of Scotland Data Foundry, and the University of North Carolina at Chapel Hill's On the Books: Jim Crow and Algorithms of Resistance. ARCH joins this pool of effort with an initial focus on supporting computational research with web archive collections. Moving forward ARCH will expand scope, leveraging the Internet Archive's non-profit owned infrastructure and open source software to facilitate responsible computational use of all digital collection content types (e.g., text, video, audio, images). Furthermore, with a diverse set of partners the Internet Archive will work to enable aggregation and/or ingest of multi-institutional digital collections for use with ARCH.

By expanding ARCH to support work with more digital collection content types and enabling a multi-institutional collection aggregation and/or ingest function ARCH aims to do the following:

> Support a broad disciplinary community with a range of collection content types - text, image, audio, video and more;
>
> Foster an inclusive path to institutional participation in supporting computational research that is sensitive to variation in institutional capacity - success in this space cannot remain the province of primarily well-resourced institutions; and
>
> Enable the creation of a multi-institutional multi-content type research corpus that saves the time of researchers and cultural heritage workers at scale - streamlining researcher access and reducing resource intensive replication of effort.

Keywords: Internet Archive, infrastructure, researcher use, digital humanities