Arch-It

Helge Holzmann[1] , Nick Ruest[2] , Jefferson Bailey[1] , Alex Dempsey[1] , Samantha Fritz[3] , Ian Milligan,[3] , and Kody Willis[1]

[1] Internet Archive
[2] Digital Scholarship Infrastructure Department, York University
[3] Department of History, University of Waterloo

ABSTRACT

Over the past quarter-century, web archive collection has emerged as a user-friendly process thanks to cloud-hosted solutions such as the Internet Archive's Archive-It subscription service. Despite advancements in collecting web archive content, no equivalent has been found by way of a user-friendly cloud-hosted analysis system. Web archive processing and research require significant hardware resources and cumbersome tools that interdisciplinary researchers find difficult to work with. In this paper, we present ARCH (Archives Research Compute Hub)[1], an interactive interface, closely connected with Archive-It, engineered to provide analytical actions, specifically generating datasets and in-browser visualizations. It efficiently streamlines research workflows while eliminating the burden of computing requirements. Building off past work by both the Internet Archive (Archive-It Research Services) and the Archives Unleashed Project (the Archives Unleashed Cloud), this merged platform achieves a scalable processing pipeline for web archive research.

1 INTRODUCTION

While collecting web archive content has matured into a user-friendly process, thanks in no small part to cloud-hosted solutions such as the Internet Archive's Archive-It service, this ease-of-use has not been matched on the analysis side. We accordingly need a user-friendly system that can enable the creation of research datasets from web archives so that researchers can work with material at scale.

In this paper we present an overview of the Archives Research Compute Hub (ARCH) as opposed to our more robust examination of the platform [2]. As this paper is intended as a demonstration at WADL, we draw heavily on that paper. ARCH is a production system tightly integrated with the Internet Archive infrastructure and services. It grew out of the Archives Unleashed Cloud: a proof-of-concept platform that demonstrated the ability of a web browser-based system to power backend Apache Spark-driven jobs on web archival datasets [6]. Powered by the Archives Unleashed Toolkit and the Internet Archive's Sparkling data processing library[2], the ARCH platform will become a complementary component of the Internet Archive's Archive-It system.

The Archives Unleashed project aims to address this problem [7] by being for web archive analysis as Archive-It is for web archive capture: powerful, scalable, and above all, accessible and intuitive for users. The Archives Unleashed Cloud (2017-2020) provided user access to the features of the Archives Unleashed Toolkit in a cloud-hosted environment [6]. The Cloud worked with Archive-It collections, using APIs to transfer data from the Internet Archive to Compute Canada cloud-hosted infrastructure. Yet the initial approach of having a separate analysis service presented shortcomings. When a user wished to carry out analysis, data had to be transferred. More importantly, connections between Archive-It and the Cloud required a complicated interplay of APIs, bulk data transfers, and other workflows, leaving a separate analysis service vulnerable to network disruptions or changing standards. These factors combined to make it an interesting proof-of-concept but one that presented considerable sustainability challenges. Our goal, then, was to integrate Archives Unleashed tools with the Internet Archive's Archive-It service.

2 RELATED WORK AND PROJECT CONTEXT

Established in 2017, the Archives Unleashed project recognizes the collective need among researchers, librarians and archivists for analytical tools, community infrastructure, and accessible web archival interfaces. To this end, the project aspires to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past. Between 2017 and 2020, the project focused on developing the "Archives Unleashed Cloud," a web-based interface for working with web archives at scale using the Archives Unleashed Toolkit and Apache Spark [6]. This work built on the project's long-standing interests in building exploratory search interfaces for web archive collections [3]. Similar noteworthy work includes the SolrWayback project from The Royal Danish Library. Combining Apache Solr with OpenWayback or pywb, SolrWayback provides search and discovery of web archive collections, as well as replay, and a number of analysis and visualization features [5].

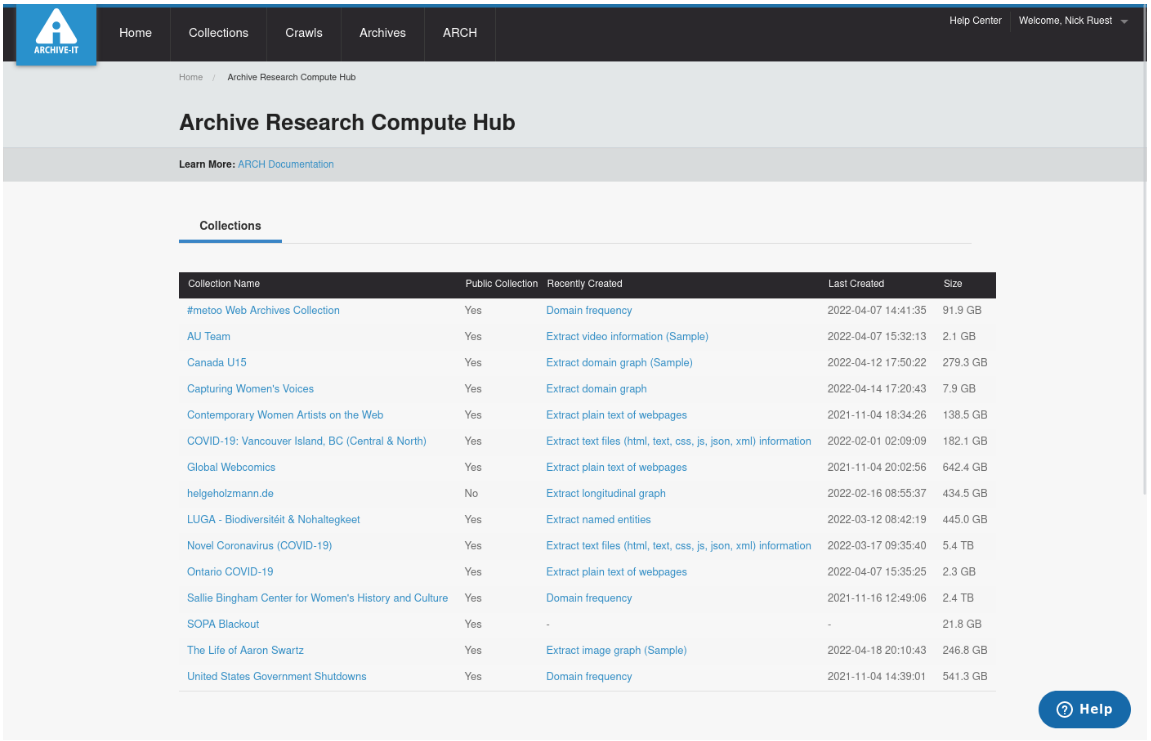
In 2020, the project's first phase was completed. The next phase involved exploring integration and collaboration with the Internet Archive [7]. We were influenced by the global adoption of the Internet Archive's Archive-It subscription service and the stability of the Apache Spark platform [1].

Since the launch of the Internet Archive's subscription service in 2006, over 700 institutions from 23 countries have used Archive-It to preserve over two petabytes of data consisting of over 40

[1] https://github.com/internetarchive/arch

[2] https://github.com/internetarchive/Sparkling

WADL'22, June 20, 2022, Cologne, Germany
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



Archive Research Compute Hub

Learn More: ARCH Documentation

Collection Name	Public Collection	Recently Created	Last Created	Size
#metoo Web Archives Collection	Yes	Domain frequency	2022-04-07 14:41:35	91.9 GB
AU Team	Yes	Extract video information (Sample)	2022-04-07 15:32:13	2.1 GB
Canada U15	Yes	Extract domain graph (Sample)	2022-04-12 17:50:22	279.3 GB
Capturing Women's Voices	Yes	Extract domain graph	2022-04-14 17:20:43	7.9 GB
Contemporary Women Artists on the Web	Yes	Extract plain text of webpages	2021-11-04 18:34:26	138.5 GB
COVID-19: Vancouver Island, BC (Central & North)	Yes	Extract text files (html, text, css, js, json, xml) information	2022-02-01 02:09:09	182.1 GB
Global Webcomics	Yes	Extract plain text of webpages	2021-11-04 20:02:56	642.4 GB
helgeholzmann.de	No	Extract longitudinal graph	2022-02-16 08:55:37	434.5 GB
LUGA - Biodiversitéit & Nohaltegkeet	Yes	Extract named entities	2022-03-12 08:42:19	445.0 GB
Novel Coronavirus (COVID-19)	Yes	Extract text files (html, text, css, js, json, xml) information	2022-03-17 09:35:40	5.4 TB
Ontario COVID-19	Yes	Extract plain text of webpages	2022-04-07 15:35:25	2.3 GB
Sallie Bingham Center for Women's History and Culture	Yes	Domain frequency	2021-11-16 12:49:06	2.4 TB
SOPA Blackout	Yes	-	-	21.8 GB
The Life of Aaron Swartz	Yes	Extract image graph (Sample)	2022-04-18 20:10:43	246.8 GB
United States Government Shutdowns	Yes	Domain frequency	2021-11-04 14:39:01	541.3 GB

Figure 1: ARCH main collections page.

billion born-digital, web-published records in over 12,000 public collections. It is a successful service. A survey by the National Digital Stewardship Alliance reported that by 2017, 94% of surveyed institutions were using Archive-It to preserve web material – and an additional 4% were using other services provided by the Internet Archive [4]. Archive-It is thus effectively the de-facto platform for web archiving, used by nearly all Association of Research Library members, hundreds of other higher education, memory institutions, public libraries, governments, and non-profit organizations.

Despite this widely-accepted solution for the capture of web material, the problem of analysis remains. By this, we refer to at-scale explorations of data that require more than the replay interface of the Wayback Machine. While web archive data is captured and preserved in the ISO-standard WARC file format, the formation of a scholarly ecosystem around web archive analysis has been slow.[3]

3 ARCHIVES RESEARCH COMPUTE HUB

In this section, we present our interface and its broader context within Archive-It. As of December 2021, ARCH has both feature parity with the earlier Archives Unleashed Cloud, and also additional functionality to generate several additional datasets. As functionality from the earlier Cloud was ported, all features were redesigned and reimplemented. We addressed known issues, fixed existing bugs, and more importantly, implemented an approach that scales to meet our needs.

3.1 Design Considerations

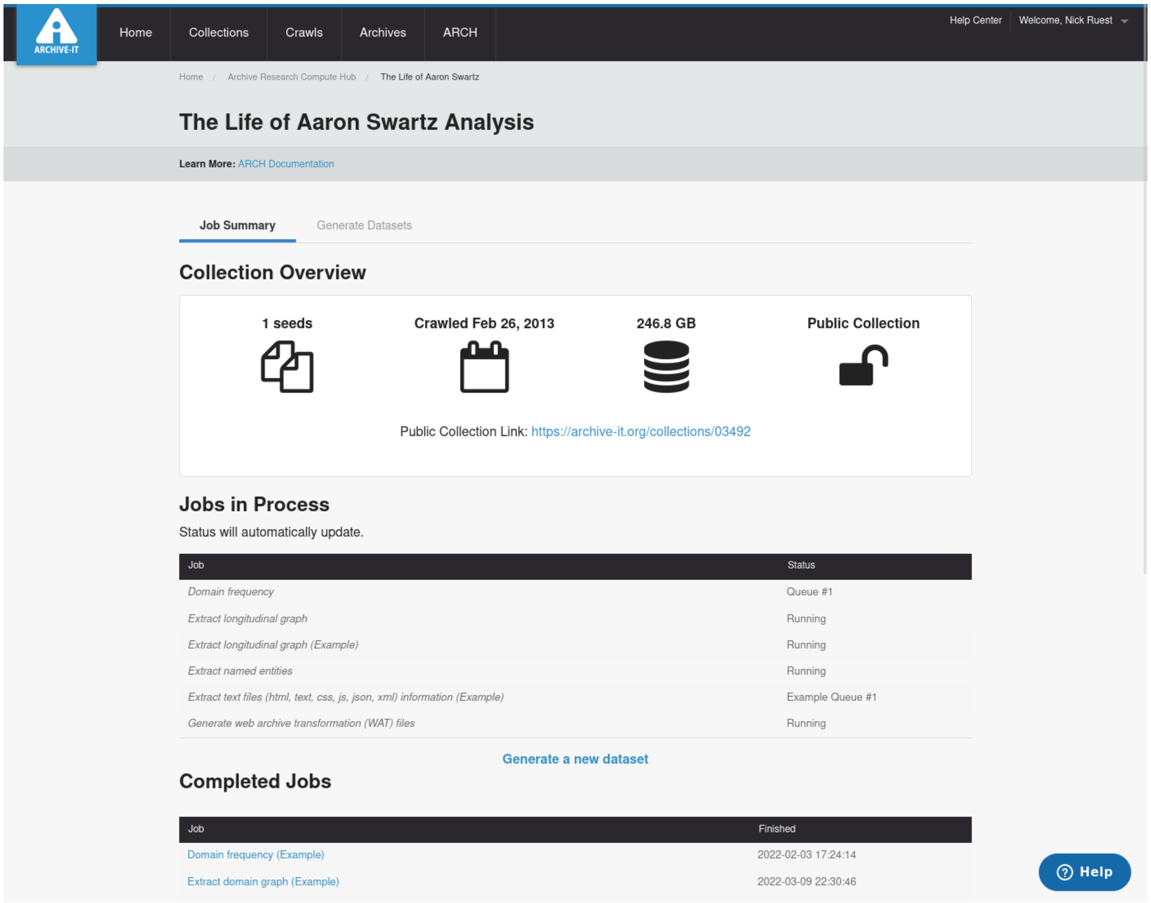
ARCH now runs on an infrastructure that is physically connected to Archive-It servers and computing infrastructure, mitigating the need to copy data before processing. As not all Archive-It data is kept in its dedicated computing cluster, ARCH is connected to the Internet Archive's long-term storage system (the "Petabox") to fetch missing data. In addition, we implemented a smart caching mechanism to avoid re-fetches for consecutive access to the same data. Cognizant of researcher needs beyond Archive-It collections, we also support custom collections which can be located on ARCH's own cluster.

Given the sensitive nature of web archival collections, we have implemented a user and permissions system. There are two authentication providers: Archive-It user accounts and dedicated ARCH users. For Archive-It users, we rely on Archive-It's internal permissions process. We have also implemented a permission control access that allows ARCH and Archive-It users to cross-access additional Archive-It collections (pending permission from the data collector) and ARCH custom collections.

To control jobs and enable the downloading of files via different tools (browser-based downloads for smaller files, command line for larger ones), we provide multiple APIs and authentication methods. While the actual implementation details are beyond the scope of this paper, ARCH is a native Scala application built using Scalatra[4]. The underlying toolkit is based on the Archives Unleashed Toolkit (previously known as Warcbase) as well as the Internet Archive's Sparkling library. Jobs and queues are controlled via APIs,

[3] The best place to learn about available tools is the "Web Archiving Awesome List" maintained by the International Internet Preservation Consortium and researchers across the field. See https://github.com/iipc/awesome-web-archiving.

[4] https://scalatra.org/



The Life of Aaron Swartz Analysis

Learn More: ARCH Documentation

Collection Overview

1 seeds — Crawled Feb 26, 2013 — 246.8 GB — Public Collection

Public Collection Link: https://archive-it.org/collections/03492

Jobs in Process

Status will automatically update.

Job	Status
Domain frequency	Queue #1
Extract longitudinal graph	Running
Extract longitudinal graph (Example)	Running
Extract named entities	Running
Extract text files (html, text, css, js, json, xml) information (Example)	Example Queue #1
Generate web archive transformation (WAT) files	Running

Completed Jobs

Job	Finished
Domain frequency (Example)	2022-02-03 17:24:14
Extract domain graph (Example)	2022-03-09 22:30:46

Figure 2: ARCH job summary page.

enabling Spark jobs to be chained with post-processing jobs, as well as separate queues for example/full jobs, Spark operations, and post-processing.

3.2 User Interface

ARCH's interface consists of four levels. These guide users to interact with their collections by generating datasets for analysis and engaging with in-browser features. The goal of ARCH is to provide an efficient, streamlined workflow without burdening users with computing requirements or actions.

The first level is the **main collections** page. All of a user's Archive-It collections are presented in a table (Figure 1), accompanied by information about the most recent analysis conducted and other collection-based metadata. Each collection title provides an access point for conducting analysis.

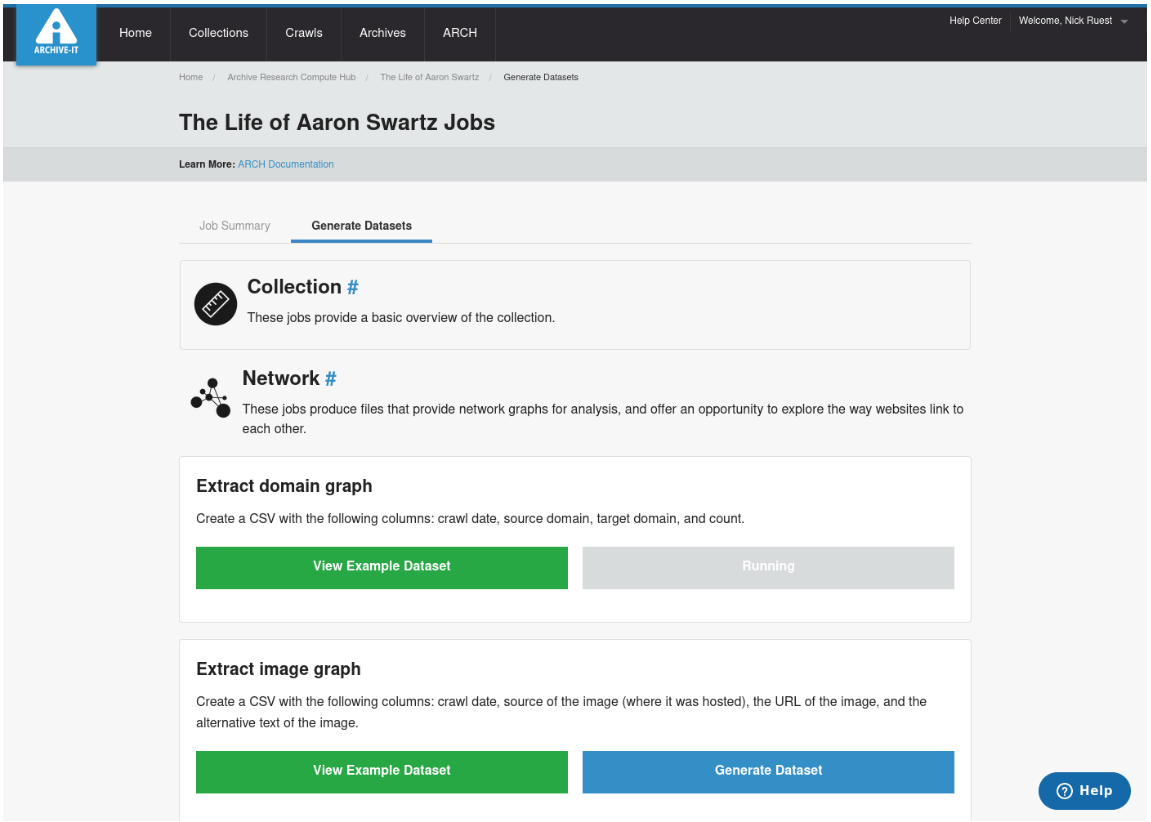
The second is a **job summary** page, where users can generate, download, and monitor derivative datasets. An overview of the collection identifies basic metadata about the collection, including collection size and whether it is a public or private collection. The second main feature of this space provides tables that summarize "Jobs in Process" - the stage and queue of any current jobs being run - and a "Completed Jobs" table identifying all datasets previously generated, noting an accompanying date/time stamp (Figure 2).

The third level is the **generation of datasets** (Figure 3). As a core feature of ARCH, users can generate sixteen different datasets for scholarly exploration. These datasets are categorized into four main themes of analysis (Table 1). This supports different dataset generation jobs based on a generic interface to start jobs, monitor their status, and explore the ensuing output.

Finally, the last level are the **derivative dataset** pages themselves. For each dataset generated, users can access an overview

Dataset Category	**Description**
Collection	Offers an overview of a collection by looking at simple statistical counts.
Network	Produces files that provide network graphs for analysis and offer an opportunity to explore the way websites link to each other.
Text	Allows the user to explore text components of a web archive, including extracted "plain text" HTML, CSS, and other web elements.
File formats	Provides files that contain information on certain types of binary files found within a web archive.

Table 1: ARCH Datasets



The Life of Aaron Swartz Jobs

Learn More: ARCH Documentation

Collection #

These jobs provide a basic overview of the collection.

Network #

These jobs produce files that provide network graphs for analysis, and offer an opportunity to explore the way websites link to each other.

Extract domain graph

Create a CSV with the following columns: crawl date, source domain, target domain, and count.

Extract image graph

Create a CSV with the following columns: crawl date, source of the image (where it was hosted), the URL of the image, and the alternative text of the image.

Figure 3: The "generate datasets" page in the ARCH interface.

page of the dataset, which provides metadata (file name, file size, results count, and date completed), download options, a preview of up to 100 lines, and the option to re-run any job. An example of this can be seen in Figure 8. Where possible, in-browser visualization and charts present a summary of the data. For instance, the *extract web graph* dataset page offers an interactive network graph that users can explore using simple functionalities like zooming in and out on modes and clusters and exporting a high-resolution image. These datasets are intended be downloaded and further explored with other analytical tools and methods.

4 CONCLUSION AND EVALUATION

We have presented ARCH, the Archives Research Compute Hub, a novel data processing platform for web archives, closely integrated with Archive-It.

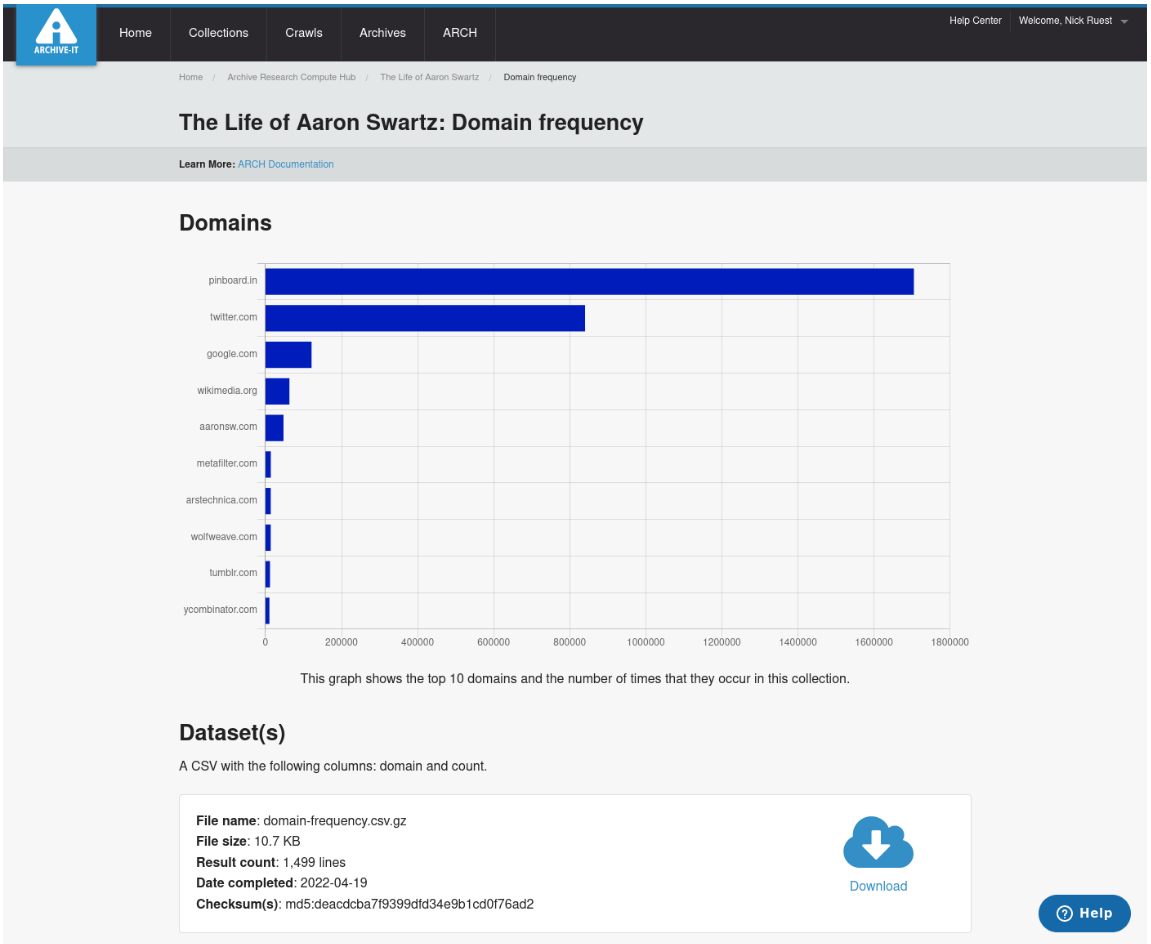
The design process for ARCH involved a variety of interconnected stages, from designing wireframes to building infrastructure to connecting backend processes to the user interface. User experience (UX) evaluations were essential for measuring and understanding the needs of researchers. As such, the team conducted iterative and multi-staged user testing and surveying to assess user needs and experience. By engaging with Archive-It power users and Archives Unleashed Cloud alumni in five closed user testing rounds, our team gathered feedback and initial impressions of ARCH. Testing was primarily conducted through surveys, which collected qualitative and quantitative data to determine user satisfaction and experience.

Findings from the survey were translated into actionable tickets to provide action-based tasks for development cycles. We were able to implement the majority of action items, with some needing further planning and only a few that fell outside of our scope of work.

As a multi-stage UX testing process, each subsequent round of testing served as another opportunity to review and refine impressions of prior development and enhancements — improving our accuracy and capacity to match user needs at each stage. Our final rounds of testing concluded in early 2022. This final process served two purposes. First, we expanded testing to include a larger group (approximately 100 participants) to serve as a stress test. As this was our largest testing group to date, this offered an opportunity to verify ARCH's robustness, capacity, and efficiency while noting any bottlenecks or areas for improvement. Second, we conducted focused interviews with a small group of researchers who have extensively used ARCH since August 2021. These researchers were ideal for understanding the real-life application and use cases of the web archives research journey.

ACKNOWLEDGEMENTS

This research was supported by the Andrew W. Mellon Foundation's Public Knowledge program, the Social Sciences and Humanities Research Council of Canada, as well as Start Smart Labs, the University of Waterloo, and York University.



The Life of Aaron Swartz: Domain frequency

Learn More: ARCH Documentation

Domains

This graph shows the top 10 domains and the number of times that they occur in this collection.

Dataset(s)

A CSV with the following columns: domain and count.

File name: domain-frequency.csv.gz
File size: 10.7 KB
Result count: 1,499 lines
Date completed: 2022-04-19
Checksum(s): md5:deacdcba7f9399dfd34e9b1cd0f76ad2

Figure 4: One of ARCH's Dataset Results Pages.

REFERENCES

[1] Helge Holzmann, Vinay Goel, and Avishek Anand. 2016. ArchiveSpark: Efficient Web Archive Access, Extraction and Derivation. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. ACM, Newark New Jersey USA, 83–92. https://doi.org/10.1145/2910896.2910902

[2] Helge Holzmann, Nick Ruest, Jefferson Bailey, Alex Dempsey, Samantha Fritz, Peggy Lee, and Ian Milligan. 2022. ABCDEF - The 6 key features behind scalable, multi-tenant web archive processing with ARCH: Archive, Big Data, Concurrent, Distributed, Efficient, Flexible. *International Journal of Digital Humanities*. https://doi.org/10.1145/3529372.3530916

[3] Andrew Jackson, Jimmy Lin, Ian Milligan, and Nick Ruest. 2016. Desiderata for Exploratory Search Interfaces to Web Archives in Support of Scholarly Activities. (2016). https://doi.org/10.1145/2910896.2910912 Accepted: 2016-05-09T11:57:37Z.

[4] Katherine Kim, Wayne Graham, Paige Walker, National Digital Stewardship Alliance (NDSA), Carol Kussmann, and Aliya Reich. 2018. 2017 Web Archiving in the United States - A 2017 Survey. (Oct. 2018). https://doi.org/10.17605/OSF.IO/3QH6N Publisher: OSF.

[5] The Royal Danish Library. 2021. SolrWayback. https://github.com/netarchivesuite/solrwayback

[6] Nick Ruest, Samantha Fritz, Ryan Deschamps, Jimmy Lin, and Ian Milligan. 2021. From archive to analysis: accessing web archives at scale through a cloud-based interface. *International Journal of Digital Humanities* 2, 1 (Nov. 2021), 5–24. https://doi.org/10.1007/s42803-020-00029-6

[7] Nick Ruest, Jimmy Lin, Ian Milligan, and Samantha Fritz. 2020. The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (JCDL '20)*. Association for Computing Machinery, New York, NY, USA, 157–166. https://doi.org/10.1145/3383583.3398513