Archives Unleashed

Summer 2020 Newsletter

Welcome to the summer edition of the Archives Unleashed Newsletter!

Special Announcement

The [Andrew W. Mellon Foundation](#) has awarded a $1,084,087 CAD grant to the University of Waterloo to support the "Integrating Archives Unleashed Cloud with Archive-It" project. Led by Professor [Ian Milligan](#), from the University of Waterloo's Department of History, alongside co-investigators [Jimmy Lin](#), Professor and Cheriton Chair at Waterloo's Cheriton School of Computer Science, [Nick Ruest](#), Digital Assets Librarian in the Digital Scholarship Infrastructure department of York University Libraries, and Jefferson Bailey, Director of Web Archiving & Data Services at the Internet Archive, this project represents the next stage of the [Archives Unleashed](#) Project. With this funding from The Andrew W. Mellon Foundation, the project hopes to bridge the current gulf between web archiving collection, access, and data-driven analysis.

[Read more about our new project's phase here.](#)

We look forward to sharing additional details over the coming weeks and

months, and continuing to work with the web archiving community.

What's Been Happening? ⚙️

The majority of our work has focused on wrapping up tasks from the first phase of the project:

- Nick Ruest has been hard at work to bring the community the **Archives Unleashed Toolkit** [0.80.0](#) **release**.
- We've added **Spark 3.0.0 and Java 11 support** to the Toolkit, which is available in the main branch of the project.
- Jimmy Lin has continued to offer guidance for development cycles and has been a coordinator with Ian Milligan of our team's **scholarly activities**.

Led by Samantha Fritz as first author, our team has been working on **two manuscripts:**

- "[Building community at distance: a datathon during COVID-19](#)", reflects on the **roles of digital libraries during the COVID-19** pandemic, specifically looking at the implications of transitioning events to virtual platforms. It has recently been published as an open-access article in a special issue of [Digital Library Perspectives](#) on "[Digital libraries and COVID-19: responding to a global emergency](#)."

- The second piece reflects on **how our project fostered community engagement** through collaborative events and activities, while proposing a community engagement model that can be adapted by the broader web archiving community. We look forward to sharing this pieces with the community over the next few months.

Featured Articles

Datathon Report

Datathons combine the terms "data" and "marathon" and are generally short, intensive events where individuals have access to datasets to solve, answer, and/or investigate questions.

Our team reflects on the moments and individuals who've contributed to these four successful datathons as part of our Archives Unleashed Project.

(Read More...)

Community Report

Supported by the Andrew W. Mellon Foundation, the Archives Unleashed Project launched in 2017 with the aim of making petabytes of historical internet content accessible to scholars and others interested in researching the recent past.

We are pleased to share a summary of our activities and milestones through our Community Report.

(Read More...)

spAUTlight

Here's a quick overview of Archives Unleashed Project activities and accomplishments between 2017 and 2020.

Archives Unleashed Project
spAUTlight 2017 - 2020

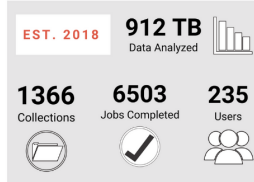*Making petabytes of historical internet content accessible*

Archives Unleashed Tool Development

Archives Unleashed Cloud

EST. 2018 | **912 TB** Data Analyzed

1366 Collections
6503 Jobs Completed
235 Users

Cloud User Distribution

59 Unique Institutions

10 Countries

Archives Unleashed Toolkit

EST. 2017 | **15** Releases

37 Implemented Enhancements
211 Merged Pull Requests
137 Closed Issues

Community Engagement

Newsletter
202 Subscribers
41 Blog Posts
8 Issues

Learning + Resources
35 Datasets
7 Learning Guides
5 Videos

Scholarly Activities
17 Conference Presentations
8 Invited Talks/Lectures

Publications
7 Conference Publications
4 Conference Posters & Demos
3 Journal Articles
1 Book

Datathons & Scholarly Activites

Archives Unleashed Datathons
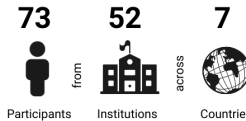
April 2018
University of Toronto

November 2018
Simon Fraser University

March 2019
George Washington University

March 2020
Columbia University

73 Participants
from
52 Institutions
across
7 Countries

(click image to zoom in)

Project Resources 👓

Over the past three years, our team has developed several resources that can be used in conjunction with our tools and platforms. Here's a quick recap of all the resources currently available to you.

Introducing the Archives Unleashed Project

The Archives Unleashed Project

Making Historical Internet Content Accessible

Learn about our project's beginnings and the different ways you can use the Archives Unleashed Toolkit in this [short video](#).

Archives Unleashed Tools and Platforms

Archives Unleashed Toolkit

An open-source platform for analyzing web archives built on [Apache Spark](#), which provides powerful tools for analytics and data processing.

- Current Release: [0.80.0](#)
- [User Documentation](#)
- [Introduction to the Toolkit](#)
- [Github Page](#)

Archives Unleashed Cloud

The Archives Unleashed Cloud is a one-stop portal that allows scholars to access their collections and execute a number of analyses.

- [Learning guides](#) walk through using Cloud derivatives for further analysis
- [User Documentation](#)
- [Cloud Tour](#)
- [Github Page](#)

Archives Unleashed Notebooks

Example notebooks for working with web archives with the Archives Unleashed Toolkit and derivatives generated by the Archives Unleashed Toolkit.

- [Github Page](#)
- Parquet and PySpark example notebooks are available for use
- We also used notebooks for our recent New York datathon, using [Colaboratory](#)

Warclight

Warclight is a [Project Blacklight](#) based [Rails engine](#) that supports the discovery of web archives held in the WARC and ARC formats. It is designed to work with web archive data that is indexed via the UK Web Archive's [webarchive-discovery](#) project.

- [Github Page](#)
- [Demo Site](#) - explore faceted full-text search, record view, and other advanced discovery options
- [Documentation](#) will help to get you started on your own Warclight application
- Featured posts appear through our Medium blog:
    - [See a Little Warclight](#)

- Warclight and WALK: Providing a Framework for Search and Discovery of Web Archives
- Thumbnails in Warclight

Archives Unleashed Engagement

Archives Unleashed Datathons

Datathons provided an opportunity to build community around Archives Unleashed tools, scholarly discussion, and training for scholars with limited technical expertise to explore archived web content.

Final team presentations are available: Toronto, Vancouver, Washington, and New York

We also have opening presentations and materials that were used during the 2020 New York online datathon:

- Video
- Presentation slides
- Archives Unleashed Quick Guide: Setting up Colab Notebooks

Communications and Publications

Our team has connected with the community through a variety of platforms, including Slack, Twitter, our quarterly newsletter, and our Medium blog.

We've written over 40 blog posts, available through our Medium blog.

We also have a full publications list that includes all of our scholarly work, including pre-prints, journal articles, peer-reviewed conference materials, and invited talks and lectures.

In collaboration with institutions, we have also made over 35 datasets available via the Web Archives for Historical Research Group on Zenodo and Dataverse.

Words of Thanks

As we round out the first phase of the Archives Unleashed Project, we would like to extend our sincere thanks to our advisory board members who have given us invaluable guidance throughout the project. Many thanks to Jefferson Bailey, Nathalie Casemajor, Robert H. McDonald, Matthew Weber, Michele Weigle, Nicholas Worby,

We would also like to thank our community - all of you who have supported the project through your contributions, participation, and enthusiasm. We can't wait to continue our work with you!

The work done by the Archives Unleashed Project has been primarily supported by The Andrew W. Mellon Foundation. Other financial and in-kind support was given by the Social Sciences and Humanities Research Council, Compute Canada, the Ontario Ministry of Research, Innovation, and Science, York University Libraries, Start Smart Labs, and the Faculty of Arts and David R. Cheriton School of Computer Science at the University of Waterloo.

Get Involved

Interested in getting involved with the Archives Unleashed Project? Connect with our team and help grow our community

Follow us on Twitter
Join our Slack group
Participate on Github
Subscribe to our newsletter

Connect and tell us how you've used our tools and platforms

Share our news with colleagues and friends

Twitter   GitHub   Website   Email

The Archives Unleashed Project aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past. Supported by a grant from the Andrew W. Mellon Foundation, the project will develop web archive search and data analysis tools to enable scholars and librarians to access, share, and investigate recent history since the early days of the World Wide Web.

The Archives Unleashed Project
200 University Avenue West | Waterloo, ON N2L 3G1