Welcome to the Archives Unleashed Project Community Report!

Our team is excited to share a summary of activities and outcomes that have been completed by the Archives Unleashed Project between 2017 and 2020.

We sincerely appreciate your support, participation, and enthusiasm as our team has worked towards lowering barriers to access for working with web archives.
The Archives Unleashed Project recognizes the critical role that web archives play for scholars studying the 1990s onward.

The project grew out of a series of datathons held between 2016-2017, which identified the need for better analytics tools, community infrastructure, and accessible web archival interfaces.

In 2017, the Archives Unleashed Project was awarded a grant from the Andrew W. Mellon Foundation.

From 2017 - 2020 the project has developed tools, resources, and hosted datathons to bridge an accessibility and usability gap to web archives.
The Archives Unleashed Project

Project Team

Ian Milligan  
History  
University of Waterloo  
Primary Investigator

Nick Ruest  
Library  
York University  
Co-Investigator

Jimmy Lin  
Computer Science  
University of Waterloo  
Co-Investigator

Samantha Fritz  
AU Project  
University of Waterloo  
Project Manager
The Archives Unleashed Project

Collaborators

Advisory Board

Jefferson Bailey
Nathalie Casemajor
Robert H. McDonald
Matthew Weber
Michele Weigle
Nicholas Worby

Contributors

Ryan Deschamps
Sarah McTavish
Rebecca MacAlpine
Jeremy Wiebe
Gursimran Singh
Boris Lin
Joseph Zhou
Titus An
Billy Jin
Daniel Hopper
The Archives Unleashed Project

Goals

The Archives Unleashed Project is dedicated to lowering barriers to access for working with web archives at scale, with a focus in three areas:
The Archives Unleashed Project

Goals

- We know that access remains a significant barrier in the use of Web archives.
- By access/accessibility, we refer to the ability to make use of something, or capability of being reached, used, understood or appreciated.
- Archives Unleashed incorporates the spirit of access/accessibility by:
  a. Providing multiple access points for exploring web archives via development of the Toolkit, Cloud, and additional platforms;
  b. Ensuring Tools and platforms are created as user friendly as possible; and
  c. Creating documentation and resources to support training and learning.
The Archives Unleashed Project

Goals

- Building community has been a vital component and goal of the project.

- Archives Unleashed datathon events have been a primary activity through which we’ve developed a community of users around our tools.

- Our team has also invested and participated in the wider web archival community through additional scholarly activities, such as institutional collaborations, conferences, and meetings.
The Archives Unleashed Project

Goals

- Sustainability planning speaks to the long term lifecycle of the project.

- The primary goal is to ensure a project’s survival and continued efforts once the grant cycle has ended.

- Archives Unleashed has developed tools and platforms with sustainability in mind, specifically by adopting widely adopted and stable programming languages and best practices.

- Our team has also engaged in collaboration and partnerships with various institutions.
The Archives Unleashed Project

Deliverables

Aligned with the aforementioned goals, the Archives Unleashed Project outlines three main deliverables under the 2017-2020 Mellon funding:

- Toolkit
- Cloud
- Datathons
The Archives Unleashed Project

Deliverables

1. Develop the Archives Unleashed Toolkit
   ● Apply modern big data analytics infrastructure to scholarly analysis of web archives.

2. Deploy the Archives Unleashed Cloud
   ● Provide a one-stop, web-based portal for scholars to ingest their Archive-It collections and execute a number of analyses with the click of a mouse.

3. Organize Archives Unleashed Datathons
   ● Build a cohesive and sustainable user community by bringing the core project team members together with librarians, archivists, and other interested researchers.
The Archives Unleashed Toolkit is an open-source platform for analyzing web archives, built on Apache Spark.

**Accomplished:**

- **0.80.0 Release**
- Added Python as an additional analytics language
- Developed a modular analysis architecture to align with scholarly workflows
  - Supports both RDD (Resilient Distributed Datasets) and DF (DataFrame) outputs
  - Number of UDF in Scala (RDD & DF) and Python
- Developed a scholarly workflow approach, **FEAV Cycle**
- Created comprehensive **documentation**
The Archives Unleashed Project

*Deliverables - Deploy the Cloud*

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

**Accomplished:**

- Launched in 2018
- Operationalized FEAV Cycle
- Developed GUI frontend; established connection to underlying Toolkit codebase
- Conducted analysis to understand the “Cost of a Warc”
- Analyzed just under a petabyte of data
- Designed and deployed analytics dashboard for monitoring
- Adopted by individuals from 59 unique institutions across 10 countries
The Archives Unleashed Project

**Deliverables - Host Regional 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.

Accomplished:

➔ Hosted **four** datathon events in North America:
  ◆ 2018 Toronto
  ◆ 2018 Vancouver
  ◆ 2019 Washington, DC
  ◆ 2020 New York (Online due to COVID-19)

➔ Collaborated with hosting institutions to provide increased access to web archive datasets

➔ Inspired continued research collaborations

Check out Datathon Projects: [https://archivesunleashed.org/events/](https://archivesunleashed.org/events/)
The Archives Unleashed Project

Additional Outcomes

Notebooks

The project has produced several notebooks, which offer methods for exploring and visualizing web archive derivatives created through by Toolkit. This approach offers interoperability between Archives Unleashed outputs and external tools such as Jupyter Notebooks and Google Colab.

Learning Resources

Resources have been created to support, encourage, empower and instil confidence in scholars approaching new tools and methods to explore web archives.

• 7 Learning guides on how to use and explore derivatives: network, web page text, and domain files.
• 35 Datasets via the Web Archives for Historical Research Group on Zenodo and Dataverse.
The Archives Unleashed Project

Final Thoughts

With the support of our community, the Archives Unleashed Team has been able to achieve our initial goals.

Acknowledgements of Institutional Support

This work is primarily supported by the Andrew W. Mellon Foundation. Other financial and in-kind support comes from 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.