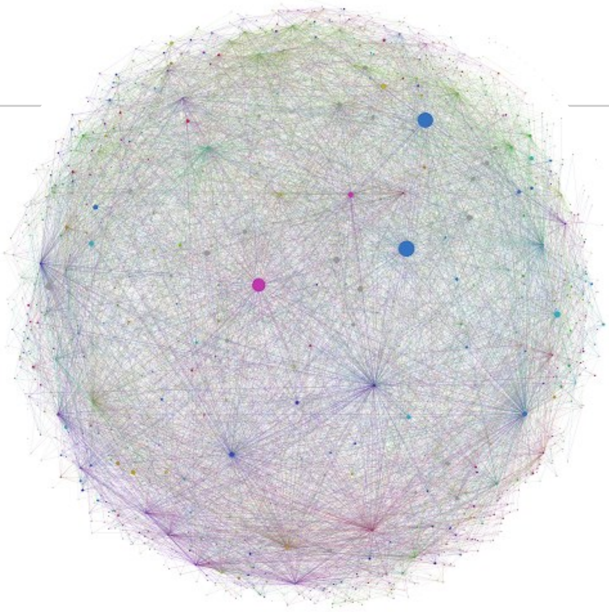
Providing Computational Access to Web Archives: The Archives Unleashed Project

Ian Milligan on behalf of the Archives Unleashed Project

Lowering the Barriers for Computational Access for Digital Archivists



The Web

- → is a reflection of society
- → remains an untapped resource for research[1]
- → has been archived since mid-1990s and led to an abundance of material
- → provides a new context for research data in the form of web archives

Citation:

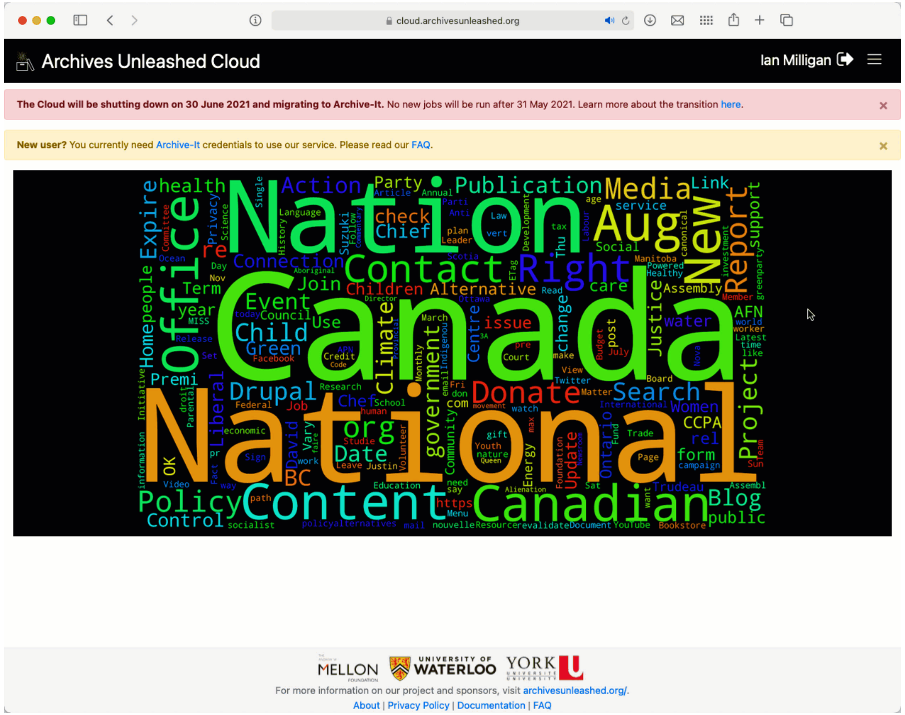
[1] Schroeder, R., & Brügger, N. (2017). Introduction: The web as history. In R. Schroeder & N. Brügger (Eds.), *The Web as History: Using Web Archives to Understand the Past and the Present* (pp. 1–20). UCL Press. https://doi.org/10.2307/j.ctt1mtz55k.6

The Challenge

available analytics tools, community infrastructure, and inaccessible web archival interfaces present **high barriers for conducting research with web archives at scale.**

Archives Unleashed I (2017-2020)

- → Recognizes the critical role of web archives for scholars studying the 1990s onward
- → Developed the Archives Unleashed Cloud
  - ◆ An interface to sync Archive-It collections, analyze them, generate scholarly derivatives, and work with them
- → Standalone system that demonstrated how a web browser interface could power the underlying Apache Spark-based Archives Unleashed Toolkit



The Cloud will be shutting down on 30 June 2021 and migrating to Archive-It. No new jobs will be run after 31 May 2021. Learn more about the transition here. ×

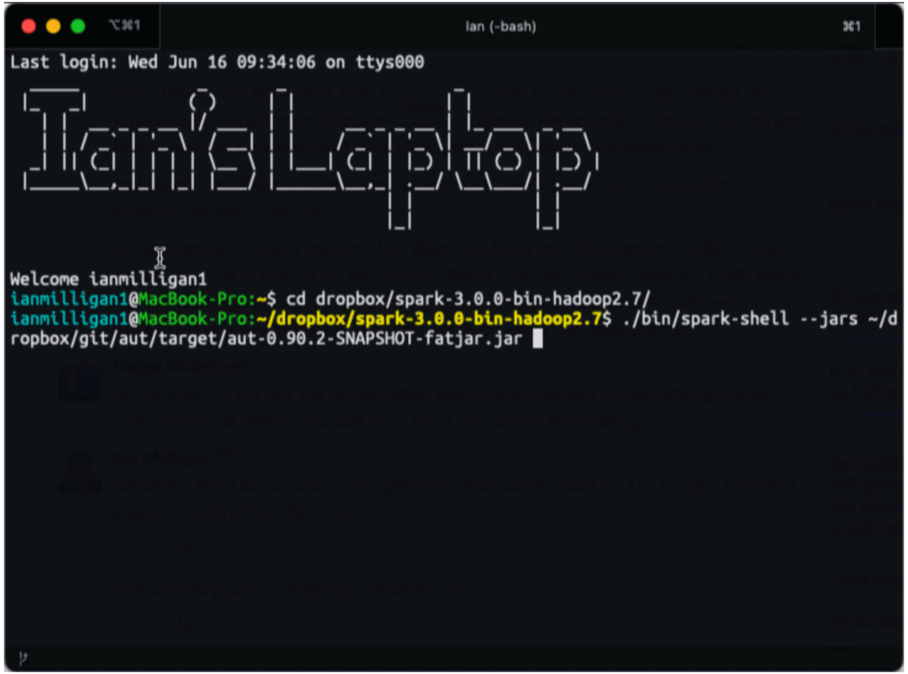
New user? You currently need Archive-It credentials to use our service. Please read our FAQ. ×

For more information on our project and sponsors, visit archivesunleashed.org/.

About | Privacy Policy | Documentation | FAQ

Archives Unleashed I (2017-2020)

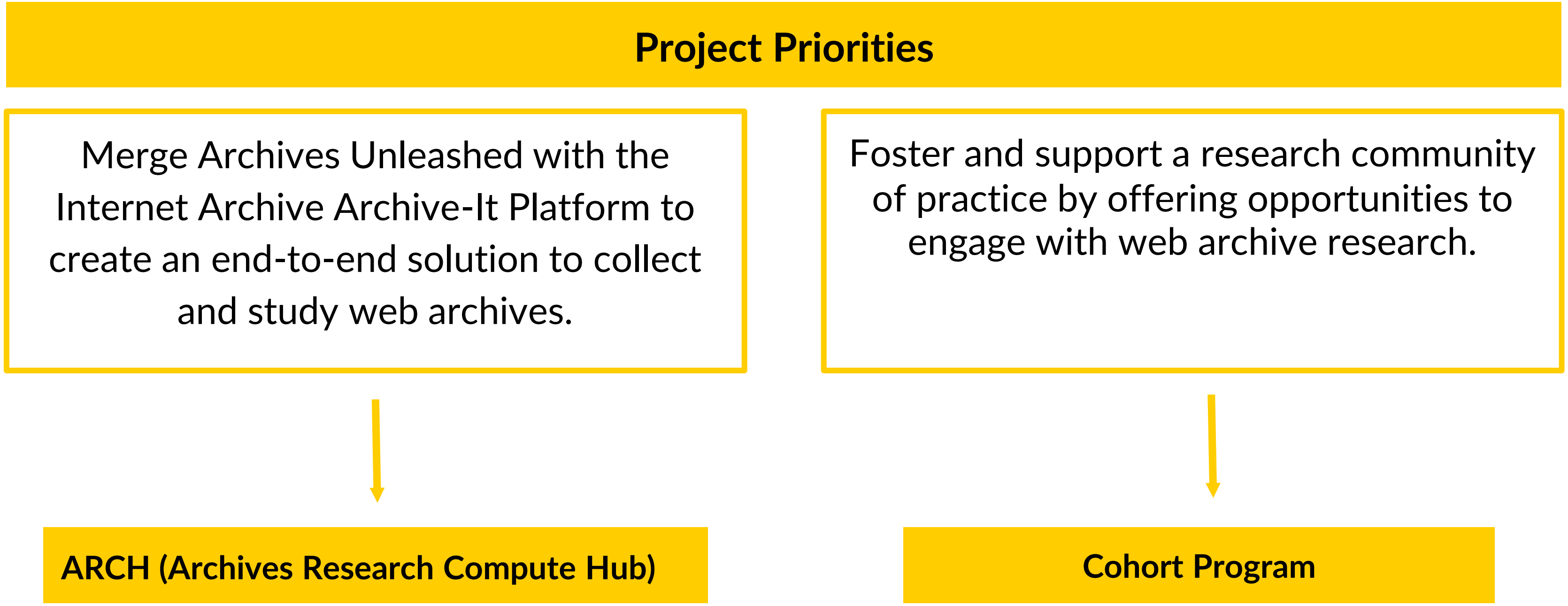
- ➔ Refined the Archives Unleashed Toolkit
- ➔ An Apache Spark library for working with W/ARC files and analyzing them
- ➔ User documentation offers dozens of pre-built scripts to explore web archives and extract information



```
Last login: Wed Jun 16 09:34:06 on ttys000

Welcome ianmilligan1
ianmilligan1@MacBook-Pro:~$ cd dropbox/spark-3.0.0-bin-hadoop2.7/
ianmilligan1@MacBook-Pro:~/dropbox/spark-3.0.0-bin-hadoop2.7$ ./bin/spark-shell --jars ~/d
ropbox/git/aut/target/aut-0.90.2-SNAPSHOT-fatjar.jar
```

Archives Unleashed II (2020-2023)



Project Priorities

Merge Archives Unleashed with the Internet Archive Archive-It Platform to create an end-to-end solution to collect and study web archives.

Foster and support a research community of practice by offering opportunities to engage with web archive research.

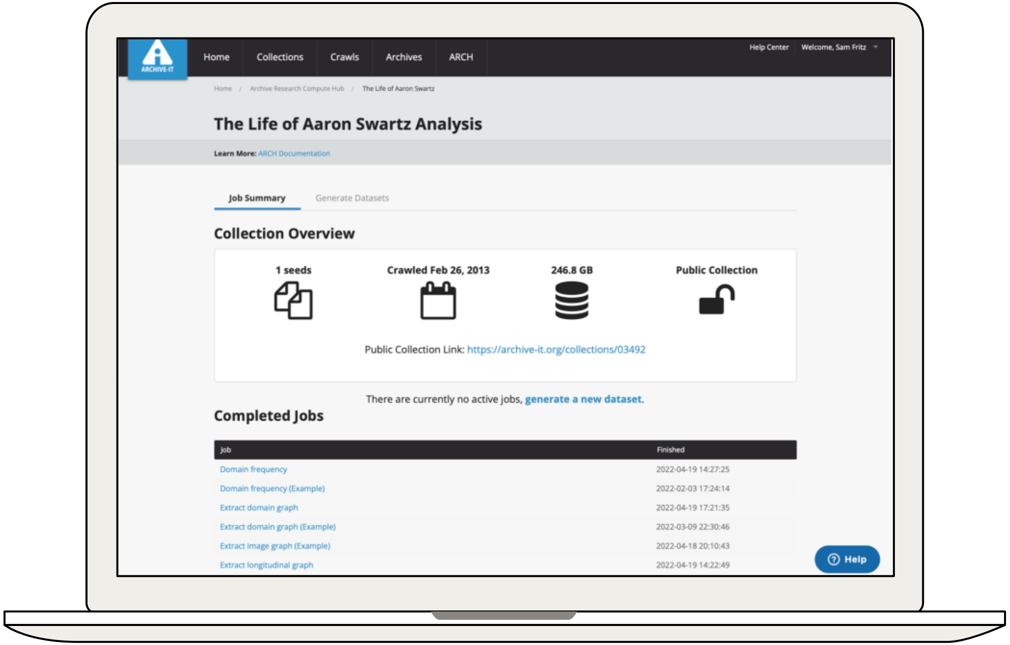
ARCH (Archives Research Compute Hub)

Cohort Program

Introducing the ARCH Platform

- ★ Develop a scalable analysis platform within the Archive- It environment

- ★ ARCH allows users to delve into the rich data within web archival collections for further research



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

There are currently no active jobs, **generate a new dataset**.

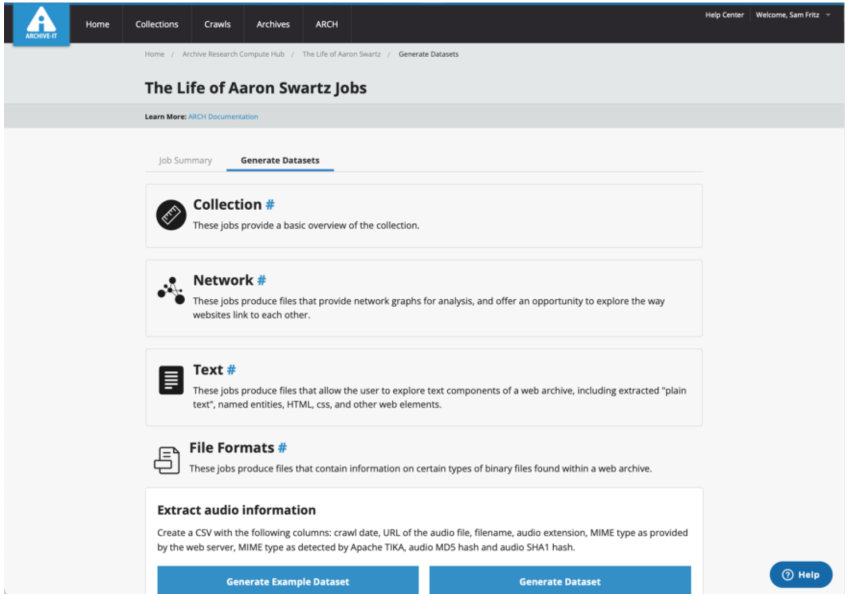
Completed Jobs

Job	Finished
Domain frequency	2022-04-19 14:27:25
Domain frequency (Example)	2022-02-03 17:24:14
Extract domain graph	2022-04-19 17:21:35
Extract domain graph (Example)	2022-03-09 22:30:46
Extract image graph (Example)	2022-04-18 20:10:43
Extract longitudinal graph	2022-04-19 14:22:49

Introducing the ARCH Platform

Features:

- Interactive, familiar environment for current Archive-It subscribers
- Addresses first steps in analysis
- Generate and download over a dozen datasets
- In-browser visualizations and data previews that presents a glimpse into collection content
- Located in the Internet Archive data center, ARCH has quick access to the petabytes of content collected



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.

Text #

These jobs produce files that allow the user to explore text components of a web archive, including extracted "plain text", named entities, HTML, css, and other web elements.

File Formats #

These jobs produce files that contain information on certain types of binary files found within a web archive.

Extract audio information

Create a CSV with the following columns: crawl date, URL of the audio file, filename, audio extension, MIME type as provided by the web server, MIME type as detected by Apache TIKA, audio MD5 hash and audio SHA1 hash.

Switching to Live Demo here

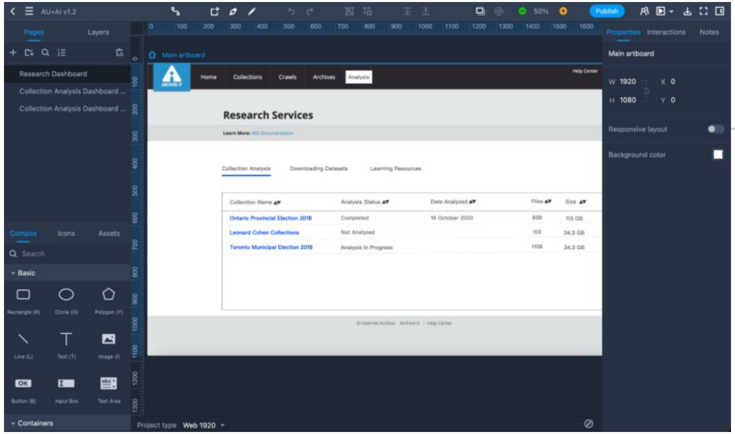
Building for Scalable Analysis

First Steps

Ideation: Identifying existing Archives Unleashed and Archive-It services – overlaps and differences?

Creation: A half-dozen paper drawings to an interactive prototype (using MockPlus) - sketching wireframe

Iteration: Showing teams storyboards, thinking about how to make for an intuitive and friendly workflow



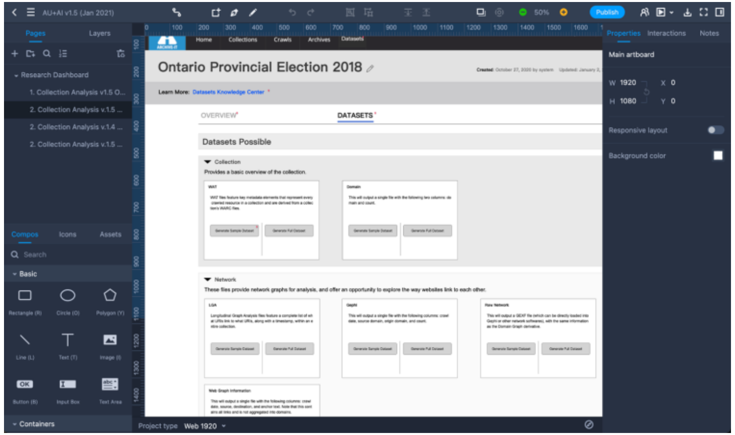
Research Services

Learn More: ARS Documentation

Collection Analysis | Downloading Datasets | Learning Resources

Collection Name	Analysis Status	Date Analyzed	Files	Size
Ontario Provincial Election 2018	Completed	19 October 2020	939	113 GB
Leonard Cohen Collections	Not Analyzed		159	34.3 GB
Toronto Municipal Election 2018	Analysis In Progress		1108	34.3 GB

© Internet Archive Archive-It | Help Center



Ontario Provincial Election 2018

Created: October 27, 2020 by system Updated: January 2,

Learn More: Datasets Knowledge Center

OVERVIEW | DATASETS

Datasets Possible

Collection

Provides a basic overview of the collection.

WAT

WAT files feature key metadata elements that represent every crawled resource in a collection and are derived from a collection's WARC files.

Generate Sample Dataset | Generate Full Dataset

Domain

This will output a single file with the following two columns: domain and count.

Generate Sample Dataset | Generate Full Dataset

Network

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

LGA

Longitudinal Graph Analysis files feature a complete list of what URIs link to what URIs, along with a timestamp, within an entire collection.

Generate Sample Dataset | Generate Full Dataset

Gephi

This will output a single file with the following columns: crawl date, source domain, origin domain, and count.

Generate Sample Dataset | Generate Full Dataset

Raw Network

This will output a GEXF file (which can be directly loaded into Gephi or other network softwares), with the same information as the Domain Graph derivative.

Generate Sample Dataset | Generate Full Dataset

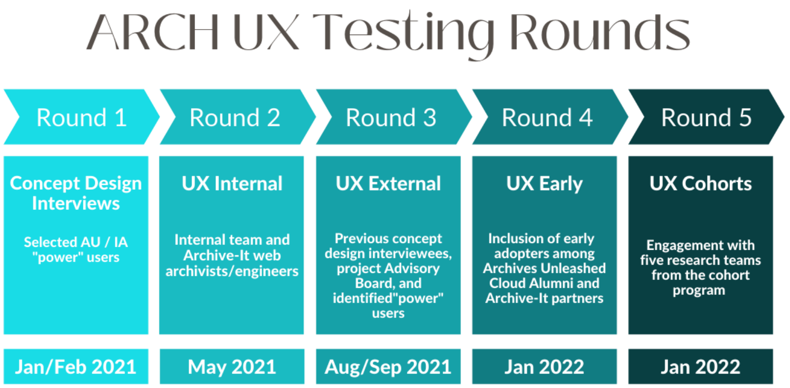
Web Graph Information

This will output a single file with the following columns: crawl date, source, destination, and anchor text. Note that this contains all links and is not aggregated into domains.

User Experience Testing

seeks to understand the impressions, experience, and feelings a user expresses while interacting with a product prototype.

- ➔ Brings the creators and developers into closer alignment with their end-users
- ➔ ARCH UX Goal: understand research behaviours and the user journey while assessing what works well, what challenges arise, and identifying needs that aren't being met
- ➔ Conducted multi-staged user testing process to continually assess user sentiment and the impact with functionality and interface improvements
  - ◆ Concept Design Interviews (2021)
  - ◆ Multiple rounds of UX testing (2021-2022)
  - ◆ Focused interviews with cohort researchers (2022)

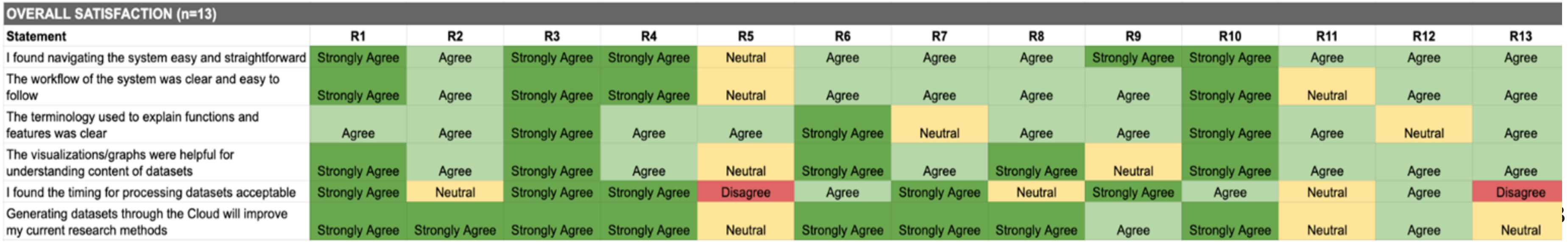


ARCH UX Testing Rounds

Round 1	Round 2	Round 3	Round 4	Round 5
Concept Design Interviews Selected AU / IA "power" users	**UX Internal** Internal team and Archive-It web archivists/engineers	**UX External** Previous concept design interviewees, project Advisory Board, and identified"power" users	**UX Early** Inclusion of early adopters among Archives Unleashed Cloud Alumni and Archive-It partners	**UX Cohorts** Engagement with five research teams from the cohort program
Jan/Feb 2021	May 2021	Aug/Sep 2021	Jan 2022	Jan 2022

User Experience Testing

- ➔ **Spectrum of confidence**: Data scientists confident in their ability to analyze data, collectors less so.
- ➔ **Better integration** into analysis pipelines (i.e. command line download)
- ➔ **Enhanced Analysis** to include additional datasets that respond to research needs
- ➔ **Increased alignment** with accessibility standards
- ➔ **Improved workflow and navigation** by providing way-finding support and prompts
- ➔ **Clearer language** for our educational users and additional documentation and learning resources



OVERALL SATISFACTION (n=13)

Statement	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13
I found navigating the system easy and straightforward	Strongly Agree	Agree	Strongly Agree	Strongly Agree	Neutral	Agree	Agree	Agree	Strongly Agree	Strongly Agree	Agree	Agree	Agree
The workflow of the system was clear and easy to follow	Strongly Agree	Agree	Strongly Agree	Strongly Agree	Neutral	Agree	Agree	Agree	Agree	Strongly Agree	Neutral	Agree	Agree
The terminology used to explain functions and features was clear	Agree	Agree	Strongly Agree	Agree	Agree	Strongly Agree	Neutral	Agree	Agree	Strongly Agree	Agree	Neutral	Agree
The visualizations/graphs were helpful for understanding content of datasets	Strongly Agree	Agree	Strongly Agree	Agree	Neutral	Strongly Agree	Agree	Strongly Agree	Neutral	Strongly Agree	Agree	Agree	Agree
I found the timing for processing datasets acceptable	Strongly Agree	Neutral	Strongly Agree	Strongly Agree	Disagree	Agree	Strongly Agree	Neutral	Strongly Agree	Agree	Neutral	Agree	Disagree
Generating datasets through the Cloud will improve my current research methods	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree	Neutral	Strongly Agree	Strongly Agree	Strongly Agree	Agree	Strongly Agree	Neutral	Agree	Neutral

Connection and Integration

- Front-end development
- Connecting back-end process
- Integration of product into the production environment
- Technical Choices
  - Scalatra
  - Apache Spark
  - HDFS
  - Sparkling
  - Archives Unleashed Toolkit

Continual Improvement

- ARCH in beta, with pilot users from the Cohort Program
- Monitoring
- Final year developments
  - Pre-Filtering (user defined queries)
  - Thoughtful access and use for non AI subscribers

Lessons Learned

Reflecting on Lessons Learned

Lesson 1

If you build it, they won’t come. You need to actively work to create an environments where users feel comfortable.

Lesson 2

Work to meet your users. This doesn’t necessarily mean that you will make all of them happy, but it does mean you need to listen and be responsive through UX testing and outreach.

Lesson 3

Be ready for the unexpected! If there’s something that is 1 in a 1,000,000, you’ll run into it dozens of times in your WARCs. So be ready for error handling and continual improvement.

Acknowledgements of Institutional Support

In partnership with the Internet Archive's Archive-It, this work is primarily supported by the Andrew W. Mellon Foundation. Other financial and in-kind support has come from the Social Sciences and Humanities Research Council, Compute Canada, York University Libraries, Start Smart Labs, and the Faculty of Arts and David R. Cheriton School of Computer Science at the University of Waterloo.

THE ANDREW W. MELLON FOUNDATION

ARCHIVE-IT

Ontario
MINISTRY OF RESEARCH AND INNOVATION

UNIVERSITY OF WATERLOO
FACULTY OF ARTS

YORK
UNIVERSITÉ
UNIVERSITY

compute canada

START SMART LABS

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Thanks!

Any *questions* ?

Connect with out project team:

- @UnleashArchives
- archivesunleashed@gmail.com
- https://archivesunleashed.org