Providing Computational Access to Web Archives: The Archives Unleashed Project

Ian Milligan on behalf of the Archives Unleashed Project
The Web

→ is a reflection of society
→ remains an untapped resource for research¹
→ has been archived since mid-1990s and led to an abundance of material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:
available analytics tools, community infrastructure, and inaccessible web archival interfaces present **high barriers for** conducting research with web archives at scale.
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
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
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
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
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
User Experience Testing

→ 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)
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

<table>
<thead>
<tr>
<th>OVERALL SATISFACTION (n=13)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement</td>
</tr>
<tr>
<td>I found navigating the system easy and straightforward</td>
</tr>
<tr>
<td>The workflow of the system was clear and easy to follow</td>
</tr>
<tr>
<td>The terminology used to explain functions and features was clear</td>
</tr>
<tr>
<td>The visualizations/graphs were helpful for understanding content of datasets</td>
</tr>
<tr>
<td>I found the timing for processing datasets acceptable</td>
</tr>
<tr>
<td>Generating datasets through the Cloud will improve my current research methods</td>
</tr>
</tbody>
</table>
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 environment 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.
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.
Thanks!

Any questions?

Connect with our project team:
- @UnleashArchives
- archivesunleashed@gmail.com
- https://archivesunleashed.org