

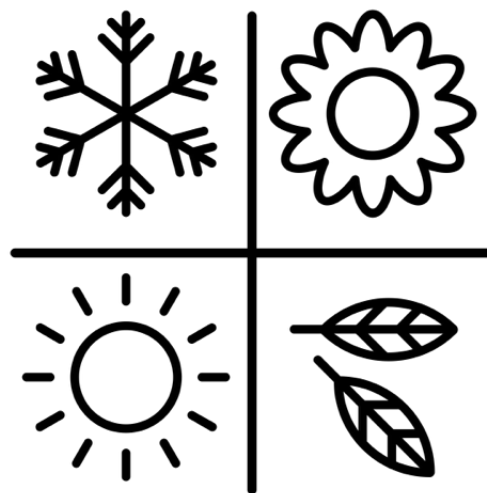


## Fall 2022 Newsletter

---

Welcome back to the Archives Unleashed newsletter! You may have noticed we took a short break, but we've been hard at work as this is the final year of our current funded project!

We are so excited to share project updates and resources from the community, so without further ado, here's a round-up of Archives Unleashed Project 2022 activities.

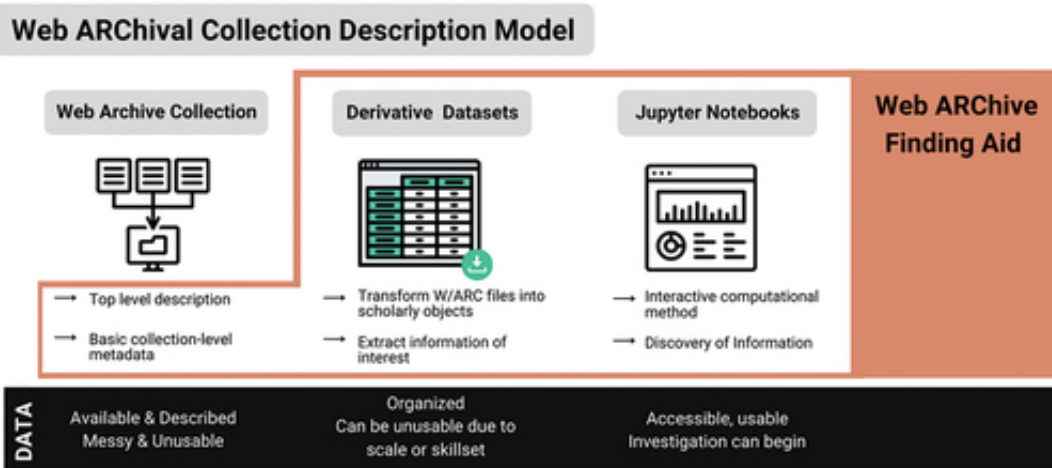


---

## Just In!

---

**[Creating order from the mess: web archive derivative datasets and notebooks](#)**



A recent article by Nick Ruest, Sam Fritz, and Ian Milligan is [#OpenAccess in Archives and Records](#).

For a quarter-century, memory institutions have been preserving web-based content. These web archives have been collected and stored in ARC and WARC (W/ARC) file formats and will form a basis for contemporary histories. Yet, these formats present significant challenges to researchers who wish to access and use web archival data. This is primarily due to the nature of collecting, storing, and providing access to these multifaceted digital objects. In other words, web archives are messy. Applying traditional archival methods of description to digital-born collections is complicated due to issues of provenance, original order, and scale. However, we believe that archival description offers a practical starting point for thinking about access. This paper argues a robust finding aid must extend beyond basic collection-level description to allow for more meaningful interactions with web archives. As such, we propose a reimagining of a traditional finding-aid model into a three-level mode of description to include computational methods, the generation of derivative datasets, and interactive code-rich notebooks. These three factors combine to ultimately contribute to the expanded access and use of web archives.

---

# What's been happening?

## ARCH Developments

- As we led another round of UX testing at the beginning of the year, **several UI improvements** were implemented based on user feedback to improve the user journey through ARCH, clarify terminology and processes, and streamline user documentation.
- The project hit a major milestone this summer with the [adoption of Sparkling](#), a data processing library for Apache Spark and has drastically affected the speed of data processing and the efficiency of loading, parsing, and storing W/ARC.
- With special thanks to Nick Ruest, we have just implemented a **new feature** in ARCH - the inclusion of **Google Colab Notebooks**, which provides inspiration for the initial exploration and analysis of ARCH derivatives!
- As our team continues to work on ARCH developments into the new year, users will soon see additional features, including the implementation of a [last-modified date](#) (available through the AU Toolkit) and user-defined filtering for collections.

## AU Toolkit

- With the **adoption of Sparkling** (developed by Helge Holzmänn) and parity with ARCH, the Archives Unleashed Toolkit hit a significant project milestone with a **1.2.0 release** in November!
- As part of its 1.0.0 release, the Toolkit fully supports DataFrames and RDD, Python and Scala implementations, and allows users to create derivatives via the spark-submit. In addition, [user documentation](#) has been updated.
- Most recently, Nick Ruest has released a new feature to help researchers analyze the [last-modified date](#) in a web archive collection rather than relying solely on the crawl date!

## Cohort Program

- In June, our team hosted cohort teams at the Internet Archives' new (and beautiful!) Canadian Headquarters at the [Permanent Library Limited](#) in Vancouver, BC.
- Members from the 2022 and 2023 cohorts groups gathered to share and collaborate on web archives research projects.
- This event allowed inaugural teams to share their research highlights and experiences while welcoming and providing encouragement to our second round of groups.
- Over the past months, we have onboarded teams to ARCH, connected them with institutional curators, and provided research consultation as they dig into analysis with various web archive collections. We're seeing a lot of interest in network analysis and topic modelling as the first steps to exploring data.

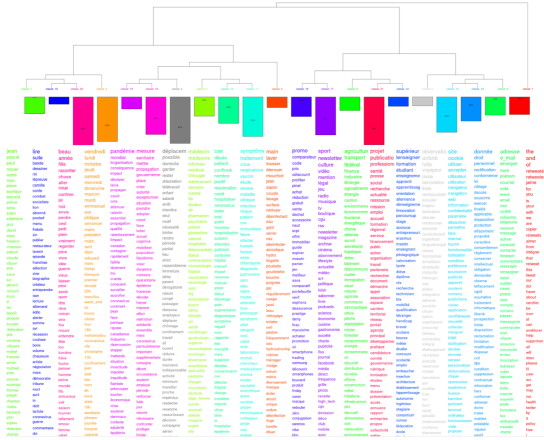
## Community Engagement

- While the pandemic has limited our opportunities to meet with friends and colleagues in person, we certainly haven't slowed down connecting with our favourite communities!
  - Team members virtually attended and presented at **JCDL**, **WADL**, and **IIPC's WAC**.
  - Based on our work with members of the cohort program, we also were invited to speak at the **Internet Archive's Library as Laboratory series** and the Fall **Archive-It Partner Meeting** to discuss the applications of web archives research.
  - Project members Ian Milligan, Nick Ruest, and Jefferson Bailey are heading to **CNI's Fall 2022 Membership Meeting** to talk about the collaborative efforts of Archives Unleashed and Archive-It in Supporting Computational Research on Large Digital Collections.
-

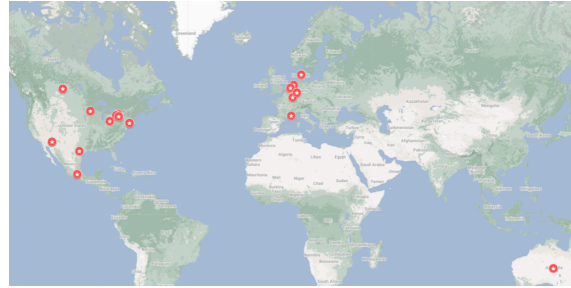
# AU Cohort Program

---

## Research Applications with Web Archives: Collaboration Among Archives Unleashed Cohorts



## Web Archives Research: Return of the Cohorts



---

## Reading + Resources



---

Don't forget to check out our [Archives Unleashed Medium](#) site for all of our greatest blogging hits!

### **NEW** [AUT and Last Date Modified](#)

Dates can be difficult to understand when trying to research web archives. While the crawl date is always there, the crawl date doesn't tell you when the content was actually created. Yet many researchers are interested

```
import io.archivesunleashed._

val data = "/sample-data/geocities/GEOCITIES-20091027143300-00114-1a400112.us"
RecordLoader.loadArchives(data, sc)
  .all()
  .select($"crawl_date", $"last_modified_date", $"mime_type_web_server")
  .show(20, false)

+-----+-----+-----+
|crawl_date|last_modified_date|mime_type_web_server|
+-----+-----+-----+
|20091027143300|200909232323454|text/html|
|20091027143259|20090913163029|image/jpeg|
|20091027143300|20020211154553|image/jpeg|
|20091027143259|19980919164703|image/jpeg|
|20091027143259|20080125150303|text/html|
|20091027143300|20010921224658|image/gif|
|20091027143258|20081009015203|image/jpeg|
|20091027143300|20080416145103|image/jpeg|
|20091027143300|20090223022835|text/html|
|20091027143300|20030928090558|image/jpeg|
|20091027143300|20091027143300|text/html|
|20091027143300|20021203212451|text/html|
|20091027143300|20040530033010|image/bmp|
|20091027143300|20090223022352|text/html|
|20091027143300|20010608202736|text/html|
+-----+-----+-----+

only showing top 20 rows
```

**The Archives Unleashed Toolkit**

The Archives Unleashed Toolkit is an open-source platform for analyzing web archives using Apache Spark, and makes use of Sparkling for parsing WARC records. The toolkit provides powerful tools for analytics and data processing. It is part of the Archives Unleashed Project.

To learn more about the Toolkit and how to use, please see our [comprehensive documentation](#).

If you would like a more in-depth look at the project, please check out the following two articles:

- Nick Ruest, Jimmy Lin, Ian Milligan, and Samantha Fitz. *The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives*. Proceedings of the 2020 IEEE/ACM Joint Conference on Digital Libraries (JCDL 2020), Wuhan, China.
- Jimmy Lin, Ian Milligan, Jeremy Wiebe, and Alice Zhou. *Warehouse: Scalable Analytics Infrastructure for Exploring Web Archives*. ACM Journal on Computing and Cultural Heritage, 10(4), Article 22, 2017.

Archives Unleashed Toolkit [README](#)

in tracing changes to content over time. Our Co-PI Nick Ruest has just implemented a new feature in the Archives Unleashed Toolkit that will allow researchers to analyze the last-modified date rather than rely on the crawl date.

## NEW [Archives Unleashed Toolkit 1.0.0: A Sparkling New Way to Explore Web Archives](#)

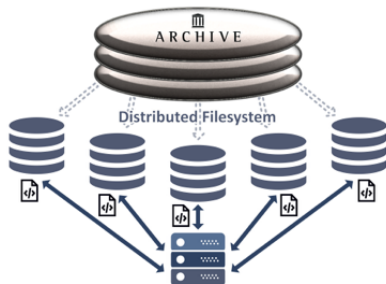
Our team is proud to present the [1.1.0 release](#) of the Archives Unleashed Toolkit! This release is the result of 5 years of development (and a lot of hard work) and is a significant milestone for the Archives

Subscribe

Past Issues

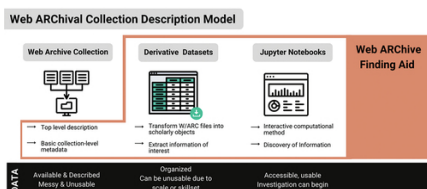
Translate ▼

the 1.1.0, including the adoption of Sparkling!

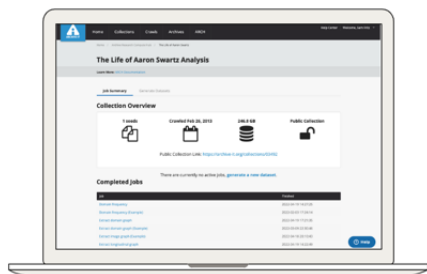


## ARTICLE [ABCDEF - The 6 key features behind scalable, multi-tenant web archive processing with ARCH: Archive, Big Data, Concurrent, Distributed, Efficient, Flexible](#)

- [Watch the JCDL 2022 Minute Madness](#)
- [Read the PrePrint](#)
- [Read the Article via IEEE](#)



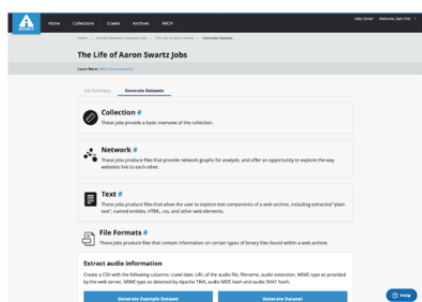
## ARTICLE [Creating order from the mess: web archive derivative datasets and notebooks](#)



## CONFERENCE [#iipcWAC22](#): Building a Computational Research Platform for [#WebArchives](#)


Members of the project team participated in a panel presentation and highlight the process underpinning the ARCH interface: from its inception and its construction to its use by a sponsored researcher.

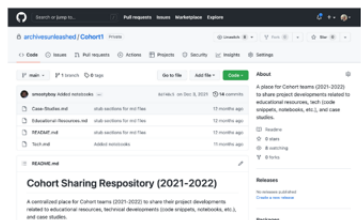
Watch the [conference session](#), including pre-recorded presentations from Jefferson Bailey, Ian Milligan, Nick Ruest, and Valérie Schafer, alongside the Q&A period.



## CONFERENCE Demo at WADL

Our project developers, Nick Ruest and Helge Holzmann, provided a live demo of the ARCH interface at WADL (Web Archiving and Digital Libraries)

 Check out Kritika Garg's [Web Archiving and Digital Libraries \(WADL\) Workshop 2022 Trip Report](#) report for all the highlights!



## TALK Archives Unleashed Cohort Program: Opportunities to Access, Explore, and Engage with Web Archives

During the 2022 Fall Archive-It Partner Meeting, Project Manager Sam Fritz provided an overview of ARCH and the [Archives Unleashed Cohort program](#) with a focus on the research applications of web archives. Cal Murgu, a 2021-2022 Cohort participant discussed his team's experience of using ARCH to [assess municipal responses to the](#)



---

# From the Community

---

We love sharing tools, articles, and resources from a variety of communities that speak touch on web archiving, digital preservation, and computational analysis (and beyond).

- [McMaster's Sherman Centre for Digital Scholarship](#) - regularly hosts virtual and in-person workshops that cover a wide range of data analysis and visualization topics. Many of their [workshops and events](#) are open to the community!
- [WARCnet Papers and Special Reports](#) - check out the most recent papers, presentations, and reports related to activities of the WARCnet network.
- In working with our cohort groups, we've found (and love!) Melanie Walsh's [Introduction to Cultural Analytics & Python](#) online textbook is helpful. It's freely accessible and provides foundational knowledge for those interested in using Python to explore data in a variety of ways.
- [Building an open-source glossary for digital preservation](#): Initiated by Andy Jackson and discussions from 2022 World Digital Preservation to create a community-owned glossary.
- [Glam Workbench](#) by Tim Sherrat: a collection of tools, tutorials, examples, and hacks to help you work with data from the GLAM sector.
- **Archiving Your Twitter:** Across many disciplines and fields, we've seen a growing concern for the persistence of Twitter as a



platform. We've noted a few tools and methods shared by our web archiving friends and colleagues to help preserve your Twitter account content.

- Michelle Weigle (ODU) - [Preserving links from your Twitter Archive](#) (and yes, we did save this [thread](#) through the Wayback Machine)
- Internet Archive - [How to Archive Your Tweets with the Wayback Machine](#)



[DocNow](#) has published a collectively curated catalogue of Twitter datasets that cover a variety of historically significant topics.

Recent datasets released include:

- [Teamsters and Teamsters Locals](#) (Vakil Smallen, Daniel Kerchner)
- [The Social Archive of the Polish Web](#) (Marcin Wilkowski)
- [#retweetthe8th: 2018 Referendum to repeal the 8th Amendment of the Constitution of Ireland](#) (Emmet Ó Briain, Jennifer Foster)



## Archives Unleashed + Collaborators

If you're more interested in [web archive-focused datasets](#), Archives Unleashed has collaborated with several academic libraries, collection curators, and Archive-It to offer collection derivatives. Over two dozen datasets are available with accompanying citation information.

<https://archivesunleashed.org/publications/#datasets>

Additional datasets, for instance, Twitter datasets can also be accessed directly through our Web Archives for Historical Research Group on [Zenodo](#) and [Dataverse](#).

We've also collaborated with folks at Archive-It to extend opportunities to explore #webarchives with a series of pre-processed datasets. All you have to do is grab a dataset and dive in!

- Geocities, web archive derivatives of the GeoCities collection from the Internet Archive (v3) are provided and were created using the Archives Unleashed Toolkit (1.2.0) <https://archive.org/details/geocities-webarchive-collection-derivatives>
- [Friendster](#) (2003–2015), the datasets provided are generated from the larger Friendster web archive collection in the Internet Archive. This provides a great opportunity for researchers to start exploring Friendster data without being overwhelmed by the full 10TB collection. <https://archive.org/details/friendsterdatasets>
- Early Web Language Datasets (1996–1999), various datasets created for research, scholarly, or general use, were generated from the "early web" era (1996-1999) of the Internet Archive's global web archive collection. <https://archive.org/details/earlywebdatasets>

---

## 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](#)

Submit to our [datathon](#)

Share our news with colleagues and friends

Twitter



GitHub

Website

Email

---

The Archives Unleashed Project, aim 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

Copyright © 2018 The Archives Unleashed Project. All Rights Reserved.

To change how you receive these, you can update your preferences or unsubscribe from this list.

Twitter



GitHub

Website

Email

---

The Archives Unleashed Project, aim 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

Copyright © 2018 The Archives Unleashed Project. All Rights Reserved.

To change how you receive these, you can update your preferences or unsubscribe from this list.

---

This email was sent to <<Email Address>>

[why did I get this?](#) [unsubscribe from this list](#) [update subscription preferences](#)

Archives Unleashed · 200 University Ave W · Waterloo, On N2L 3G1 · Canada

