Archives Unleashed

Spring 2020 Newsletter

We know it's been a challenging start to the year, but we hope everyone is staying healthy, staying safe, and – if possible – staying home!

Introducing

AUT 0.60.0 Release

The latest release of the Archives Unleashed Toolkit is now available!

The 0.60.0 also marks a major milestone as the migration from RDD to DataFrames is complete. Check out our documentation!

It also means that the team is closer to Archives Unleashed Toolkit 1.0,

Spark-submit

Part of the AUT 0.60.0 release provides a new feature to users, the ability to easily use the Toolkit with spark-submit!

With this new functionality, users can deploy a variety of extraction jobs (domain frequency, domain graph, plain text, etc.) and spark-submit offers the ability to run

so keep posted for announcements.

Check out the 0.60.0 release notes for full details on new features and implementations.

multiple jobs in parallel.

We've provided a documentation page with some recipes to help get you started.

Datasets

Over the past three years, our team has been working with several institutions and collaborators to provide scholarly access to derivative data of over **30 web archive collections.**

These **freely available** datasets were created using the Archives Unleashed Cloud and new functionality in the Archives Unleashed Toolkit, in collaboration with the individuals who curate the web archive collections, and the institutions which support web archiving activity.

A running list of datasets is available through our publications page.

You are also welcome to search through the Zenodo and Dataverse **Web Archives for Historical Research Group** communities to access derivative files and citable DOI.

We'd like to acknowledge the institutions who have provided an opportunity for scholars to explore their web archive collections by allowing us to generate derivatives:

- Columbia University Libraries
- Ivy Plus Libraries Confederation
- Bibliothèque et Archives nationales du Québec
- Institutions from the Web Archives for Longitudinal Knowledge (WALK) project include:
    - University of Toronto
    - University of Alberta
    - University of Victoria
    - University of Winnipeg
    - Dalhousie University
    - Simon Fraser University

What's been happening? 

The Short List

- **Archives Unleashed Toolkit**
  - Published two releases: 0.50.0 and the current 0.60.0
  - Completed the migration from Resilient Distributed Dataset (RDD) to DataFrames (a **major milestone**); and
  - Implemented a new feature: the spark-submit app;
- **Archives Unleashed Cloud**
  - We've currently processed **889TB**!
  - The focus has been on sketching out methods of exploring new derivatives and effectively engaging derivatives and collections with external platforms to increase access and analysis methods.
- **Archives Unleashed Notebooks**
  - Set up a series Notebooks (for the NYC-gone-virtual Datathon) which provides example approaches to working with web archive derivatives in Parquet format.
  - Connected and launched Archives Unleashed Notebooks to the Google Colab environment!
- Participated in a number of conferences and scholarly meetings.
- Hosted our final datathon online with support from our colleagues at Columbia University Libraries and the Ivy Plus Libraries Confederation.

Featured Articles 

Archives Unleashed Project: 2019 Progress Report

In 2019 the Archives Unleashed team made significant progress. Checkout the developments in our annual progress report.

(Read More...)

Cloud-hosted web archive data: The winding path to web archive collections as data

Co-PI Nick Ruest provides lessons, insights & reflections on lowering access barriers to web archives.

(Read More...)

So You Want to Move Your Event Online?

In light of COVID-19, our team transitioned our final datathon to an online model. If you're part of a group looking at potential solutions, check out the insights and lessons our team learned.

(Read More...)

Using the Archives Unleashed Toolkit at the Munich DigitiZation Center

Katharina Schmid shares insights into how the Munich DigitiZation Center (MDZ) at the Bavarian State Library analyzed a 2.5TB web archive collection (European Parliamentary elections) using the Archives Unleashed Toolkit!

(Read More...)

Check It Out

In North American, many of us are heading into the seventh week of physical-distancing and a shift to working from home (for those that can). Amid the global pandemic, there have been several projects that have been initiated by a need to examine the uncharted territory we've all faced.

We just wanted to take a moment to highlight some of the inspiring projects being within the web archiving field:

Novel Coronavirus (COVID-19) Collection

The Content Development Group of the IIPC and Archive-It continue to collaborate on a web archive collection that aims to preserve web content related to the Novel Coronavirus (COVID-19) outbreak. Currently, it holds 12,574,647 documents in 39 languages.

Help them collect websites by visiting: http://netpreserve.org

Neural Covidex

This system applies network modeling and AI techniques to answer techniques using the COVID-19 Open Research Dataset (CORD-19) - which contains over 47,000 scholarly articles about coronavirus-related research. This project is led by one of our Co-PIs Jimmy Lin (University of Waterloo) and Kyunghyun Cho (NYU).

Documenting the Now

This project develops tools and platforms to respond to the civic and scholarly collection, preservation, and use of social media content to chronicle historically significant events.

The recently redesigned DocNow Catalog provides a friendly interface to community contributed Twitter datasets.

The project welcomes new collections and is currently seeking to add contributions related to COVID-19 twitter datasets.

Library of Congress

This year marks a special occasion - the 220th anniversary of the Library of Congress, and the 20th anniversary of web archiving at LoC. You can check out Abbie Grotke's post on The Signal, which reflects on the early years of the Web Preservation Project Pilot.

You can also access a recent article, "Meet Your Meme Lords," by the New York Times, which features the dedicated web archiving team at the Library of Congress (and a guest appearance by one of our principal investigators too!).

Awesome Web Archiving

National Emergency Library

Have you ever wished for a list of web archiving resources when it comes to training, tools/software, and community resources? Look no further, because the IIPC has a great list for getting started with web archiving!

The Internet Archive launched a temporary collection to help support emergency remote teaching, research activities, and independent scholarship, amid closures of schools, universities, and libraries.

Recent Workshops and Presentations

While COVID-19 put a kibosh on travel for conferences and workshops, we wanted to take the opportunity to share resources created for our scholarly activities over the past few weeks.

CHIIR

Our three principal investigators collaborated with some other colleagues on a recent article for the ACM CHIIR 2020 (Conference on Human Information Interaction and Retrieval) proceedings: "**We Could, but Should We?: Ethical Considerations for Providing Access to GeoCities and Other Historical Digital Collections.**"

CHIIR '20: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval March 2020 Pages 135–144 https://doi.org/10.1145/3343413.3377980

The preprint is also available: http://hdl.handle.net/10315/36947

[Virtual] New York Datathon

We transitioned our final datathon to an online environment amid a global pandemic. Special thanks to our co-organizers Pamela Graham, Alex Thurman (Columbia University), and Samantha Abrams (Ivy Plus Libraries Confederation). These organizations also provided access to their web archive collections for participants to explore.

Resources Available:

- **Opening Presentations**: Google Slides + Video
- **Web Archive Collection Derivatives**: Want to explore the data yourself? You're in luck! Nick Ruest worked with folks from Columbia University Libraries and Ivy Plus to generate derivatives and create DOIs for access, use, and citation. This means you are free to access and use the derivatives in your own research! Check out the Zenodo and Dataverse Web Archives for Historical Research Group.
- **Quick Guide: Setting up Colab Notebooks** Video demonstrates how datathon teams set up sample Notebooks through via Google Colab.
- Check out final projects from our datathon teams!

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 I Waterloo, ON N2L 3G1