

Winter 2020 Newsletter



Happy New Year!

Welcome back - let's catch up and talk #WebArchiving.

Introducing —

Tweet Archives Unleashed Toolkit (TWUT)

One of the newest features of Archives Unleashed, analyzing lineoriented JSON Twitter archives with Apache Spark.

Try it out:

https://github.com/archivesunleashed/twut

Archives Unleashed Notebooks + Google Colab

You can now use Archives
Unleashed Toolkit derivatives (in
Apache Parquet format) in a Google
Colab environment! We have a
variety of example notebooks.

Try it out:

https://github.com/archivesunleashed/notebooks

The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives

The Archives Unleashed project aims to improve scholarly access to web archives through a multi-pronged strategy involving tool creation, process modeling, and community building - all proceeding concurrently in mutually-reinforcing efforts. As we near the end of our initially-conceived three-year project, we report on our progress and share lessons learned along the way.

Preprint Available: https://arxiv.org/abs/2001.05399

Celebrating with our Collaborators

The Archives Unleashed Team would like to extend special thanks to two collaborators we've had an opportunity to work with over the past few months: Jeremy Wiebe and Gursimran Singh. Their contributions have helped to move the development and implementation of DataFrames support forward in strides.

What's been happening?



The Short List

Archives Unleashed Toolkit

Significant progress in the move from **Resilient Distributed Dataset** (**RDD**) to **DataFrames**;

Restructuring of documentation organization: users are now able to access a cookbook-esque type format, which has made information more user-friendly; and

Code refactoring: keeping long-term sustainability in mind, we've

focused on refactoring our code base to simplify it and reduce dependencies.

Archives Unleashed Cloud

Processed our largest collection yet! A 17.6 TB collection (<u>2012 Summer Olympics Collection</u>) from the <u>International Internet</u> Preservation Consortium.

The Cloud officially celebrates two years in production!

Archives Unleashed Notebooks

Launched the Archives Unleashed Notebooks - a series of Jupyter Notebooks that provide example approaches to working with new web archive derivatives.

Connected Archives Unleashed Notebooks to the Google Colab environment!

Keep and eye out for future connections from our notebooks to the Archives Unleashed Cloud.

Participated in a number of conferences and scholarly meetings.

Currently planning our final datathon co-hosted with Columbia University

(Butler Library) and Ivy Plus Libraries Confederation; as well as an additional datathon after IIPC's Web Archiving Conference.

For a comprehensive list of project developments see our recently published post:

Archives Unleashed Project: 2019 Progress Report

Featured Articles



Archives Unleashed: A Year in Review (2019)

It's hard to believe that we are already two and a half years into the <u>Archives Unleashed Project!</u> So as we begin this new year, a new decade, and head into the final six months of this project, let's reflect on the work and milestones the Archives Unleashed

The Archives Unleashed Toolkit as a Finding Aid Utility

So, a finding aid. How could we create a finding aid for a web archive collection with relatively minimal labour? With the Archives Unleashed Toolkit, we can create a more robust finding aid. But, what should it be?

Team and community has reached.

(Read More...)

(Read More...)

GeoCities and the spacer.gif

Reflecting on a recent article by <u>Trevor</u>

Owens and <u>Grace Thomas</u>, Archives

Unleashed developer Nick Ruest
explores a large GeoCities dataset!

The catch, he uses the <u>Archives</u>

<u>Unleashed Toolkit</u> to
identify 121,371,844 images for
analysis.

(Read More...)

twut. Wait, wut? Twut?

Early-on, and up until earlier this year, the Archives Unleashed Toolkit had functionality built in for reading Twitter data into RDDs, and processing them with a number of small methods that overlapped a bit with the utilities that twarc ships with. In an effort to simplify the Toolkit's codebase and dependencies, we removed all of that functionality.

(Read More...)

Check It Out 60

The Archives Unleashed Toolkit: Latest Documentation

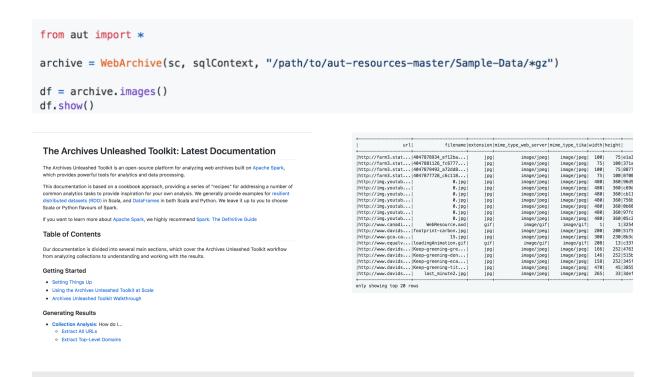
Significant development updates on the Archives Unleashed Toolkit means we've revamped our documentation to help inspire exploration of your web archive collections.

Our documentation is divided into several main sections, which cover the Archives Unleashed Toolkit workflow from analyzing collections to understanding and working with the results.

We've taken a cookbook approach, which provide a series of "recipes" for addressing a number of common analytics tasks. You'll find examples for Resilient Distributed Datasets

(RDD) in Scala, and DataFrames in both Scala and Python.

Take the Archives Unleashed Toolkit for a spin: aut-docs/current



Recent Workshops and Presentations

Feel free to check out presentation slides from some of the most recent presentations/workshops given by Archives Unleashed team members.



September 2019

Project Manager, Samantha Fritz presented at ACCESS on identifying and applying theme of access and usability within open source projects, using the Archives Unleashed Project



October 2019

Samantha Fritz and Sarah McTavish presented a mini workshop at <u>CEDWARC</u>, to engage with library and archive professionals in exploring web archiving analysis. Presentations walk

View Slides

View Overview Slides

View Workshop Slides



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

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.