The Archives Unleashed Notebook:
Madlibs for Jumpstarting Scholarly Explorations of Web Archives

Ryan Deschamps[1], Nick Ruest[2], Jimmy Lin[1], Samantha Fritz[1], and Ian Milligan[1]

[1] University of Waterloo, [2] York University

Problem Statement

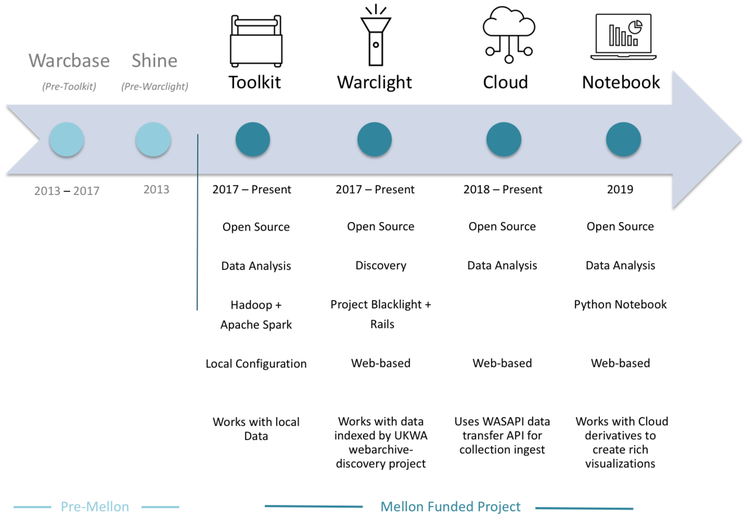
How can the Archives Unleashed Project help support scholars – particularly humanists and social scientists – cope with challenges of analyzing web archives at scale?

Motivations

- Web archives are a critical resource for those exploring the recent past.
- Humanists and social scientists face several challenges with web archives: scholarly access and scale.
- We engaged computer scientists and historians to co-design an analytics framework that would be usable by humanities scholars and social scientists with no formal computer science training. This is the FAAV cycle [1].
    - Filtering (i.e. by date, domain, subject)
    - Analyzing (finding pages, images, documents, etc. that fit certain criteria)
    - Aggregating (counting or performing statistical operations), and
    - Visualizing (from tables to graphs or Word Clouds)
- This project involves building notebooks that interactively guide scholars through sample analyses and inviting further engagement using a fill-in-the-blanks 'madlibs' approach.

Platform Development: Toolkit > Cloud > Notebook

The Archives Unleashed team has been tackling the challenge of scale by tools to bridge the gap between vast web archive collections (hundreds of gigabytes or even terabytes) and intuitive, easily-accessible analytics tools.



To provide guidance, in previous work we proposed a model for scholarly interactions that starts with a question and proceeds iteratively through four main steps. Community feedback and consultation with librarians, scholars, computer scientists, and other stakeholders [2] has:

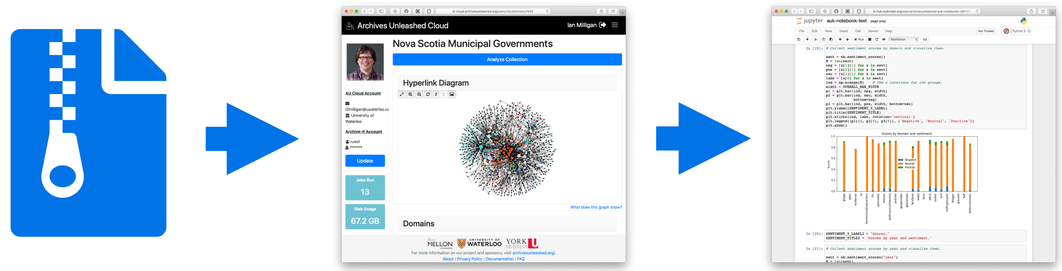
- Provided environmental scan of needs
- Informed technical direction
- Identified that scholars are frequently interested in the same types of derivatives as starting points to their analyses: domain crawl distributions, full text and associated metadata, and the domain-to-domain network graph
- Demonstrated that scholars are often unsure where to even begin….

References

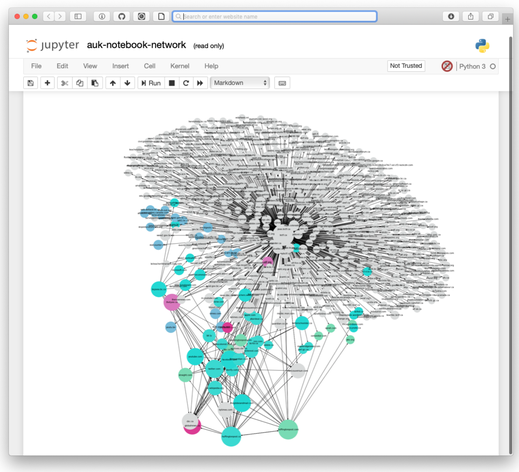
[1] J. Lin et al. 2017. Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. J. Comput. Cult. Herit. 10, 4.
[2] I. Milligan et al. 2019. Building Community and Tools for Analyzing Web Archives through Datathons. In JCDL.

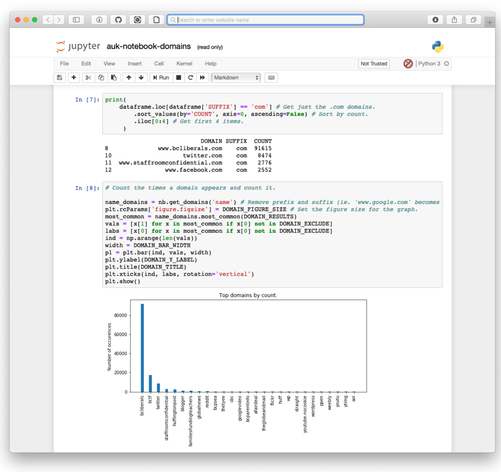
The Archives Unleashed Notebook

- Jupyter notebooks allow for code fragments and execution results (i.e., graphs and figures) to be side by side to support rapid interaction.
- Our notebooks guide scholars through sample analyses, each corresponding to a potential research question, using a fill-in-the-blanks 'madlibs' approach.
- Notebooks work with derivative datasets from the Archives Unleashed Cloud.
- Opens up integration with cloud storage such as Google Drive (via Colab) or Amazon S3.



Web Archive (Stored as WARC Files) → Processed with the Archives Unleashed Cloud → Analyzed with the Archives Unleashed Cloud Notebook





Conclusions

- Our goal is simple: to show how a web archive collection can support scholarly inquiry.
- We do so by making the transition between web archive derivative data and research questions "gentler" — i.e. guiding scholars from guided exploration to creative expression.
- Such engagement is the key to building a robust community of scholars and content curators, which will help ensure that our digital world can be studied well into the future.

Thanks to our funders and supporters!

THE ANDREW W. MELLON FOUNDATION
Social Sciences and Humanities Research Council of Canada
Conseil de recherches en sciences humaines du Canada
Canada
UNIVERSITY OF WATERLOO
YORK UNIVERSITÉ UNIVERSITY
START SMART LABS