

Web Archives Analysis at Scale with the Archives Unleashed Cloud

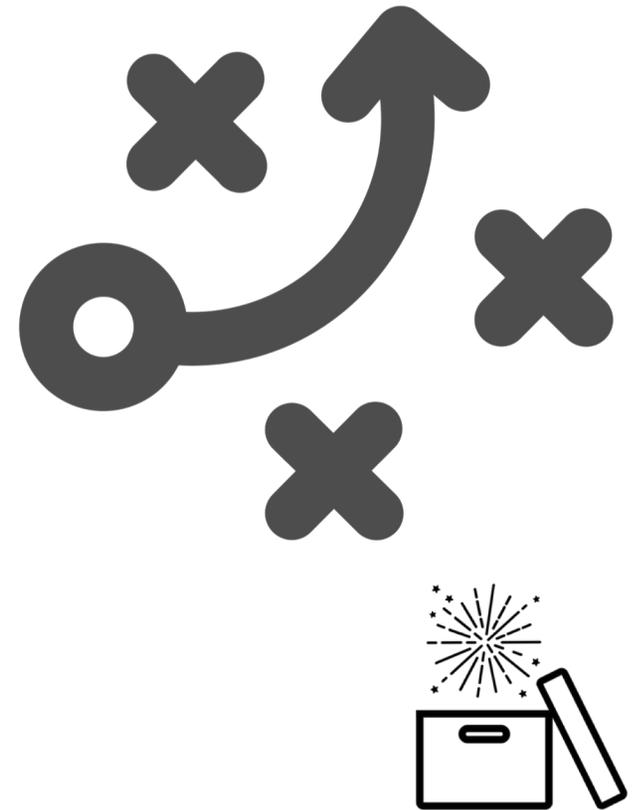
Nick Ruest (York University)

Ian Milligan (University of Waterloo)



Plan for The Talk

- Introduction
- The Problem
- Our Interdisciplinary Team
- Analysis at Scale with Archives Unleashed Tools
 - Toolkit
 - Cloud
 - Notebooks
- Caveats
- Conclusions

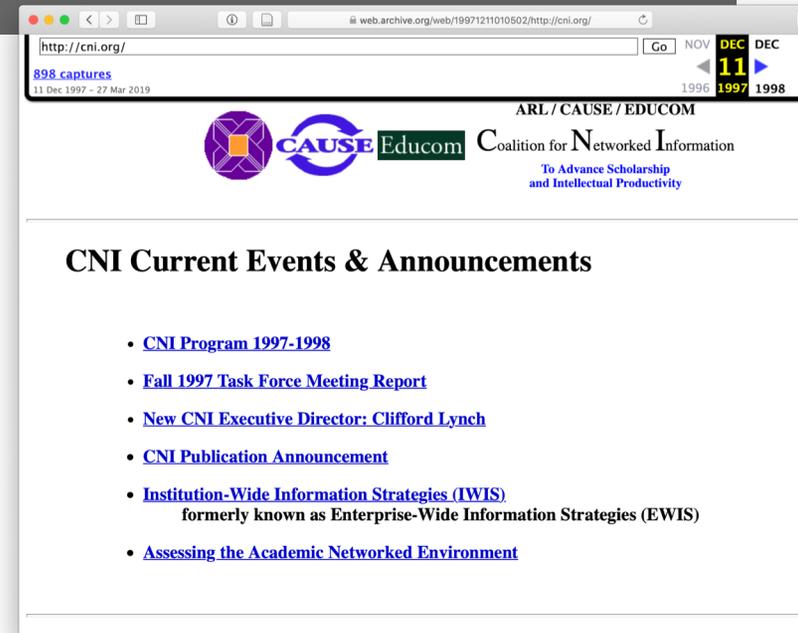


The Problem



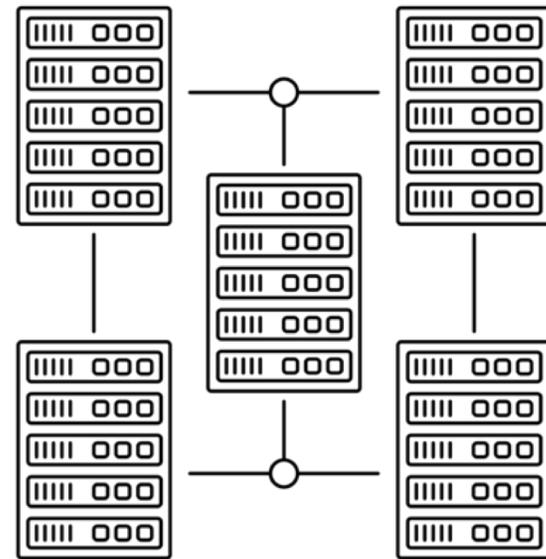
Web Archives are Important

- How we preserve and disseminate cultural information has dramatically changed;
- Since 1996, and the advent of web archiving at the Internet Archive and national libraries, how we remember has dramatically altered:
 - In scope
 - In speed
 - In scale



... in other words ...

- More data than ever before is being preserved;
- And it's being saved and delivered to us in very different ways...



All historians who want to study periods after the 1990s will have to use web archives.



**You can't study the 1990s
without web archives.**

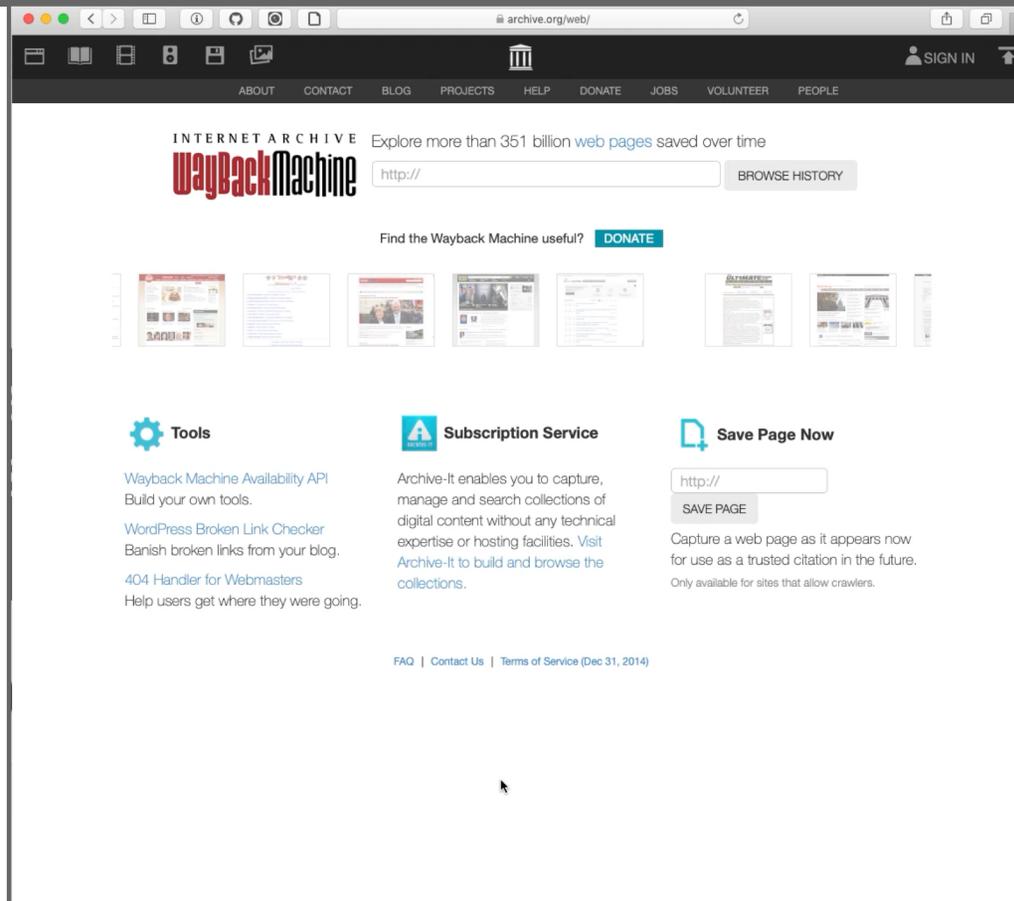
And historians aren't ready...



Access at scale has lagged.



Option One: The Wayback Machine



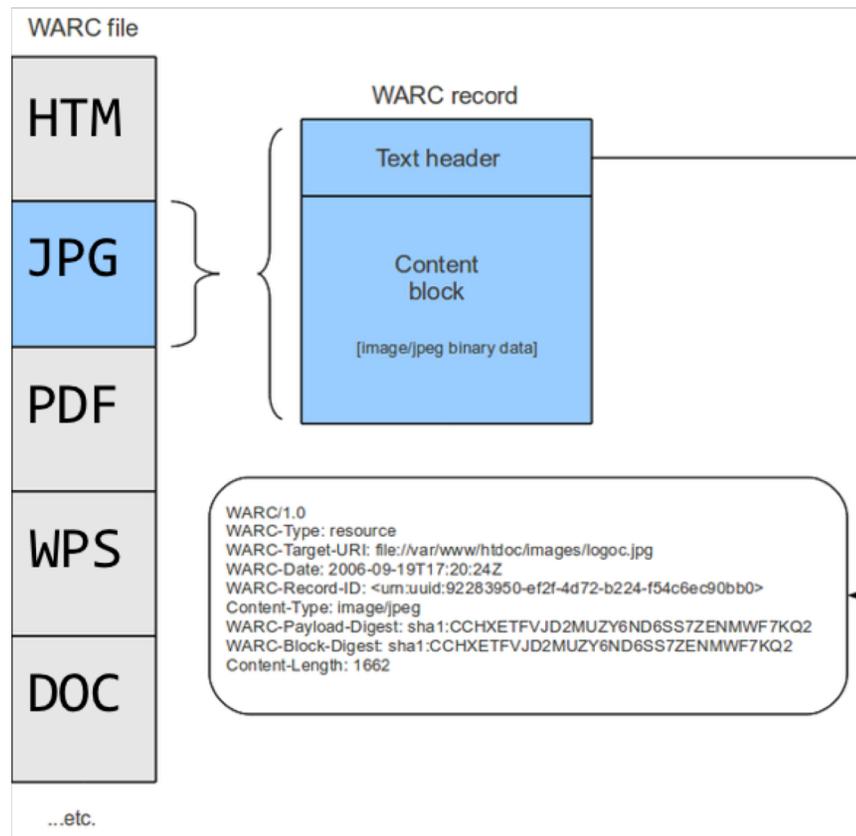
Option Two: Working with the Underlying Data



WebARChive (WARC) File



Option Two: Working with the Underlying Data



Option Two: Working with the Underlying Data

● Potential

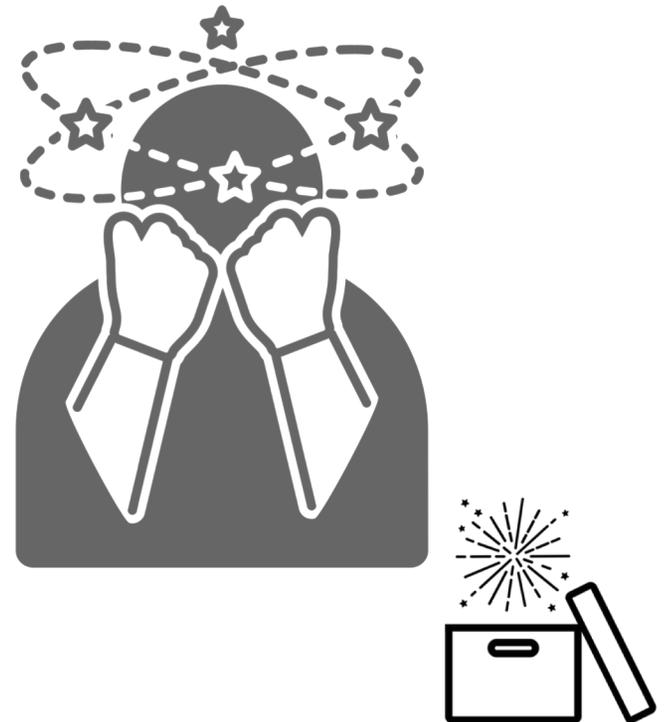
- Text analysis at scale;
 - Finding particular mentions of keywords, people, organizations, concepts, etc. over time
 - Finding patterns over time (i.e. culturomics or other forms of cultural analytics)
 - Other text mining applications
- Network analysis at scale;
 - Leveraging hyperlinks to see how people link to each other differently over time;
 - Finding pages of interest through historical applications of PageRank and other network concepts;
- Moving between “distant” and “close” scales



Option Two: Working with the Underlying Data

- **Downsides**

- Difficulty of tools to work with WARCs (humanists might be used to working with text at scale... they're not used to WARC files);
- Size of datasets (small web archives are in the tens of GBs; medium ones are in the 100GB-1TB range; large ones can easily begin to exceed 10TB);
- Lack of a research community.



In other words, researchers need to explore web archives beyond the Wayback Machine... but the tools and infrastructure aren't there.



**Enter the Archives
Unleashed Project**

Archives 
Unleashed

Our Team



Ian Milligan
Historian, University of Waterloo



Nick Ruest
Librarian/Archivist, York University



Jimmy Lin
Computer Scientist, University of Waterloo



Our Mission Statement

Archives Unleashed aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past.



So what do we do?



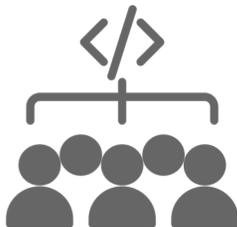
Archives Unleashed Projects



Archives Unleashed Toolkit



Archives Unleashed Cloud

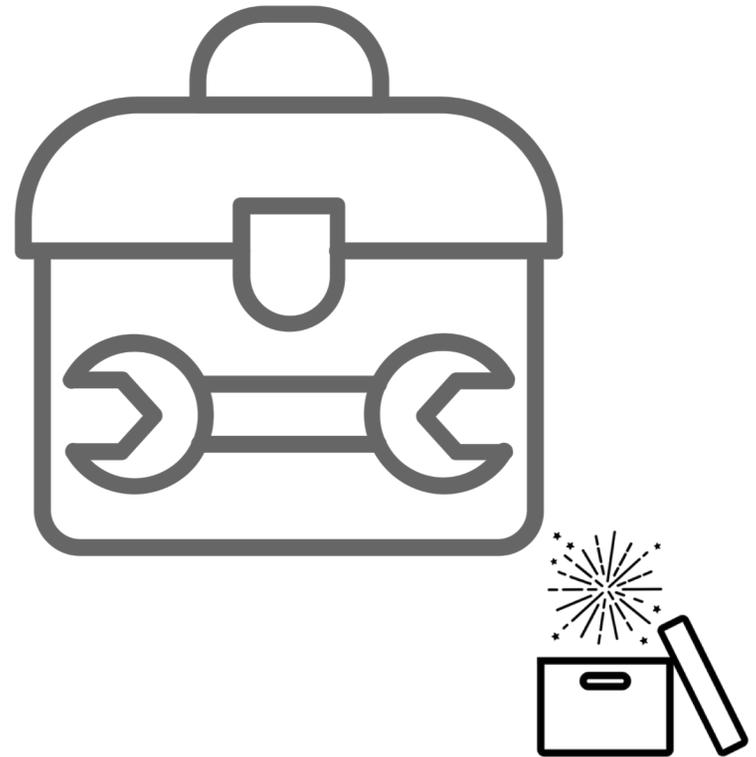


Archives Unleashed Datathons



Archives Unleashed Toolkit

- An open-source platform for analyzing web archives with Apache Spark;
- Scalable
 - Can work on a powerful cluster
 - Can work on a single-node server
 - Can work on a laptop (on MacOS, Linux, or on Windows with a Linux VM)
 - Can work on a Raspberry Pi for all your personal web archiving analysis needs

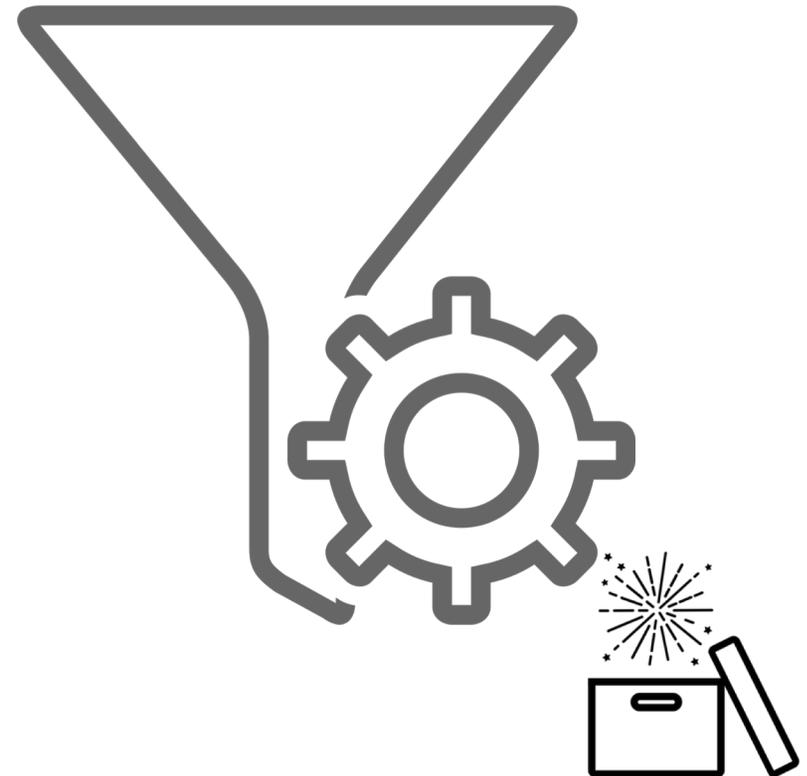


Using the Toolkit is based on the
Filter-Analyze-Aggregate-Visualize (FAAV) Cycle



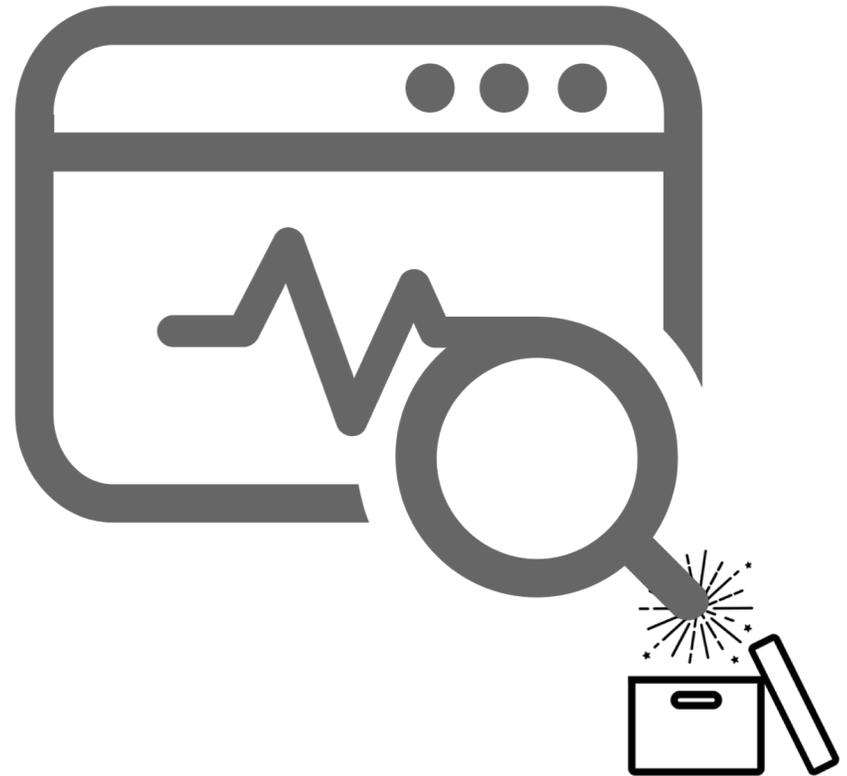
Filter

- Filter down content
 - Focus on a particular range of crawl dates;
 - Focus on a particular domain;
 - Content-based filter (“global warming”) or those who link to a given site
- Can be nested - i.e. pages from 2012 from liberal.ca that link to conservative.ca and contain the phrase “Keystone XL”



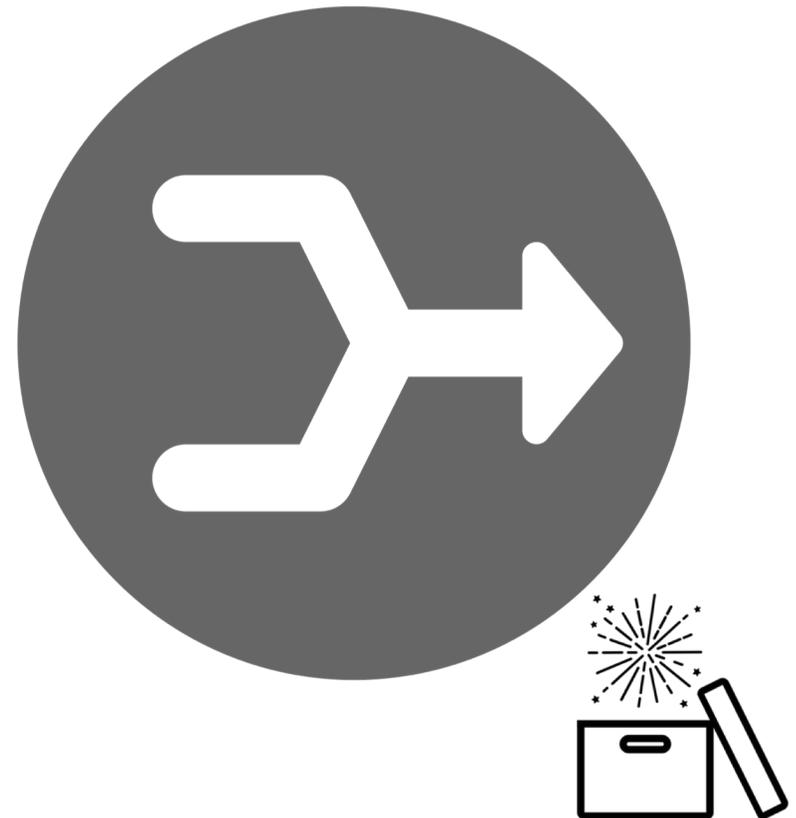
Analyze

- After filtering, want to perform analysis – extracting information of interest.
- Such as:
 - Links and associated anchor text?
 - Tagging or extracting named entities?
 - Sentiment analysis.
 - Topic modeling.



Aggregate

- Summarize the output of the analysis from the previous step.
 - Counting
 - How many times is Jack Layton or Barack Obama mentioned?
 - How many links are there from one domain to another?
- Finding maximum (page with most incoming links?)
- Average (average sentiment about “Barack Obama” or “Donald Trump”)



Visualize

- **Output data as a visualization**
 - Tables of results
 - External applications (i.e. GEXF files for Gephi)



Great!
So why doesn't everybody use the Toolkit?!?!



Our Cutting Edge Interface

```
1. ssh
x fsevent_watch #1 x bash #2 x ssh #3
our platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://rho.library.yorku.ca:4040
Spark context available as 'sc' (master = local[*], app id = local-1553805629588).
Spark session available as 'spark'.
Welcome to

 version 2.3.2

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_161)
Type in expressions to have them evaluated.
Type :help for more information.

scala> :paste
// Entering paste mode (ctrl-D to finish)

import io.archivesunleashed._
import io.archivesunleashed.matchbox._

RecordLoader.loadArchives("example.arc.gz", sc)
  .keepValidPages()
  .keepDomains(Set("www.archive.org"))
  .map(r => (r.getCrawlDate, r.getDomain, r.getUrl, RemoveHTML(r.getContentString)))
  .saveAsTextFile("plain-text-domain/")
```

In other words...

We have a wonderful platform that takes WARC files and converts them into formats that are familiar to digital humanists, computational social scientists, systems librarians, digital archivists, and beyond..

.. but you basically need to be a developer to run the simplest of commands (despite ample documentation and outreach... the command line interface is a bridge too far).

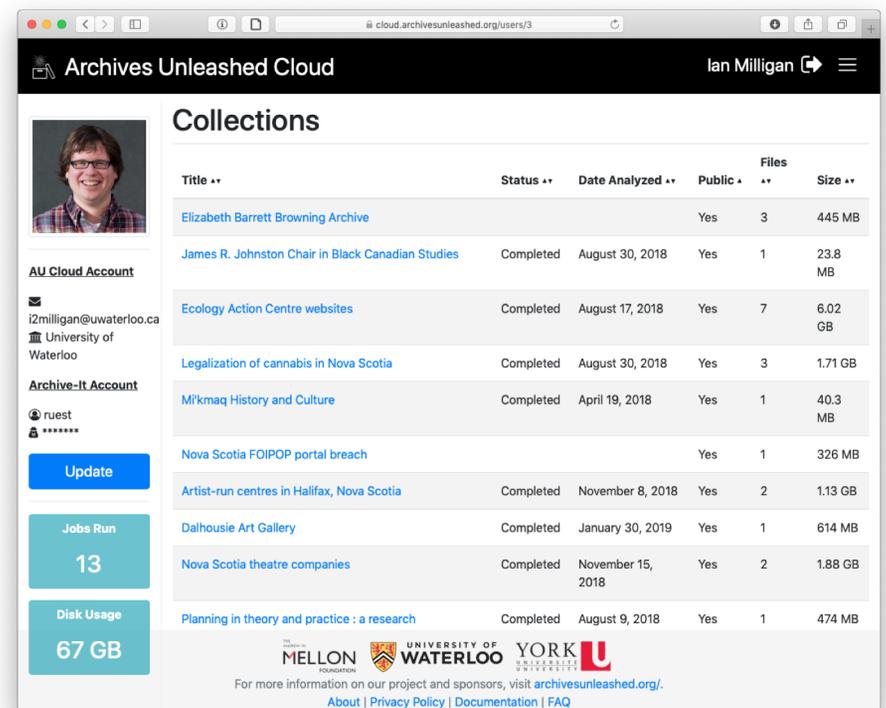


Enter the Archives Unleashed Cloud



Archives Unleashed Cloud

- A web-based front end for working with the Archives Unleashed Toolkit;
- Runs on our central servers or you can run one yourself;
- Uses WASAPI – Web Archives Systems API – to transfer data
 - Currently Archive-It supported;
 - We are exploring integration with WebRecorder.io and other WASAPI endpoints
- Generates a basic set of research derivatives for scholars to work with



The screenshot shows the Archives Unleashed Cloud interface for user Ian Milligan. The page includes a user profile section on the left with contact information and account management options. The main content area displays a table of collections with columns for Title, Status, Date Analyzed, Public, Files, and Size.

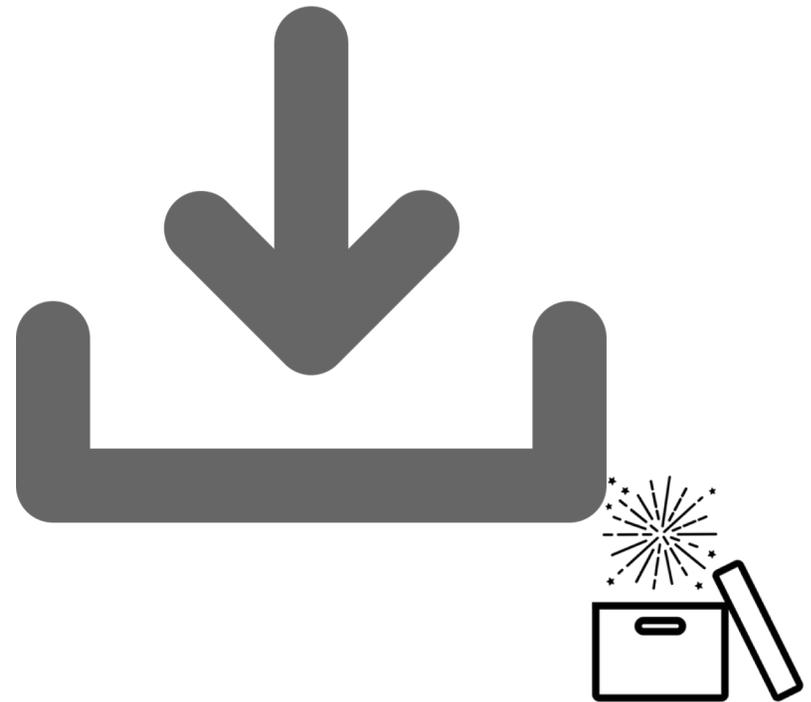
Title	Status	Date Analyzed	Public	Files	Size
Elizabeth Barrett Browning Archive			Yes	3	445 MB
James R. Johnston Chair in Black Canadian Studies	Completed	August 30, 2018	Yes	1	23.8 MB
Ecology Action Centre websites	Completed	August 17, 2018	Yes	7	6.02 GB
Legalization of cannabis in Nova Scotia	Completed	August 30, 2018	Yes	3	1.71 GB
Mikmaq History and Culture	Completed	April 19, 2018	Yes	1	40.3 MB
Nova Scotia FOIPOP portal breach			Yes	1	326 MB
Artist-run centres in Halifax, Nova Scotia	Completed	November 8, 2018	Yes	2	1.13 GB
Dalhousie Art Gallery	Completed	January 30, 2019	Yes	1	614 MB
Nova Scotia theatre companies	Completed	November 15, 2018	Yes	2	1.88 GB
Planning in theory and practice : a research	Completed	August 9, 2018	Yes	1	474 MB

At the bottom of the interface, logos for the Mellon Foundation, University of Waterloo, and York University are displayed, along with a link to the project website.

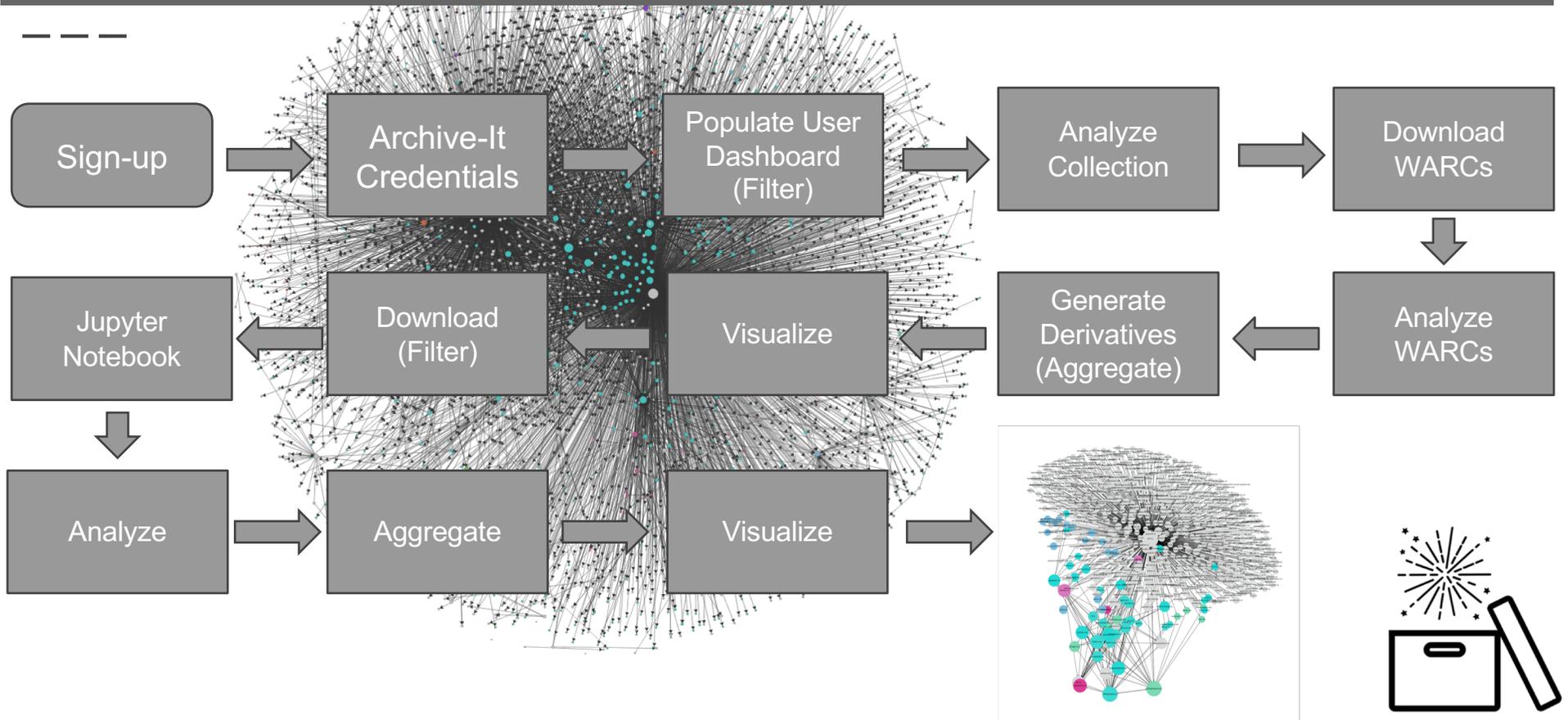


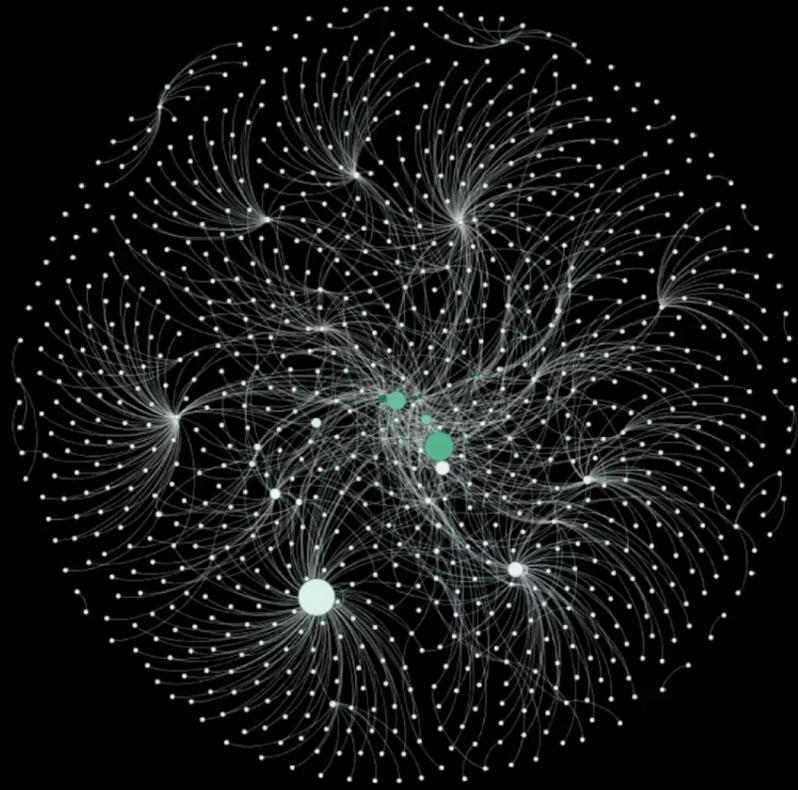
Archives Unleashed Cloud

- Download options for each collection
 - Full text of a web archive;
 - Full text of the top-ten most popular domains in a web archive;
 - Network diagram with characteristics pre-computed (Gephi);
 - Raw network diagram (origin/destination/weight);
 - Domain frequency statistics



How it works



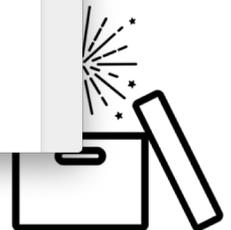
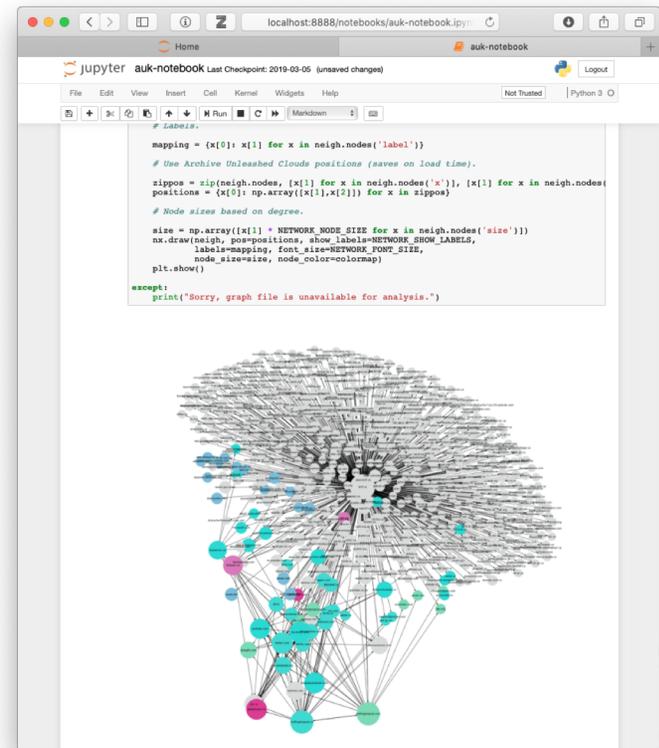


But where does our platform end... And the researcher begin?



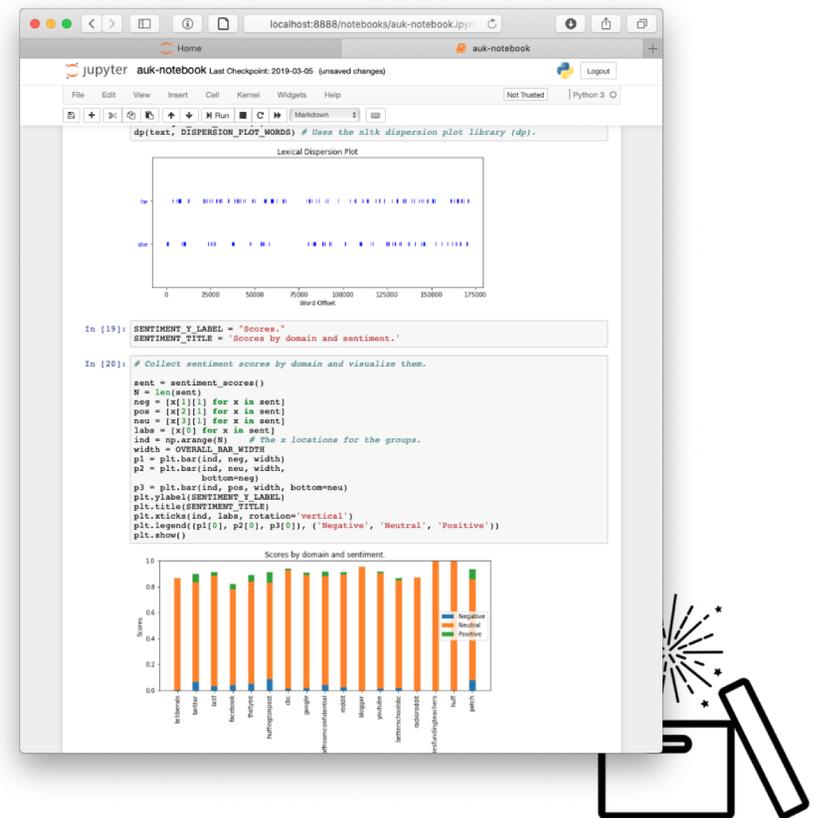
Archives Unleashed Cloud Notebooks

- Jupyter Notebooks
- One for each derivative
- A “mad-libs” approach - fill in the blanks with the variables (domains, dates, collections, etc.) that you are interested in, and it does basic computations for you
- Still under development
- Bundled with data – download, run, explore data in your browser

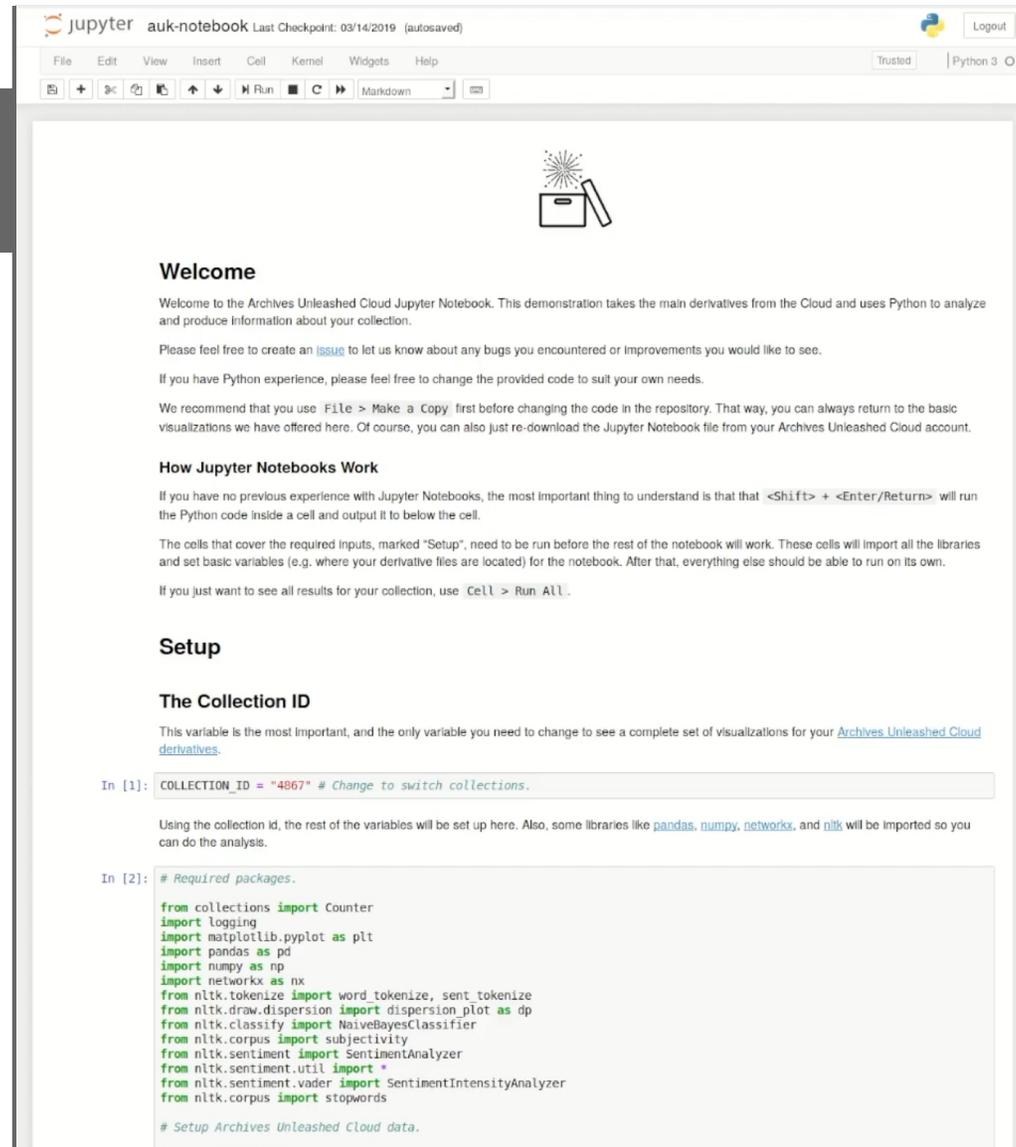


Archives Unleashed Cloud Notebooks

By giving researchers these notebooks, with data, can we begin to jumpstart the process of research question creation and imagining what they can do with the data.



Notebook Demo



The image shows a Jupyter Notebook interface for a notebook named "auk-notebook". The top bar includes the Jupyter logo, the notebook name, and a "Last Checkpoint" timestamp of "03/14/2019 (autosaved)". A "Logout" button is visible in the top right. Below the top bar is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. A toolbar contains icons for file operations and execution. The main content area displays a "Welcome" message, followed by instructions on how to use the notebook, including how to run code cells and how to set up the environment. The "Setup" section includes a code cell that sets the "COLLECTION_ID" variable and another code cell that imports various Python libraries like pandas, numpy, networkx, and nltk.

Welcome

Welcome to the Archives Unleashed Cloud Jupyter Notebook. This demonstration takes the main derivatives from the Cloud and uses Python to analyze and produce information about your collection.

Please feel free to create an [issue](#) to let us know about any bugs you encountered or improvements you would like to see.

If you have Python experience, please feel free to change the provided code to suit your own needs.

We recommend that you use `File > Make a Copy` first before changing the code in the repository. That way, you can always return to the basic visualizations we have offered here. Of course, you can also just re-download the Jupyter Notebook file from your Archives Unleashed Cloud account.

How Jupyter Notebooks Work

If you have no previous experience with Jupyter Notebooks, the most important thing to understand is that that `<Shift> + <Enter/Return>` will run the Python code inside a cell and output it to below the cell.

The cells that cover the required inputs, marked "Setup", need to be run before the rest of the notebook will work. These cells will import all the libraries and set basic variables (e.g. where your derivative files are located) for the notebook. After that, everything else should be able to run on its own.

If you just want to see all results for your collection, use `Cell > Run All`.

Setup

The Collection ID

This variable is the most important, and the only variable you need to change to see a complete set of visualizations for your [Archives Unleashed Cloud derivatives](#).

```
In [1]: COLLECTION_ID = "4867" # Change to switch collections.
```

Using the collection id, the rest of the variables will be set up here. Also, some libraries like [pandas](#), [numpy](#), [networkx](#), and [nltk](#) will be imported so you can do the analysis.

```
In [2]: # Required packages.
from collections import Counter
import logging
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import networkx as nx
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.draw.dispersion import dispersion_plot as dp
from nltk.classify import NaiveBayesClassifier
from nltk.corpus import subjectivity
from nltk.sentiment import SentimentAnalyzer
from nltk.sentiment.util import *
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from nltk.corpus import stopwords

# Setup Archives Unleashed Cloud data.
```



Archives Unleashed Usage Statistics

- Users: 144
- Collections: 792
- Files: 1,077,392
- Jobs completed: 4431
- Job time: 10631h, 29m, 42s (1.25yrs!)
- Longest job: 590h, 31m, 49s
- Largest collection: 17.6T (compressed)
- Data analyzed: 159T (compressed)

(as of this morning!)



Finally, we aim to build community around web archives.



Archives Unleashed Datathons

Helping to lower barriers;

Bringing people interested in web archiving (both collection + analysis) together;

Establishing a community through online communication and in-person work and social events;

Establishing a true community of practice around web archiving practice.



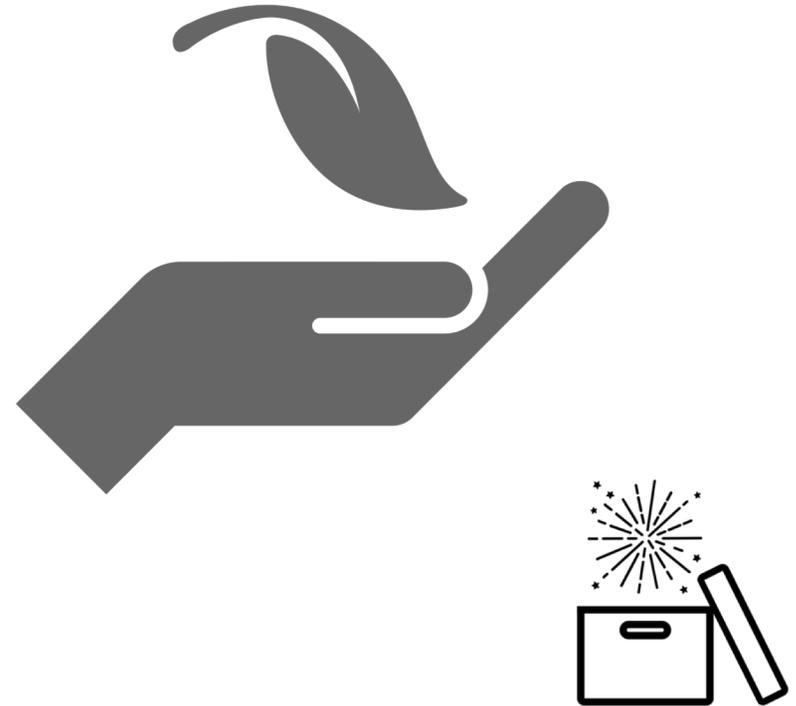
Archives Unleashed Datathons

- To date we've run (in this sequence) a series of datathons in **Toronto, Vancouver, and Washington DC**
 - a previous iteration had four events as well
- Gaining more experience with working with cultural heritage at scale



What's Next?

- **Sustainability** has been baked into our grant from the very start (thanks Mellon!).
 - Ryan Deschamps, Samantha Fritz, Jimmy Lin, Ian Milligan, and Nick Ruest. “The Cost of a WARC: Analyzing Web Archives in the Cloud.” *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, Vol. 19 (2019). Forthcoming.
 - Costs USD\$7/TB to process using the Archives Unleashed Toolkit.



What's Next?

- Actual processing costs are relatively affordable – approx. US\$7/TB to process WARC_s and generate derivatives.
 - Large collection like University of Toronto's “Canadian Political Parties and Interest Groups” would cost under US\$30 to process and generate all of our derivative types seen in the Cloud.
- But of course, computing costs aren't the crux...

The Cost of a WARC

Size	Count
≥ 1 GB, < 10 GB	10
≥ 10 GB, < 100 GB	18
≥ 100 GB, < 1 TB	15
≥ 1 TB	5
Total	48

Table 1: Sizes of the collections in our study.

Derivative	all	L	M	S
domain distribution	32	25	27	36
full text	34	28	35	34
webgraph	36	34	36	36
total	102	87	98	106

Table 2: Processing times per GB in seconds.

4 FINDINGS AND DISCUSSION

In the Archives Unleashed Project thus far, we have processed over 150 TB of web archives from our content partners. For this study we focused on 57 collections analyzed in early 2018 from six different Canadian universities, collected using the Archive-It platform. We excluded from analysis nine collections smaller than one gigabyte, as they are too small to benefit from processing by AUT (leaving 48 in total). The largest collection, at 4.3 TB in size, was the Canadian Government Information Collection (from the University of Alberta); the smallest collection, at 1.2 GB, was the University of Victoria's academic calendar. The complete distribution of collection sizes is shown in Table 1; all size figures are given in base 10 and all collection sizes refer to the raw, compressed WARC_s.

We have automated the process model described in the previous section, with scripts that start up virtual machine instances to perform the various stages of processing. For data ingestion, we used the data transfer functionalities of WASAPI (Web Archiving Systems API) provided by Archive-It. Our analysis is derived from the execution logs of these scripts.

In Table 2, we show the processing time (in seconds) per GB of source web archive for each derivative as well as the total. The column marked “all” shows analyses for all collections; we further break down results into large collections (larger than 1 TB, denoted “L”), medium collections (between 100 GB and 1 TB, denoted “M”), and small collections (less than 100 GB, denoted “S”). From these results, we make a few observations: Despite the different nature of these derivatives, running times are quite similar because the analytical queries are all dominated by the time to scan the entire collection. Extracting the webgraph is more computationally intensive, but not substantially more so. We see that total processing time for all three derivatives drops as the collection size increases, likely because the startup costs associated with AUT are amortized over longer running times. As expected, there exists a linear correlation between the raw collection size and the total amount of time required to generate all three derivatives; this is shown in Figure 2, where we observe an R^2 value of 0.970.

<https://github.com/WASAPI/Community-data-transfer-api>

WOODSTOCK '97, July 1997, El Paso, Texas USA

Figure 2: Scatter plot between collection size and total processing time, illustrating a linear relationship.

Derivative	all	L	M	S
domain distribution (KB)	0.95	0.51	0.98	1.01
full text (MB)	78.5	97.6	102.1	62.4
webgraph (KB)	76.9	85.8	122.6	50.9

Table 3: Derivative sizes per GB

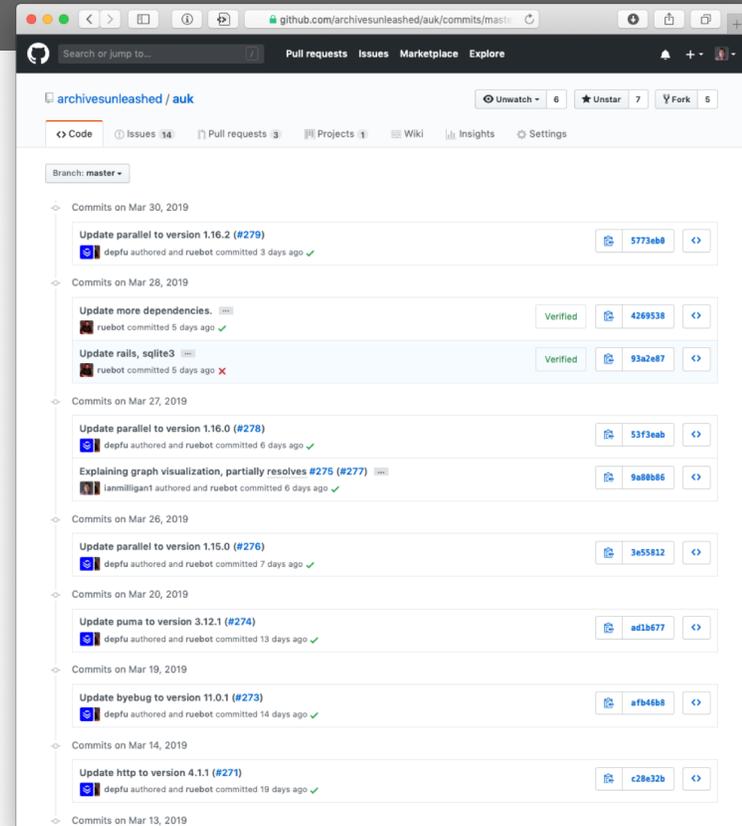
How large are these derivatives? The answer is shown in Table 3, which reports the sizes of the derivatives per GB raw archive; we report overall statistics as well as statistics broken into large, medium, and small collections (note the different units). These averages hide the fact that actual values vary by collection, depending on the nature of the crawl (e.g., wide multi-site crawls vs. narrow deep crawls, text-heavy vs. media-heavy sites, etc.). However, in rough terms, for a typical medium site, domain distribution data is usually less than 1 MB, the raw text is perhaps 10s GB, and the webgraph is 10s MB. These values support our observation that AUT provides a bridge between web archives and scholars' existing tools, since datasets of these sizes are well within the capabilities of modern laptops. Furthermore, the long-term preservation of these derivatives presents no serious challenges: they can be treated as first-class citizens in the scholarly community (e.g., given DCRs).

Next, our cost analysis is shown in Table 4, organized in the same manner as Table 2, showing the cost in USD per TB of raw web archive on Amazon's EC2 service. Based on available statistics, the instance type used in our experiments on Compute Canada aligns roughly with a c5-4xlarge instance, with 16 virtual cores and 68 GB memory, currently costing US\$0.68 per hour in the US East (Ohio) region. We assume per-minute billing (i.e., processing times are rounded up to the nearest minute) but do not account for instance startup costs. For consistency, we show cost per TB even for the small collections. These values report an macro-average, i.e., an average across individual collections. Note that our approach for computing these figures leads to inflated costs for small collections because they finish quickly (typically, only a few minutes).

All considered, a “bottom line” figure of US\$7 per TB for a typical analytics product is a fair summary of our findings. We argue that

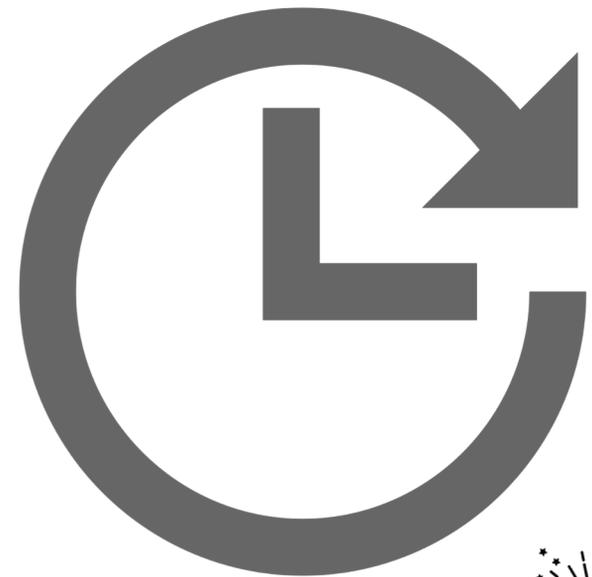
What's Next?

- Supported by Andrew W. Mellon Foundation; Compute Canada; Start Smart Labs; and some institutional support from Waterloo and York.
- Limitations (beyond computing costs):
 - Developer Time
 - Community Involvement
 - Sustainable Infrastructure



What's Next?

- We know how much it costs;
- We've forged good partnerships with institutions, including the Internet Archive, datathon hosts (Simon Fraser, Toronto, George Washington), International Internet Preservation Consortium, and others;
- Held consultations with research libraries + consortias; and
- Are exploring tangible partnerships to bring web archive analysis to a broader audience.



We look forward to
your questions and
thoughts.

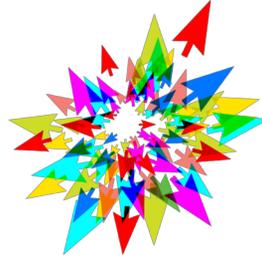


Thanks to our supporters!

THE
ANDREW W.

MELLON
FOUNDATION

compute | calcul
canada | canada



UNIVERSITY OF
WATERLOO



Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada



Links

- archivesunleashed.org
- cloud.archivesunleashed.org
- github.com/archivesunleashed
- slack.archivesunleashed.org
- news.archivesunleashed.org
- twitter.com/unleasharchives

