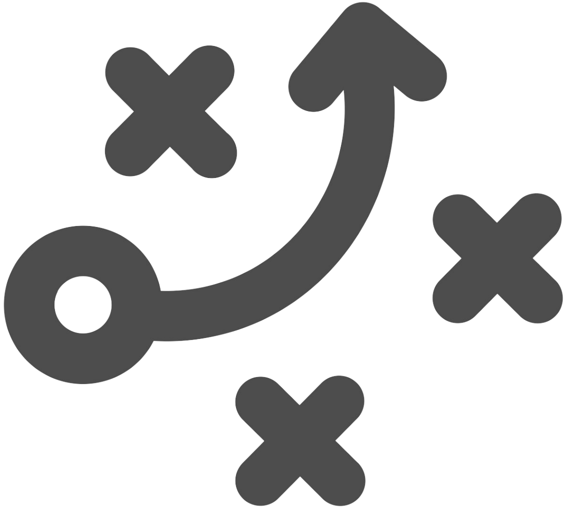
Lowering the Barrier to Access: The Archives Unleashed Project

Nick Ruest (York University)
Ian Milligan (University of Waterloo)

RESAW 2019 - Amsterdam

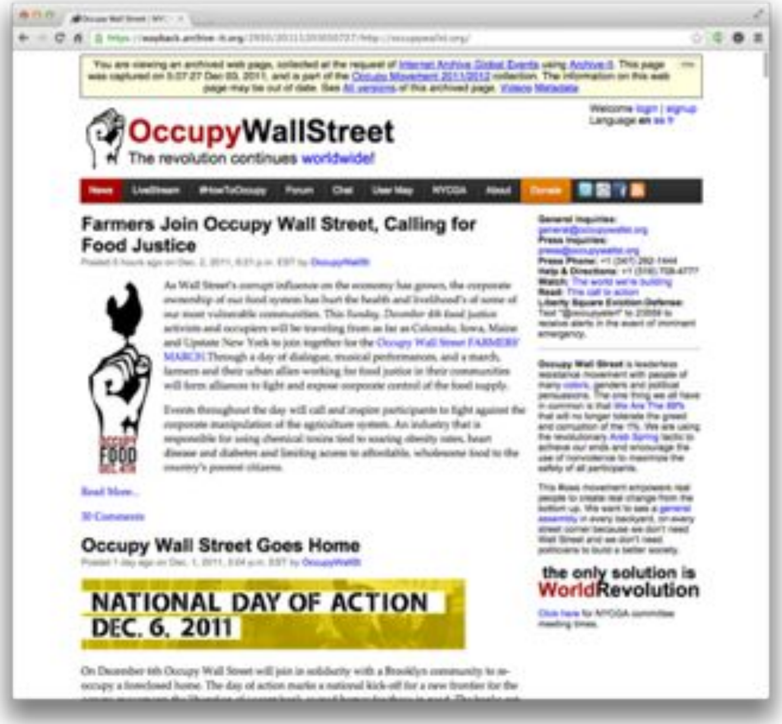
Plan for The Talk

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



The Problem

Why should we care about web archives?



Occupy Wall Street | NYC
https://wayback.archive-it.org/2950/20111203050727/http://occupywallst.org/

You are viewing an archived web page, collected at the request of Internet Archive Global Events using Archive-It. This page was captured on 5:07:27 Dec 03, 2011, and is part of the Occupy Movement 2011/2012 collection. The information on this web page may be out of date. See All versions of this archived page. Videos Metadata

Welcome login | signup
Language en es fr

OccupyWallStreet
The revolution continues worldwide!

Farmers Join Occupy Wall Street, Calling for Food Justice

Posted 6 hours ago on Dec. 2, 2011, 6:31 p.m. EST by OccupyWallSt

As Wall Street's corrupt influence on the economy has grown, the corporate ownership of our food system has hurt the health and livelihood's of some of our most vulnerable communities. This Sunday, December 4th food justice activists and occupiers will be traveling from as far as Colorado, Iowa, Maine and Upstate New York to join together for the Occupy Wall Street FARMERS' MARCH. Through a day of dialogue, musical performances, and a march, farmers and their urban allies working for food justice in their communities will form alliances to fight and expose corporate control of the food supply.

Events throughout the day will call and inspire participants to fight against the corporate manipulation of the agriculture system. An industry that is responsible for using chemical toxins tied to soaring obesity rates, heart disease and diabetes and limiting access to affordable, wholesome food to the country's poorest citizens.

Read More...

30 Comments

Occupy Wall Street Goes Home

Posted 1 day ago on Dec. 1, 2011, 3:04 p.m. EST by OccupyWallSt

NATIONAL DAY OF ACTION
DEC. 6, 2011

On December 6th Occupy Wall Street will join in solidarity with a Brooklyn community to re-occupy a foreclosed home. The day of action marks a national kick-off for a new frontier for the

General inquiries:
general@occupywallst.org
Press inquiries:
press@occupywallst.org
Press Phone: +1 (347) 292-1444
Help & Directions: +1 (516) 708-4777
Watch: The world we're building
Read: The call to action
Liberty Square Eviction Defense:
Text "@occupyalert" to 23559 to receive alerts in the event of imminent emergency.

Occupy Wall Street is leaderless resistance movement with people of many colors, genders and political persuasions. The one thing we all have in common is that We Are The 99% that will no longer tolerate the greed and corruption of the 1%. We are using the revolutionary Arab Spring tactic to achieve our ends and encourage the use of nonviolence to maximize the safety of all participants.

This #ows movement empowers real people to create real change from the bottom up. We want to see a general assembly in every backyard, on every street corner because we don't need Wall Street and we don't need politicians to build a better society.

the only solution is
WorldRevolution

Click here for NYCGA committee meeting times.

How we preserve and disseminate cultural information has dramatically changed;

Since ~1996, how we remember has dramatically altered:

- In scope
- In speed
- In scale
- And beyond...

What opportunities do they present?

The way that we preserve our culture is changing;

- **Scale**: Internet Archive has 635 billion URLs; 40PB of unique data (and non-Internet Archive collectors probably have about the same again).
- **Scope**: Data that never before would have been collected is now being collected about people who aren’t traditionally in the historical record.

Any researcher tackling post-1996 topics will realistically need to understand the vast arrays of text, image, etc. that comprise our modern cultural record.

The Wayback Machine isn’t enough; will need to explore data at scale.

You can’t study the 1990s without web archives.

And historians aren’t ready...

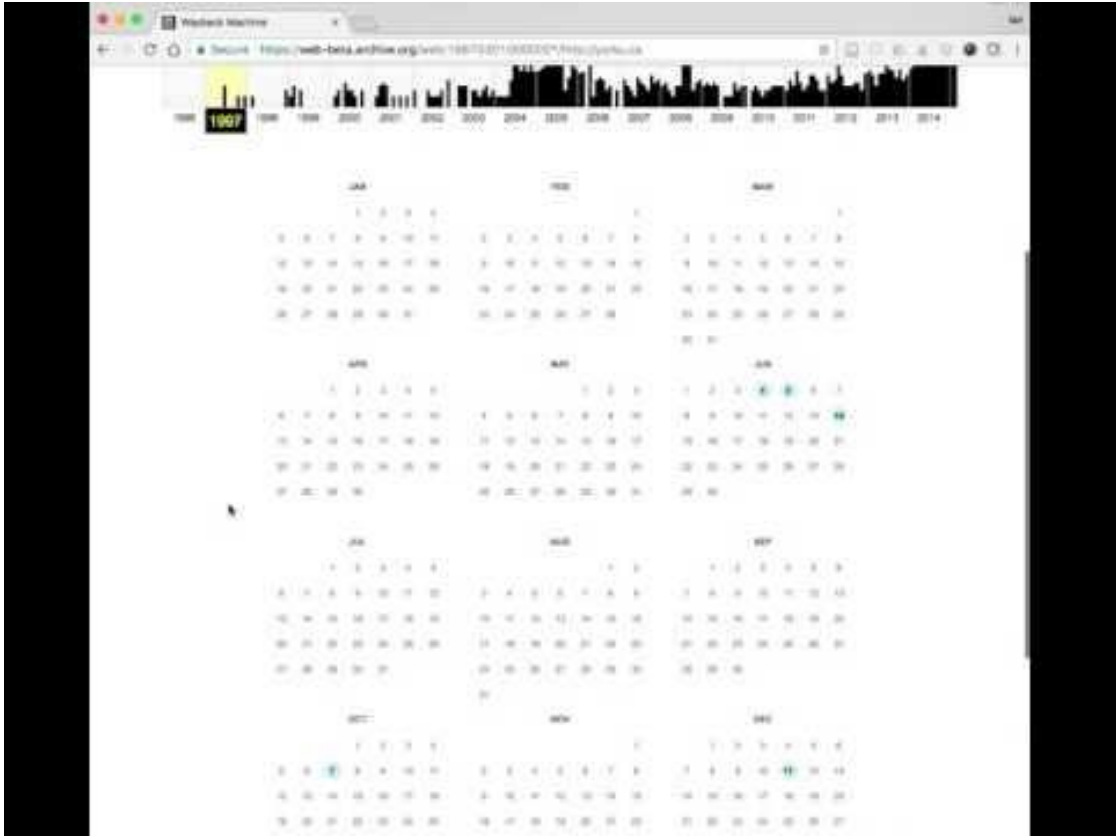
Why aren’t historians ready?

- Part of the reason that historians aren’t ready are **skills;**
  - **Skillset One**: Needing basics of working with data at scale (NLP, stats, basic data science skills)
  - **Skillset Two**: Understanding how data is constructed (i.e. why data was collected, why it wasn’t collected, basics of cleaning/normalizing data)
  - **Skillset Three**: Everything that goes into being a historian (i.e. what keeps us busy all the time right now!)
- The other part of the reason that historians aren’t ready are **platforms**;
- We need to try to help with both of these dimensions...

Let’s look at the **platform problem** first…

Access at scale has lagged.

Option One: The Wayback Machine



Option One: The Wayback Machine

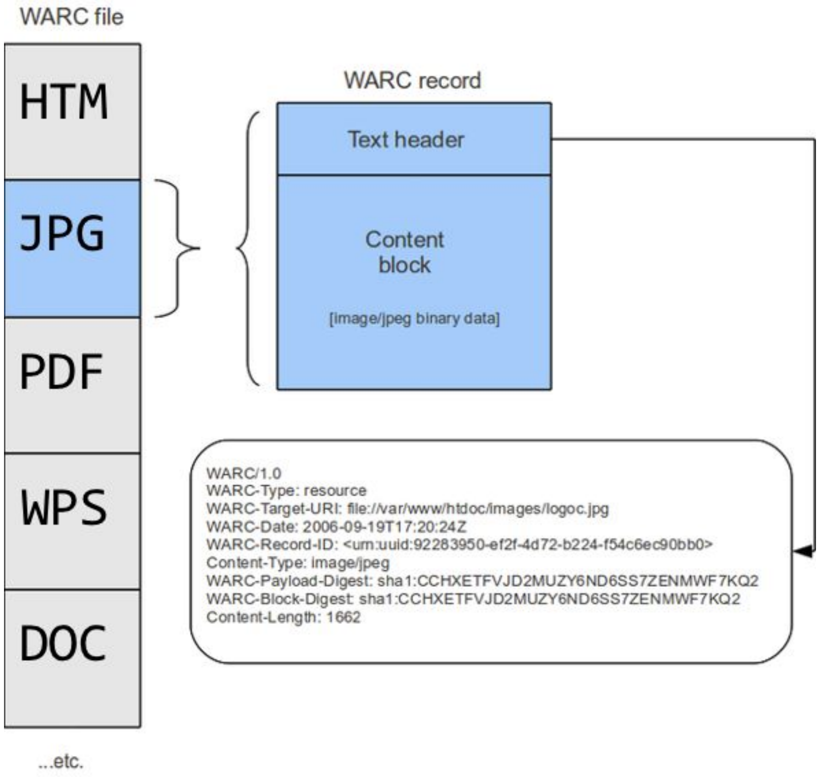
- Wayback Machine is great **if you know what you’re looking for**;
    - Ever-improving keyword search functionality
    - Represents a great stride in accessibility more generally
- But it necessarily isn’t suited for **more detailed research queries**:
    - (and it would be overkill for it to do so)
    - You may want to do complicated queries (i.e. websites that say X and link to Y);
    - You may want to do exploratory text mining;
    - You may want to work with images en masse;
    - Etc.

So to work with web archives at scale, you’re then turning to WARC files…

Option Two: Working with the Underlying Data

WebARChive (WARC) File

Option Two: Working with the Underlying Data



```
WARC/1.0
WARC-Type: resource
WARC-Target-URI: file://var/www/htdoc/images/logoc.jpg
WARC-Date: 2006-09-19T17:20:24Z
WARC-Record-ID: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>
Content-Type: image/jpeg
WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
WARC-Block-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
Content-Length: 1662
```

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.

… and neither are the skills.

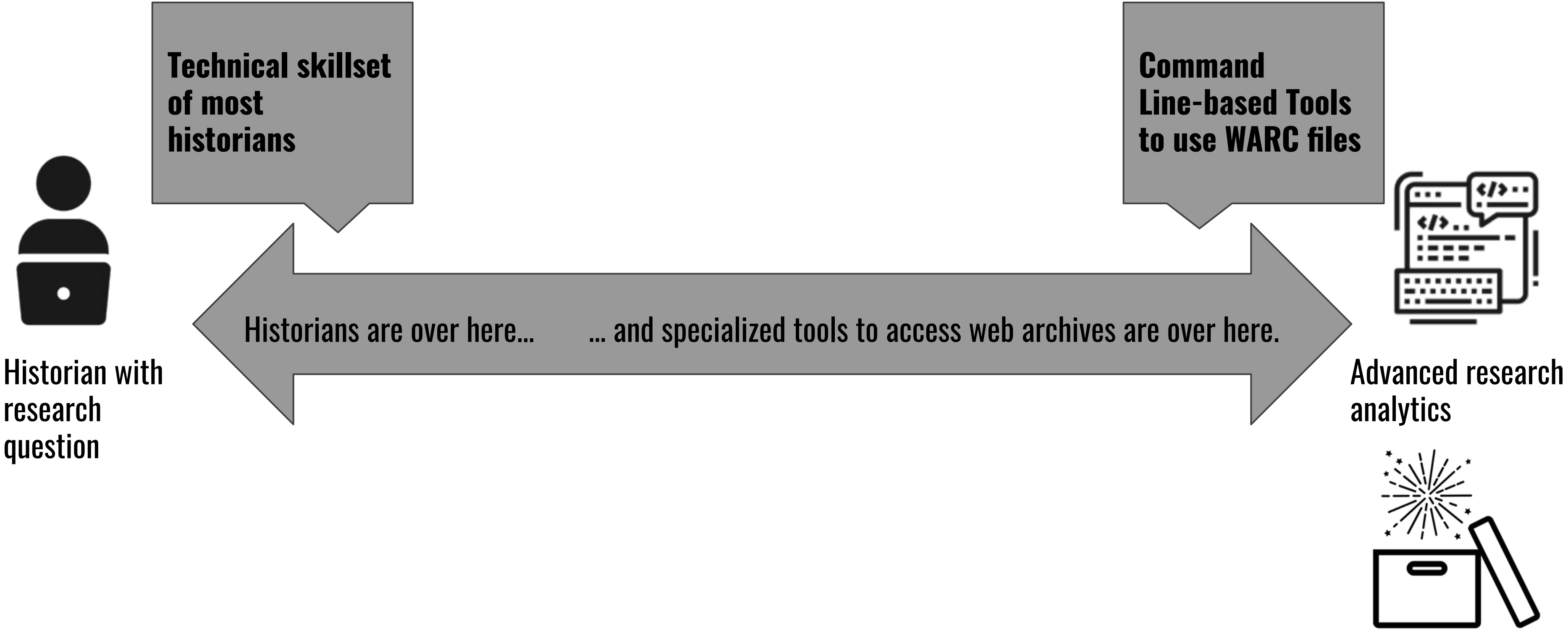
- Studies of our introductory historiography textbooks show this diminishing.
    - John Tosh, Pursuit of History
        - 1st, 2nd edition: “History by Numbers” (entire chapter)
        - By 5th edition, no quantitative history at all.
- **“Nevertheless, it is curious that at a time when both the use of and the breadth of humanities data is growing, quantitative skills … seem to no longer form a core component of our undergraduate history programmes.”** (James Baker)

https://blogs.bl.uk/digital-scholarship/2014/04/digital-history-and-the-death-of-quant.html

So let’s take stock…

- **Historians will need to understand and study the Web** in order to come to grips of history after the mid-1990s – not just for history of the Web, of course, but for the history of our society and culture as reflected on the Web
- **Existing tools like the Wayback Machine aren’t enough to tackle this problem**
- **Historians will need new skills** for working with and understanding data, plus their traditional competencies

In other words



Technical skillset of most historians

Command Line-based Tools to use WARC files

Historians are over here…

… and specialized tools to access web archives are over here.

Historian with research question

Advanced research analytics

Tools need to change and develop to move towards our users… **but what does that look like?**

Enter the Archives Unleashed Project

Archives Unleashed

Our Team

Ian Milligan
Historian, University of Waterloo

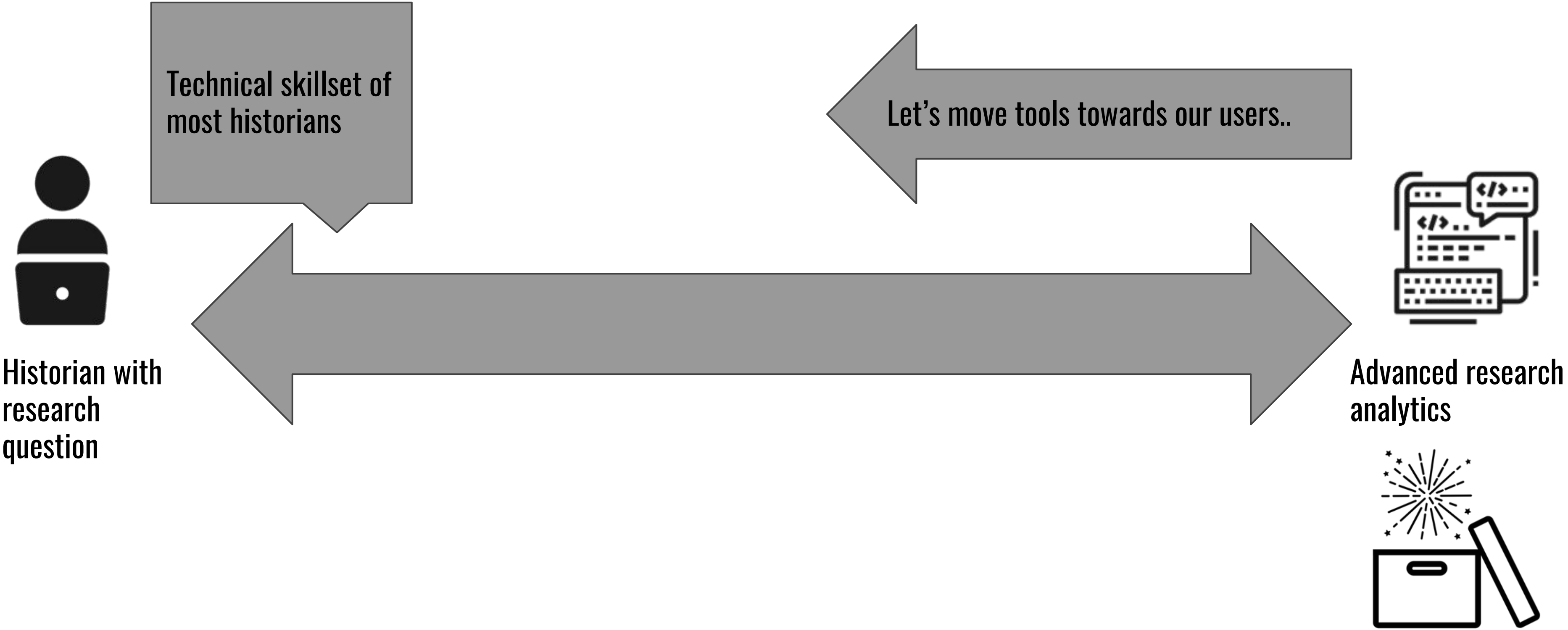
Nick Ruest
Librarian/Archivist, York University

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.

In other words



Technical skillset of most historians

Let’s move tools towards our users..

Historian with research question

Advanced research analytics

So how do we aim to do this?

Archives Unleashed Projects

Archives Unleashed Toolkit

Archives Unleashed Cloud

Archives Unleashed Datathons

Archives Unleashed Toolkit

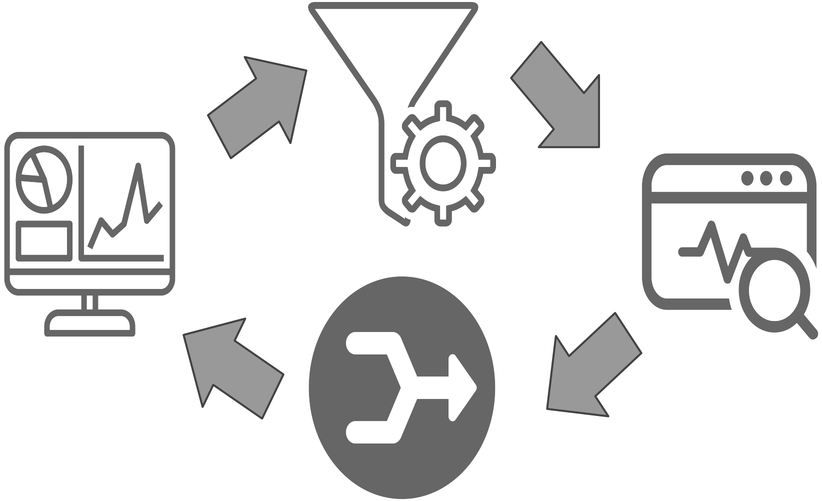
- An open-source platform for analyzing web archives with Apache Spark;
- Scalable
  - Can work on a powerful cluster
  - Can work on a single-node server
  - Can work on a laptop (on MacOS, Linux, or on Windows with a Linux VM)
  - Can work on a Raspberry Pi for all your personal web archiving analysis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

Using the Toolkit is based on the **F**ilter-**A**nalyze-**A**ggregate-**V**isualize (FAAV) Cycle

The FAAV Cycle

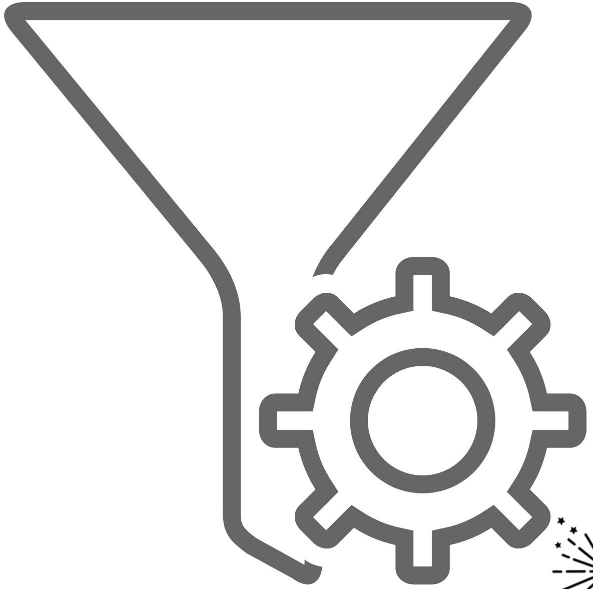
Beginning to conceptualize the workflow for historical work to do two things.

1. Inform tools development (to speak to the **platform** problem);
2. Inform pedagogical workflows and resources (to speak to the **skills** problem).



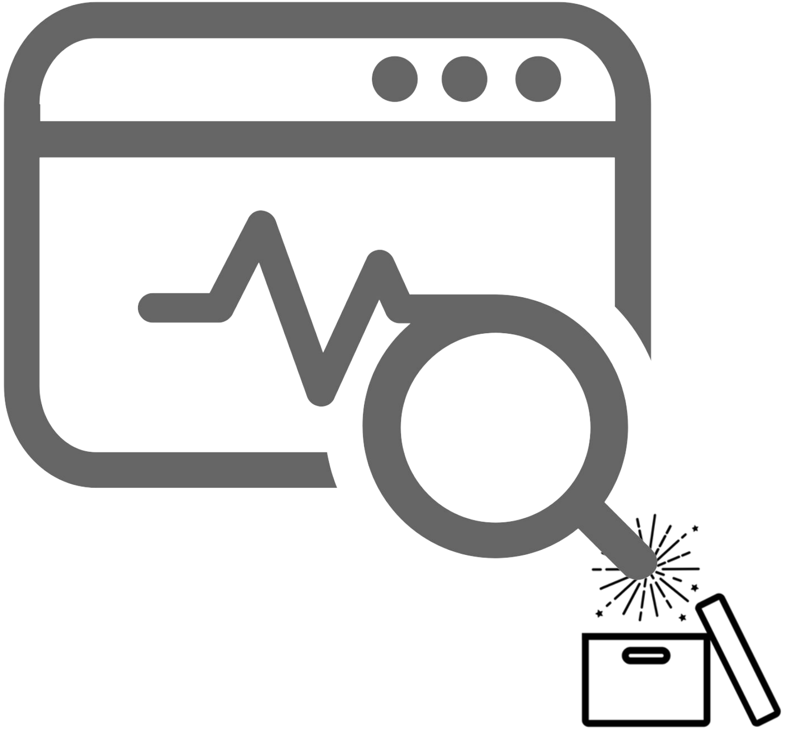
Filter

- 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.



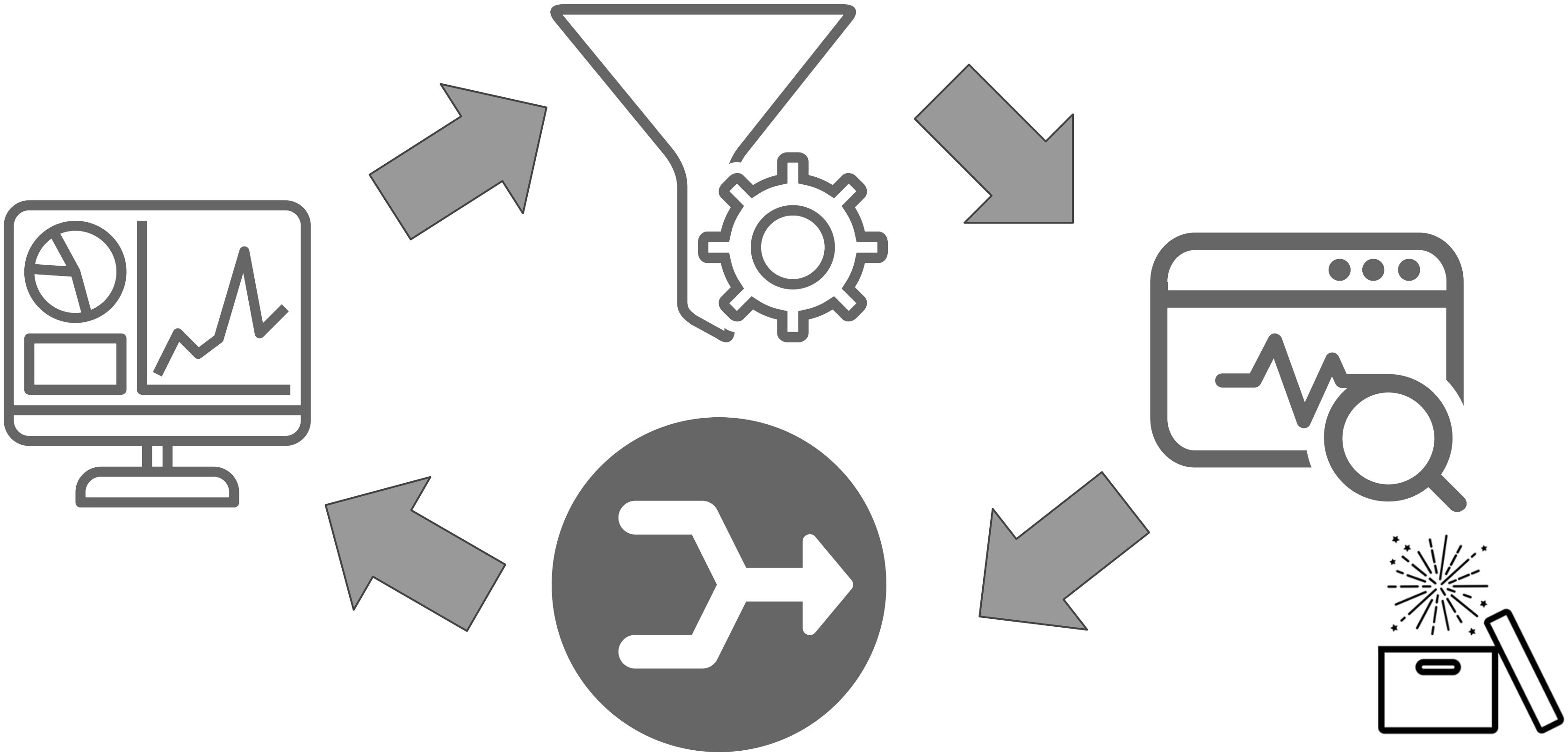
Aggregate

- 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)

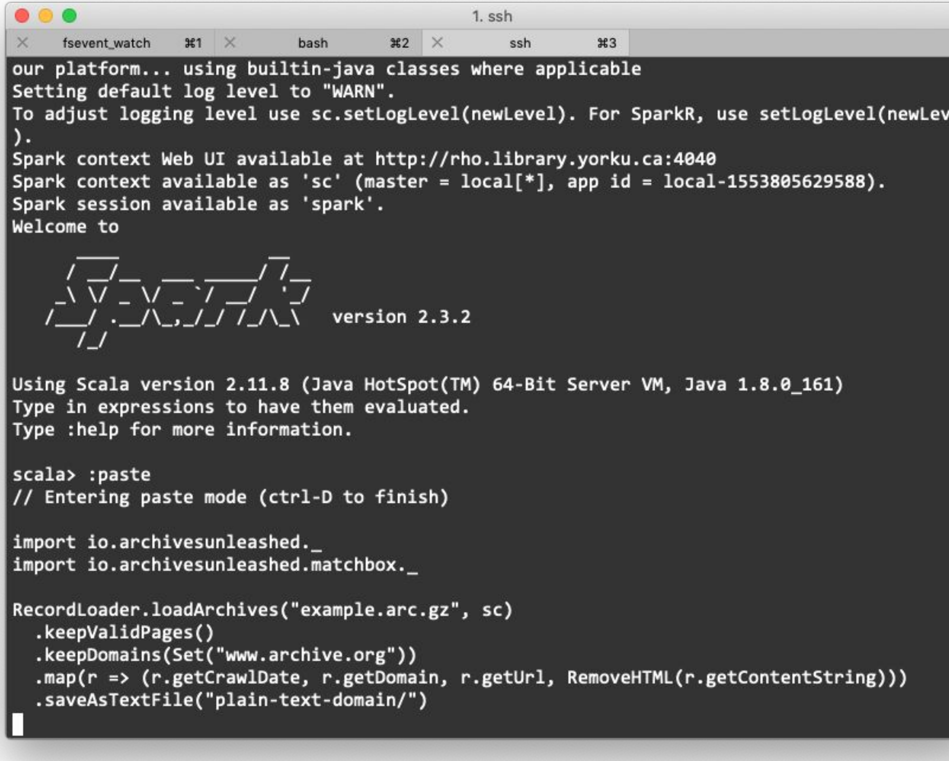
FAAV Cycle



Great!

So why doesn’t everybody use the Toolkit?!?!

Our Cutting Edge Interface



```
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(newLev
).
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/")
```

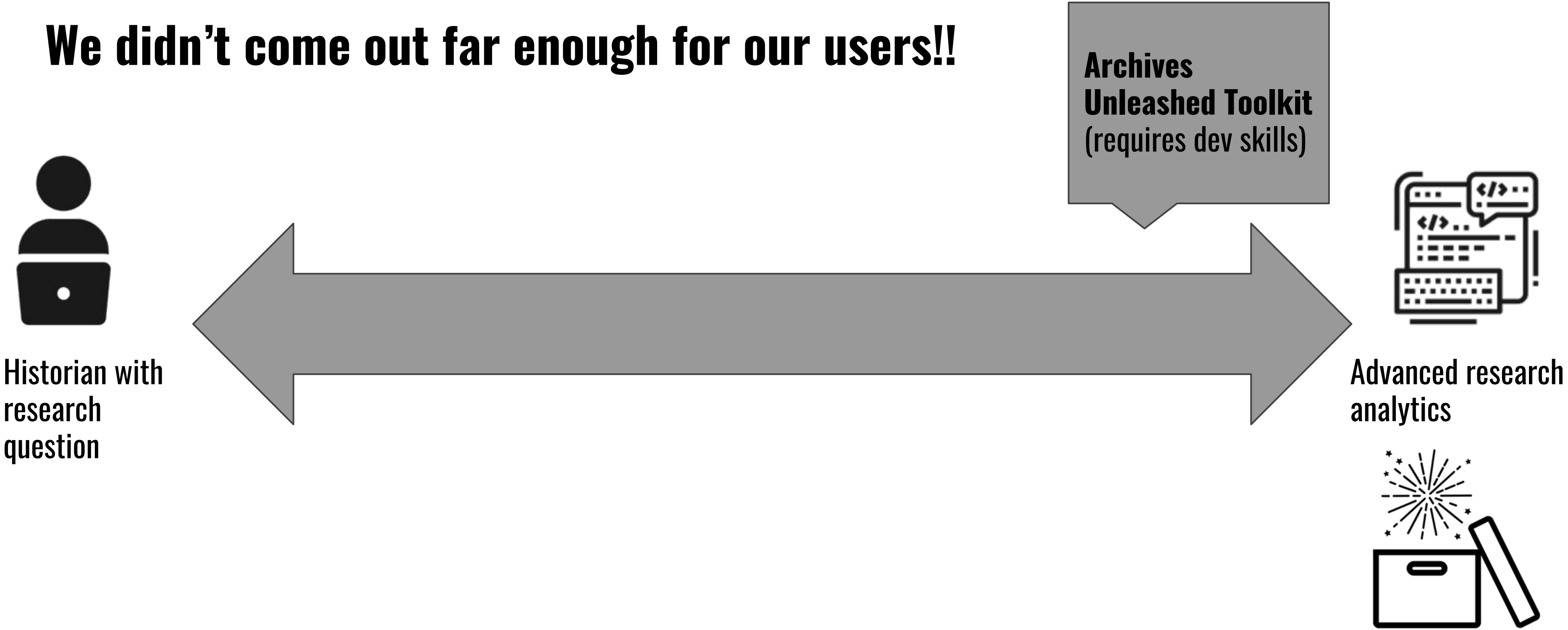
In other words...

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).

In other words...

We didn’t come out far enough for our users!!



Archives Unleashed Toolkit (requires dev skills)

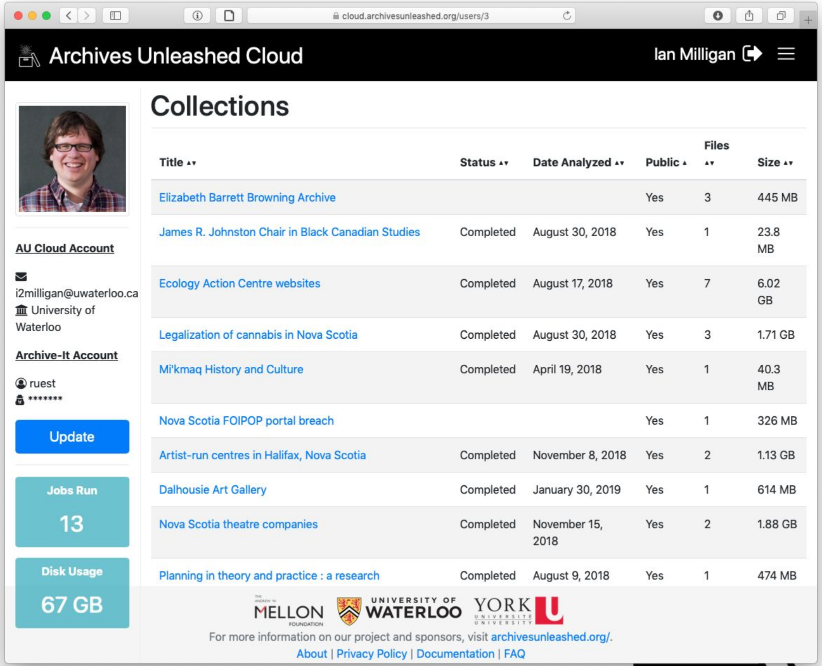
Historian with research question

Advanced research analytics

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



Collections

AU Cloud Account

✉ i2milligan@uwaterloo.ca
🏛 University of Waterloo

Archive-It Account

ruest
****

Jobs Run 13

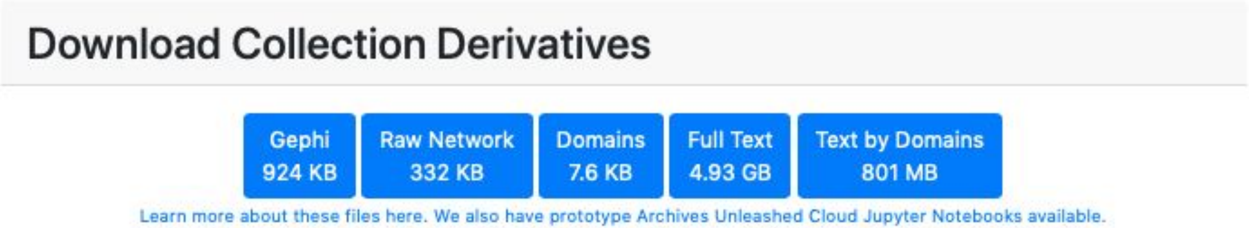
Disk Usage 67 GB

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
Mi'kmaq 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

For more information on our project and sponsors, visit archivesunleashed.org/.

About | Privacy Policy | Documentation | FAQ

What does a researcher get?



Download Collection Derivatives

Learn more about these files here. We also have prototype Archives Unleashed Cloud Jupyter Notebooks available.

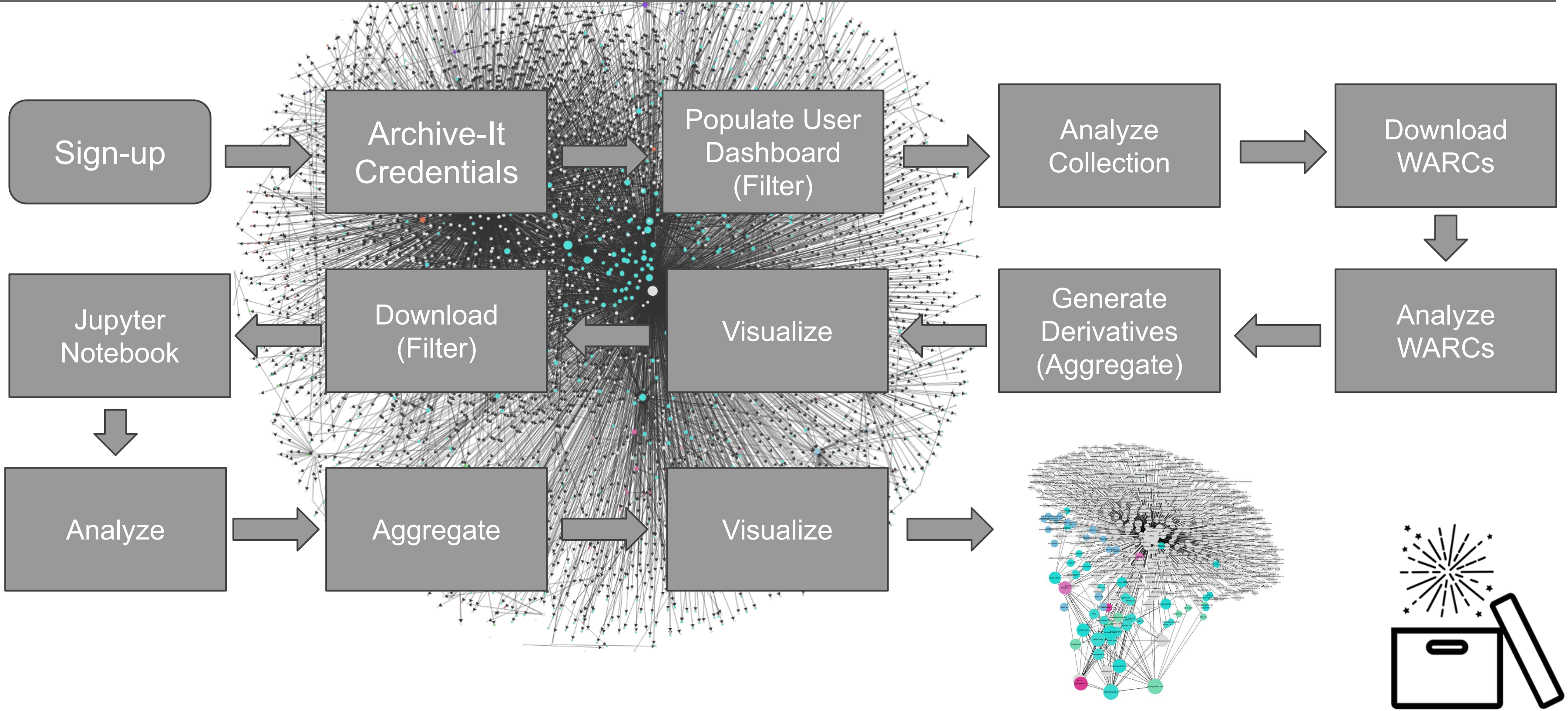
Gephi/Raw Network Files: Network diagram with characteristics pre-computed (Gephi); Raw network diagram (origin/destination/weight);

Domains: Statistical breakdown of what’s present in a collection

Full Text: The full text of the entire collection (i.e. HTML pages w/ tags stripped out for analysis), in CSV format with crawl date, domain, full URL, full text)

Text by Domains: The plain text of the ten most frequent domains.

How it works



Sign-up

Archive-It Credentials

Populate User Dashboard (Filter)

Analyze Collection

Download WARCs

Analyze WARCs

Generate Derivatives (Aggregate)

Visualize

Download (Filter)

Jupyter Notebook

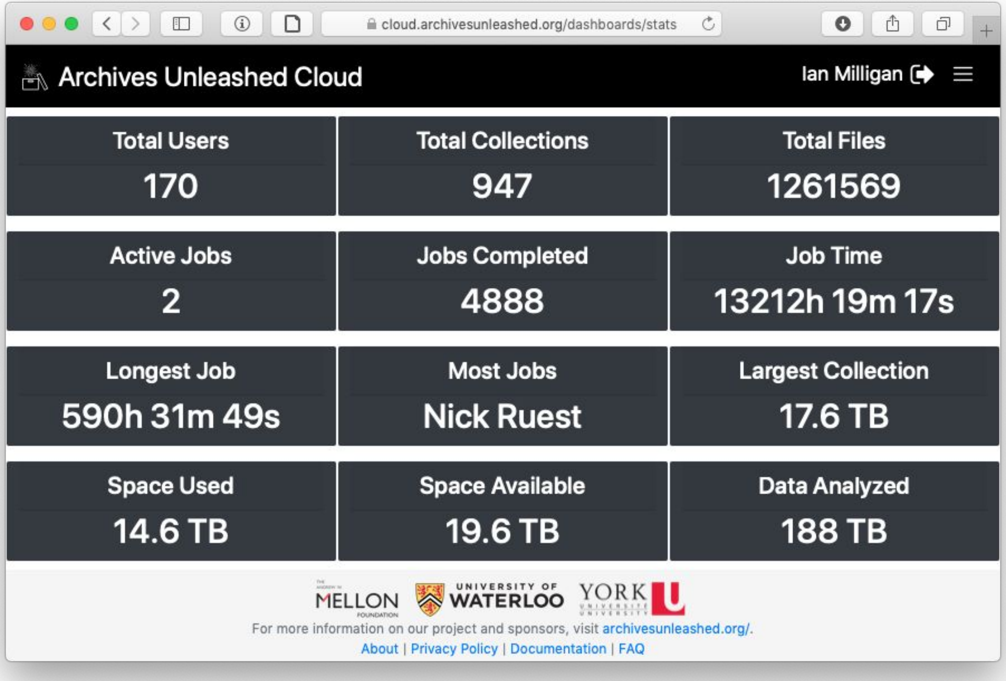
Analyze

Aggregate

Visualize

Archives Unleashed Cloud

Give me some stats!



Archives Unleashed Cloud

Ian Milligan

Total Users	Total Collections	Total Files
170	947	1261569
Active Jobs	Jobs Completed	Job Time
2	4888	13212h 19m 17s
Longest Job	Most Jobs	Largest Collection
590h 31m 49s	Nick Ruest	17.6 TB
Space Used	Space Available	Data Analyzed
14.6 TB	19.6 TB	188 TB

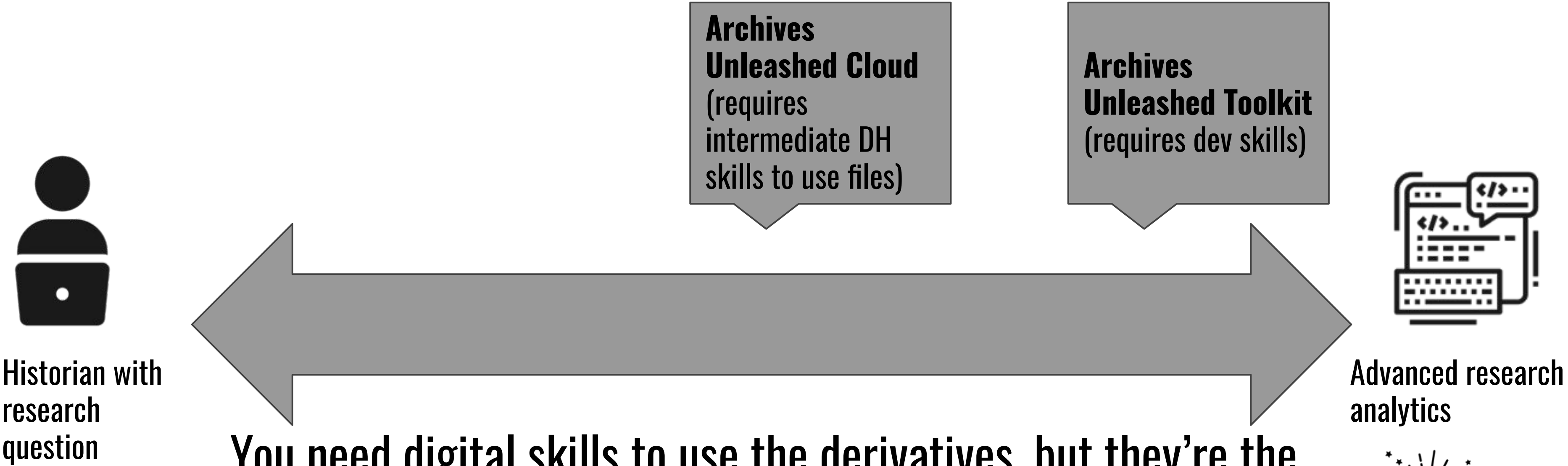
For more information on our project and sponsors, visit archivesunleashed.org/.

About | Privacy Policy | Documentation | FAQ

Stats as of Thursday, June 13th 2019

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

Right now, we're kind of here…



Archives Unleashed Cloud (requires intermediate DH skills to use files)

Archives Unleashed Toolkit (requires dev skills)

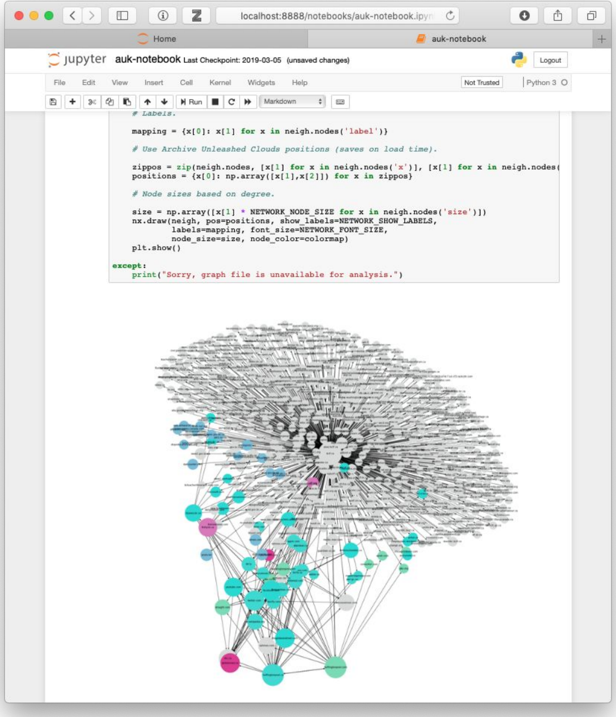
Historian with research question

Advanced research analytics

You need digital skills to use the derivatives, but they're the kind you can go to your library or main resources to use…

Archives Unleashed Cloud Notebooks

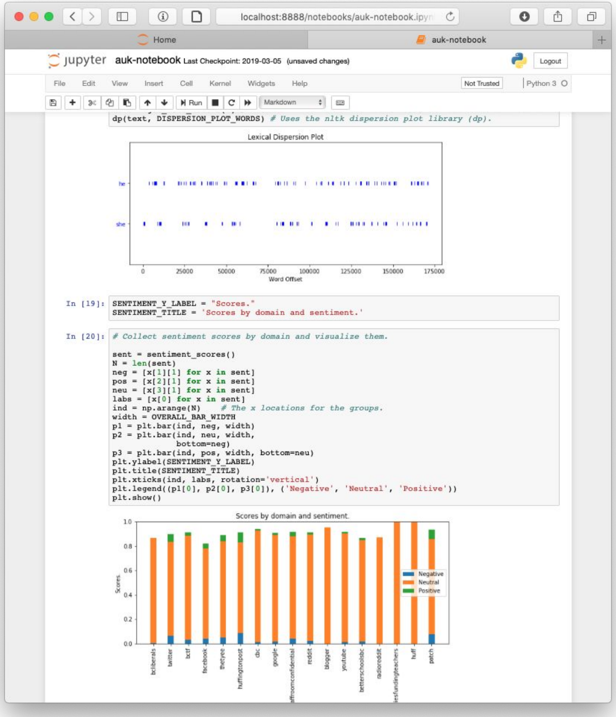
- Jupyter Notebooks
- One for each kind of derivative (domain, networks, text)
- A “**mad-libs**” approach - fill in the blanks with the variables (domains, dates, collections, etc.) that you are interested in, and it does basic computations for you
- Currently an under-development prototype service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


localhost:8888/notebooks/auk-notebook.ipyn
Home
auk-notebook
jupyter
auk-notebook Last Checkpoint: 2019-03-05 (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Markdown

```
    # Labels.

    mapping = {x[0]: x[1] for x in neigh.nodes('label')}

    # Use Archive Unleashed Clouds positions (saves on load time).

    zippos = zip(neigh.nodes, [x[1] for x in neigh.nodes('x')], [x[1] for x in neigh.nodes(
    positions = {x[0]: np.array([x[1],x[2]]) for x in zippos}

    # Node sizes based on degree.

    size = np.array([x[1] * NETWORK_NODE_SIZE for x in neigh.nodes('size')])
    nx.draw(neigh, pos=positions, show_labels=NETWORK_SHOW_LABELS,
            labels=mapping, font_size=NETWORK_FONT_SIZE,
            node_size=size, node_color=colormap)
    plt.show()

except:
    print("Sorry, graph file is unavailable for analysis.")
```

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.



```
dp(text, DISPERSION_PLOT_WORDS) # Uses the nltk dispersion plot library (dp).
```

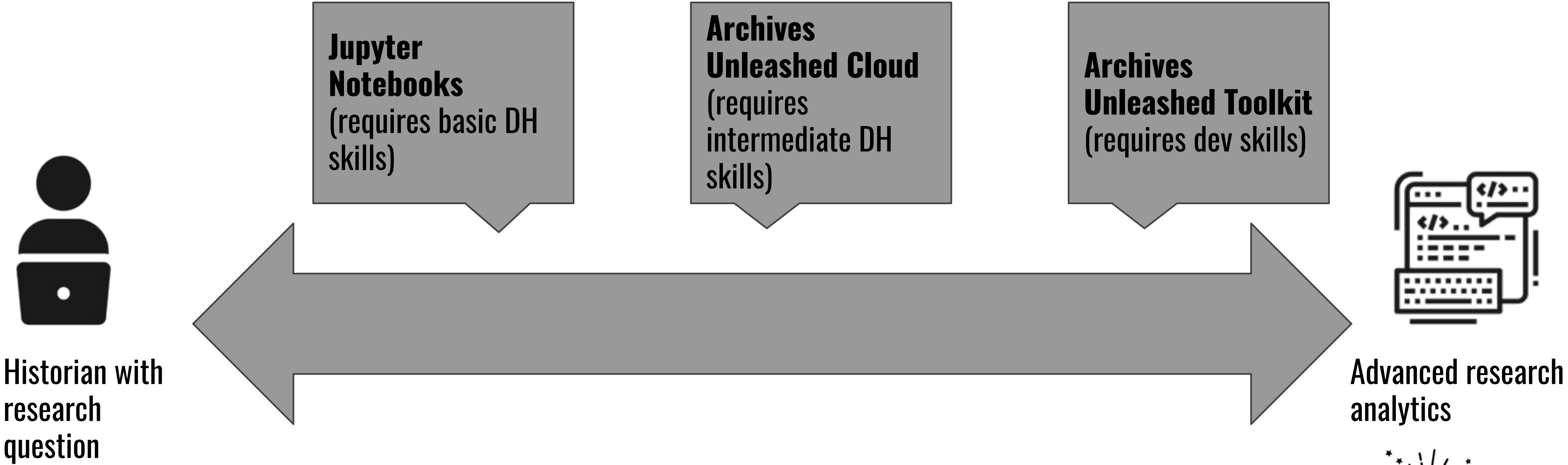
```
In [19]: SENTIMENT_Y_LABEL = "Scores."
         SENTIMENT_TITLE = 'Scores by domain and sentiment.'
```

```
In [20]: # Collect sentiment scores by domain and visualize them.

         sent = sentiment_scores()
         N = len(sent)
         neg = [x[1][1] for x in sent]
         pos = [x[2][1] for x in sent]
         neu = [x[3][1] for x in sent]
         labs = [x[0] for x in sent]
         ind = np.arange(N)    # The x locations for the groups.
         width = OVERALL_BAR_WIDTH
         p1 = plt.bar(ind, neg, width)
         p2 = plt.bar(ind, neu, width,
                      bottom=neg)
         p3 = plt.bar(ind, pos, width, bottom=neu)
         plt.ylabel(SENTIMENT_Y_LABEL)
         plt.title(SENTIMENT_TITLE)
         plt.xticks(ind, labs, rotation='vertical')
         plt.legend((p1[0], p2[0], p3[0]), ('Negative', 'Neutral', 'Positive'))
         plt.show()
```

Notebook Demo

binder

Maybe we’re helping out with the bigger spectrum?



Jupyter Notebooks (requires basic DH skills)

Archives Unleashed Cloud (requires intermediate DH skills)

Archives Unleashed Toolkit (requires dev skills)

Historian with research question

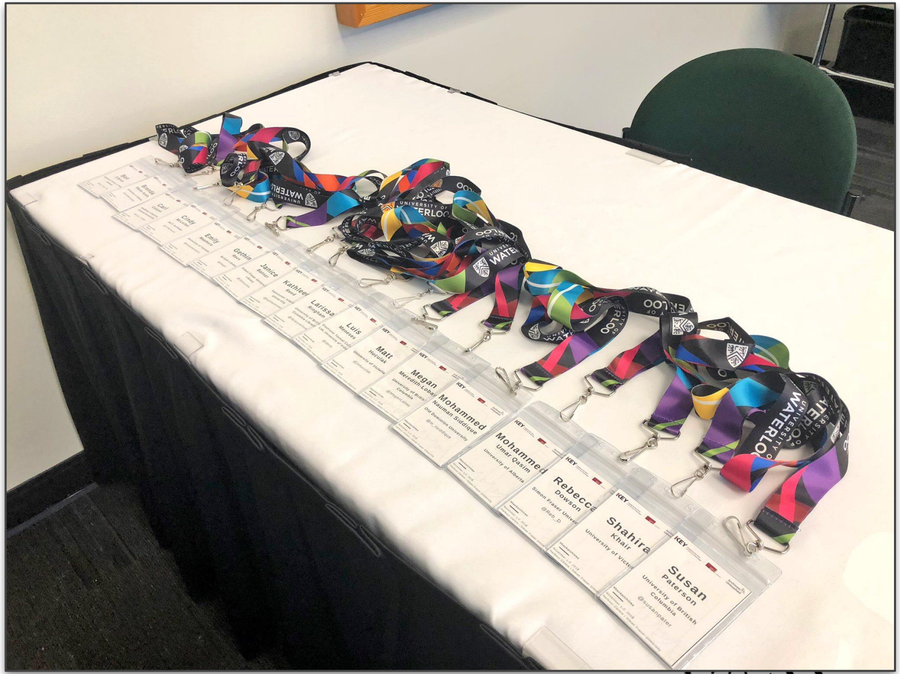
Advanced research analytics

Researchers STILL need to come out a bit, *but* we’re getting there…

Finally, we aim to build community around web archives.

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



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 arou web archiving practice.

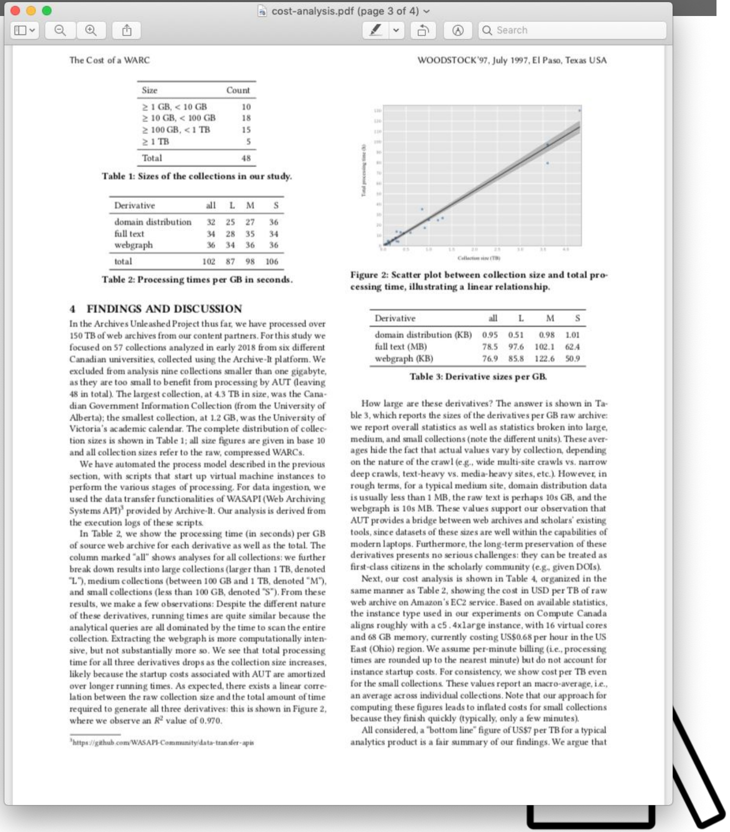
What’s Next?

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).
    - 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 WARCs 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

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

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.

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.

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 WARCs.

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)[3] 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.

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 DOIs).

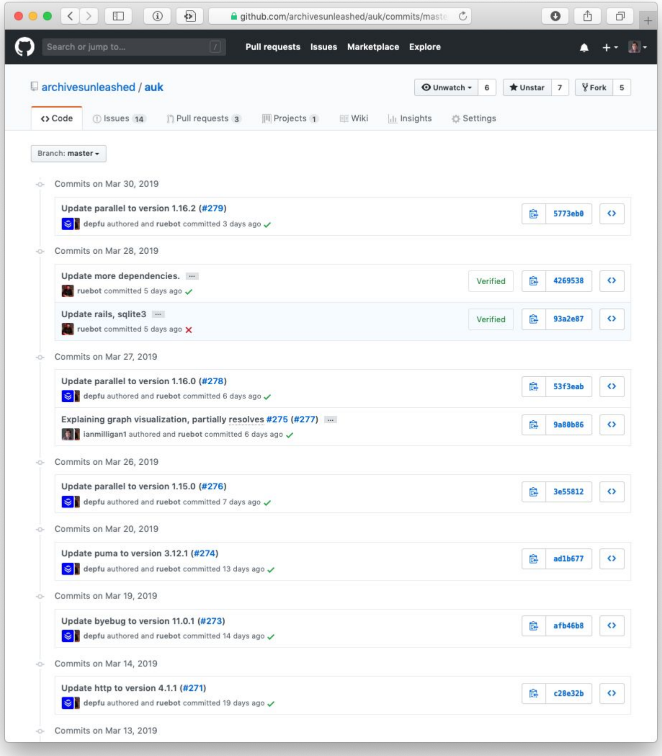
Next, our cost analysis is shown in Table 4, organized in the same manner as Table 2, showing the cost in USD per TB of raw web archive on Amazon’s EC2 service. Based on available statistics, the instance type used in our experiments on Compute Canada aligns roughly with a `c5.4xlarge` instance, with 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

[3] https://github.com/WASAPI-Community/data-transfer-apis

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



Commits on Mar 30, 2019

Update parallel to version 1.16.2 (#279)
depfu authored and ruebot committed 3 days ago ✓ — 5773eb0

Commits on Mar 28, 2019

Update more dependencies. …
ruebot committed 5 days ago ✓ — Verified — 4269538

Update rails, sqlite3 …
ruebot committed 5 days ago ✗ — Verified — 93a2e87

Commits on Mar 27, 2019

Update parallel to version 1.16.0 (#278)
depfu authored and ruebot committed 6 days ago ✓ — 53f3eab

Explaining graph visualization, partially resolves #275 (#277) …
ianmilligan1 authored and ruebot committed 6 days ago ✓ — 9a80b86

Commits on Mar 26, 2019

Update parallel to version 1.15.0 (#276)
depfu authored and ruebot committed 7 days ago ✓ — 3e55812

Commits on Mar 20, 2019

Update puma to version 3.12.1 (#274)
depfu authored and ruebot committed 13 days ago ✓ — ad1b677

Commits on Mar 19, 2019

Update byebug to version 11.0.1 (#273)
depfu authored and ruebot committed 14 days ago ✓ — afb46b8

Commits on Mar 14, 2019

Update http to version 4.1.1 (#271)
depfu authored and ruebot committed 19 days ago ✓ — c28e32b

Commits on Mar 13, 2019

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.

The TL;DR (aka Conclusions)

- Historians in the future will need to understand the Web
- We need to make sure they’re ready
  - Part of this is new, usable tools;
  - Part of this is new cultures in the humanities;
- But overall, we all need to begin to work together in a model of interdisciplinary collaboration, development, and partnership to make sure historians are equipped to do this sort of work.

Thanks to our supporters!

THE
ANDREW W.
MELLON
FOUNDATION

compute canada | calcul canada

START
SMART
LABS

YORK
UNIVERSITÉ
UNIVERSITY

UNIVERSITY OF
WATERLOO

Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada

We look forward to your **questions and thoughts.**

Links

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