Building Community and Tools for Analyzing Web Archives through Datathons

Ian Milligan, Nathalie Casemajor, Samantha Fritz, Jimmy Lin, Nick Ruest, and Nicholas Worby

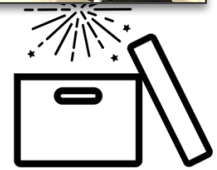
JCDL 2019 - Urbana-Champaign, IL

Why a series of web archive datathons?

Three primary objectives:

- Facilitate Scholarly Access
- Build Community
- Build Skills

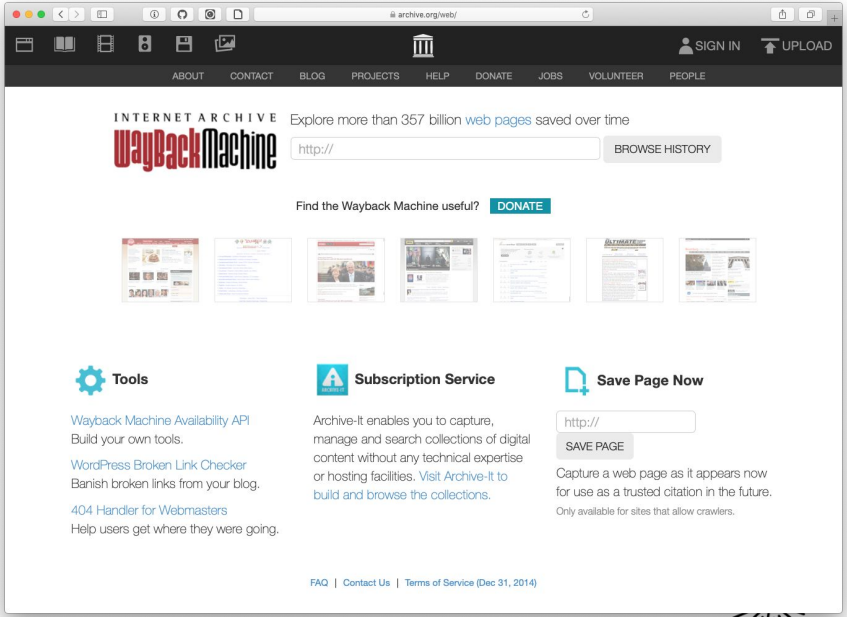
Develop the **social infrastructure** for a network of individuals and groups held together by the common goal of exploring web archives as research objects.



Why do we care about web archives?

By way of introduction:

- **Born-digital sources have the potential to reshape research** in the humanities and social sciences;
- **Research access has lagged** (beyond Wayback Machine, analysis ecosystem is mostly command-line-based tools)
- We need to expand the web archiving community and develop social infrastructure



What have we accomplished?

By way of introduction... our achievements

The tangible **development of tools and platforms** that meet demonstrated needs (i.e., better support for scholarly inquiry);

A better **understanding of the processes** by which scholars, curators, and others work with these materials, providing a reference workflow with which to evaluate future research tools;

The **building of a community**, in part supported by the continued use of datathon communication channels and standing infrastructure, as well as encouragement to attend follow-up events.

We've now run **SEVEN** datathons! (exhausting but fun)

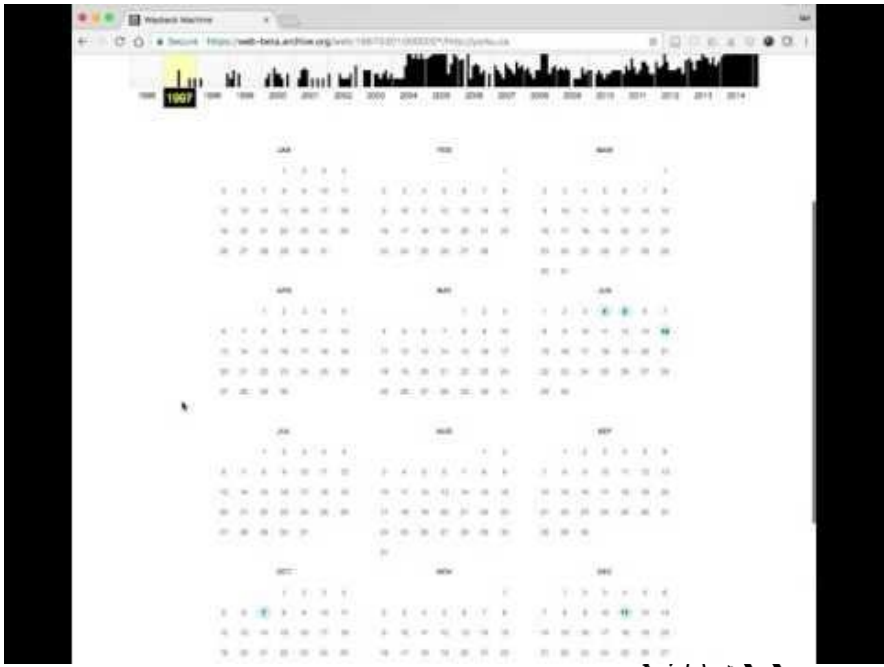
So why datathons?

The Wayback is Not Enough

The Wayback is Not Enough...

The **Wayback Machine** is incredible if you know what you're looking for...

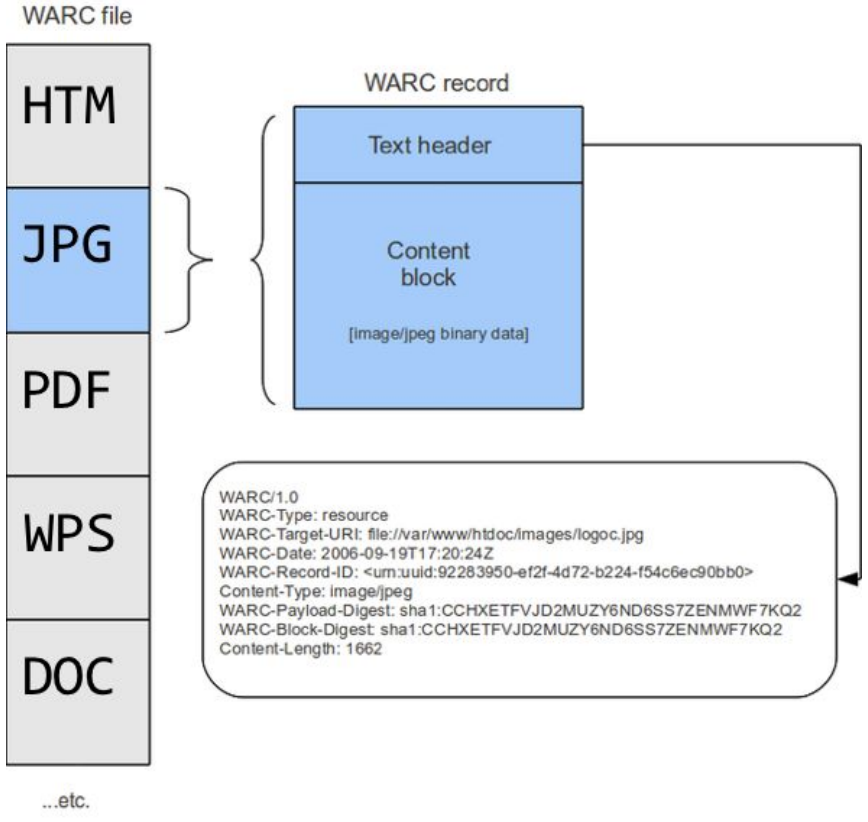
... but for more sophisticated queries (i.e. websites that say X and link to Y; finding topics; exploratory text mining; images en masse), **you need to work with the WARC files directly**



… but WARC files are hard to use!

Offer a lot of **potential**

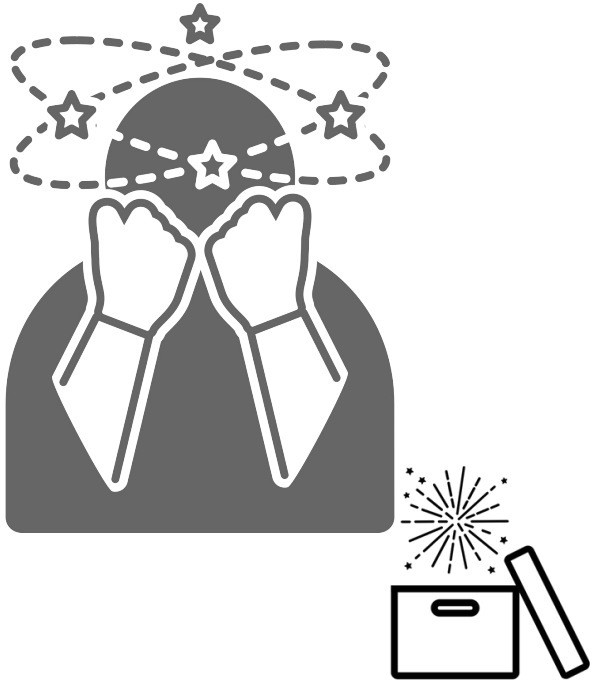
- Text analysis at scale (finding keywords, patterns, etc. over time);
- Network analysis at scale (using hyperlinks to see how links have evolved over time; PageRank, etc.)
- Supporting movement between **distant** and **close** reading



… really, really hard to use!

No, seriously, they're hard to use.

- 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 Datathons

Archives Unleashed

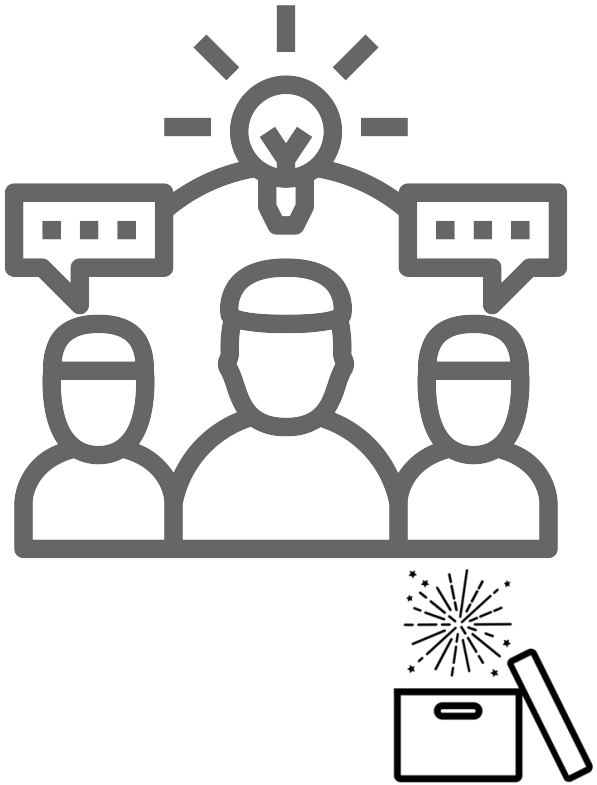
The Datathon Model

Our datathons aimed to bring together **scholars, curators, subject matter experts, developers,** and **others** interested in collaborating on a shared web archiving project.

In a dream world:

- ⅓ technical individuals (CS, developers)
- ⅓ subject matter experts (historians, political science, etc.)
- ⅓ collectors (librarians, archivists)

In general, expansive in our definitions.



Time	**Activity**	
	Day 1	*Day 2*
09:00	Breakfast and Welcome	
09:30	Introductory Remarks	Group Work
10:00	Introduction to Tools	Group Work
10:30	Coffee Break	
10:45	Sticky Notes Exercise	Group Work
11:15	Group Work	Group Work
12:30	Lunch and Lightning Talks	
13:15	Group Work	
15:00	Coffee Break	
15:30	Group Work	Awards and Closing
17:30	Evening Social	

Sample datathon schedule

The critical elements include...

Team building exercise

The most critical element of any datathon is **team formation**

- We use a “sticky notes exercise” (a technique adopted from participatory design)
- Three colours of sticky notes:
  - One colour: Research **Questions**
  - One colour: Research **Methods** or **Tools**
  - One colour: Datasets
- Stick to wall and do the rounds...

Team building exercise

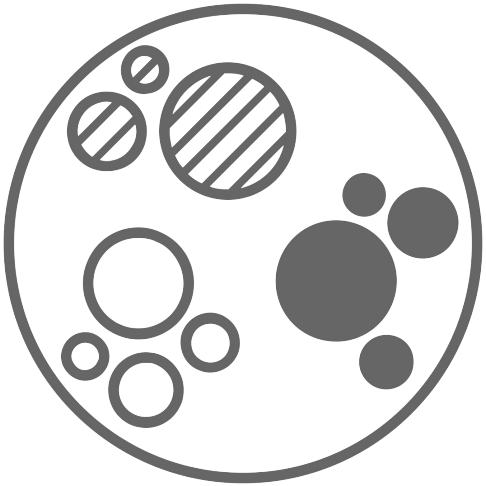
Then we **cluster** the notes together to find common themes that cut across different categories

- I.e. a cluster of people interested in discourse; hyperlink; etc.

We physically move people around the room to try out different ideas, and pay attention to two main rules:

- Teams should be smaller than six people;
- Teams should not contain individuals from the same institution.

Takes around 30 minutes.



Computing resources

We also learned (over time) that **computing resources** are critical

Compute Canada VMs, enough relatively strong servers (8/16 cores, 30-60GB RAM) so that each team can have one

Each system also has:

- The WARCs for each dataset;
- The main derivatives for each dataset (text, networks, domain stats); and
- Software packages like Archives Unleashed, libraries, etc. already loaded

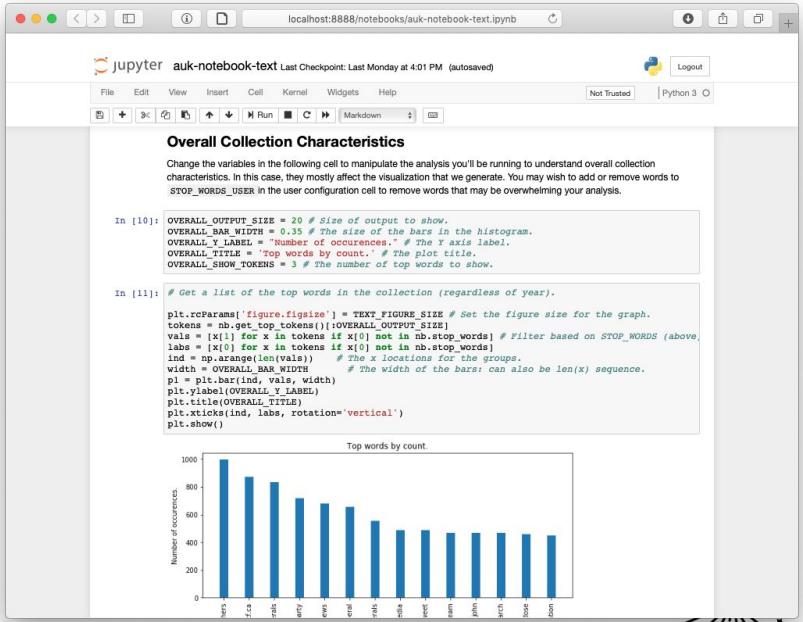
compute canada | calcul canada

Computing resources

And, our recent invention, **Archives Unleashed Jupyter Notebooks** which allow users to take those derivative files and run analyses on them using basic data science packages

The notebooks were served directly from the Compute Canada servers!

(Learn more at the poster session..)



We have now run seven events

Events to date...

Phase One

- **University of Toronto** (March 2016)
- **Library of Congress** (June 2016)
- **Internet Archive** (February 2017)
- **British Library** (June 2017)

Details are in the paper, but each saw major developments.

Events to date...

Phase Two

- **University of Toronto** (April 2018)
- **Simon Fraser University** (November 2018)
- **George Washington University** (March 2019)
- **TBA**

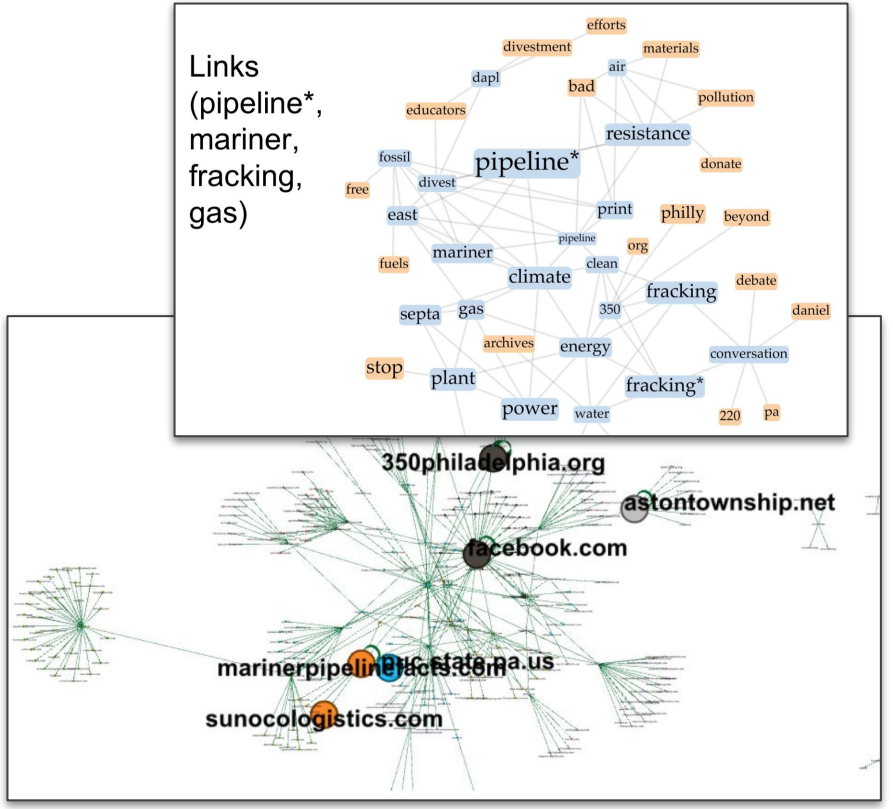
This was now a subset of organizers around a Mellon-funded project. More emphasis on particular “Archives Unleashed Tools.”

Check out some cool projects at https://archivesunleashed.org/events/

Some example projects

Team: Pipeline (Victoria, Dalhousie, Penn State, Simon Fraser University)

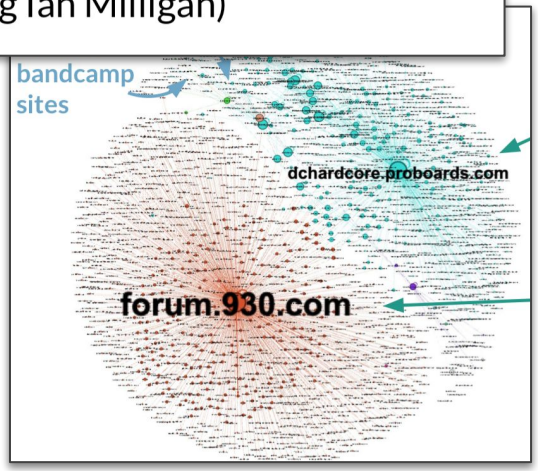
- **Worked with** Penn State WARCs; U of T WARCs; old Twitter datasets from the DocNow Catalog, and newly-generated Twitter datasets; all around pipeline activism
- **Analyzing link graphs; keywords; images; etc**.



Some example projects

Team: Punkavists (Library of Congress, Harvard, Dalhousie)

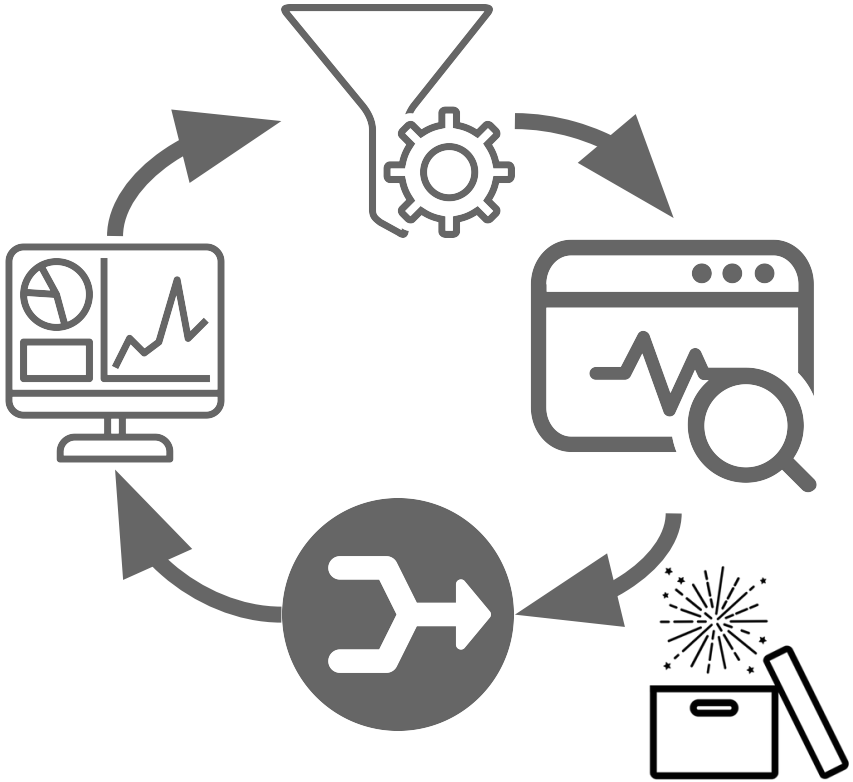
- Looking at subgenres of punk represented in collections; details around the DC punk scene;
- Looking at full text of punk collections; playing with particular domains; mapping out collections.



Again - I feel guilty that I can’t show them all off - https://archivesunleashed.org/events/

Lessons learned

Developing a research cycle for working with web archives: the **F**ilter - **A**nalyze - **A**ggregate - **V**isualize (FAAV) Cycle.



Major Lessons Learned

“More Hack, Less Yack”

- Cliche, but we do need to have room to work. Initial datathons had more formal presentations; we’ve now minimized that so teams can get to work before lunch on Day One.

“Right Mix of Participants”

- We arrived at the ⅓ technical; ⅓ subject matter; ⅓ collector/curator mix through hard experience. The right balance is critical. You also need people open to learning difficult technical things that might not perfectly work.

Major Lessons Learned

Staging Datasets in the Cloud

- We tried to run out first datathon largely on physical hard drives and USB keys. Big mistake! Too much time waiting for stuff to copy... not enough time to hack.

Pre-Processing Datasets to Provide Derivatives

- Similar lesson learned... too much time waiting to do the same three exploratory derivative generations that you'd do almost no matter what. Too much time waiting for stuff to process... not enough time to hack!

These are generalizable lessons across the digital libraries community.

In conclusion...

The more we get together, the happier* we'll be.

* By happier, we mean meeting the challenges that these new cultural datasets present...

Thanks to our supporters!

THE ANDREW W. MELLON FOUNDATION

START SMART LABS

YORK UNIVERSITÉ UNIVERSITY

compute canada | calcul canada

UNIVERSITY OF WATERLOO

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada