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?

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
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 WARC files (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)

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

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

Technical skillset of most historians

Command Line-based Tools to use WARC files

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

1. Archives Unleashed Toolkit
2. Archives Unleashed Cloud
3. 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
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”
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
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!!

Historian with research question

Archives Unleashed Toolkit (requires dev skills)

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
### What does a researcher get?

<table>
<thead>
<tr>
<th>Download Collection Derivatives</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gephi</strong> 924 KB</td>
</tr>
<tr>
<td><strong>Raw Network</strong> 332 KB</td>
</tr>
<tr>
<td><strong>Domains</strong> 7.6 KB</td>
</tr>
<tr>
<td><strong>Full Text</strong> 4.93 GB</td>
</tr>
<tr>
<td><strong>Text by Domains</strong> 801 MB</td>
</tr>
</tbody>
</table>

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

1. Sign-up
2. Jupyter Notebook
3. Archive-It Credentials
4. Populate User Dashboard (Filter)
5. Download (Filter)
6. Visualize
7. Aggregate
8. Visualize
9. Analyze Collection
10. Generate Derivatives (Aggregate)
11. Download WARC
12. Analyze WARC
Archives Unleashed Cloud
Stats as of Thursday, June 13th 2019

<table>
<thead>
<tr>
<th>Total Users</th>
<th>Total Collections</th>
<th>Total Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>170</td>
<td>947</td>
<td>1261569</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Active Jobs</th>
<th>Jobs Completed</th>
<th>Job Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4888</td>
<td>13212h 19m 17s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Longest Job</th>
<th>Most Jobs</th>
<th>Largest Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>590h 31m 49s</td>
<td>Nick Ruest</td>
<td>17.6 TB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Space Used</th>
<th>Space Available</th>
<th>Data Analyzed</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.6 TB</td>
<td>19.6 TB</td>
<td>188 TB</td>
</tr>
</tbody>
</table>
But where does our platform end... And the researcher begin?
Right now, we’re kind of here...

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 (requires intermediate DH skills to use files)

Archives Unleashed Toolkit (requires dev skills)

Historian with research question

Advanced research analytics
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
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.
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

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 around web archiving practice.
What’s Next?
What’s Next?

- **Sustainability** has been baked into our grant from the very start (thanks Mellon!).
  - Costs USD$7/TB to process using the Archives Unleashed Toolkit.
Actual processing costs are relatively affordable – approx. US$7/TB to process WARC 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...
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.
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!
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