Opening up WARCs:
The Archives Unleashed Cloud and Toolkit Projects

Ian Milligan
Associate Professor of History
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
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...
Why should we care about web archives?

▪ More data than ever before is being preserved;
▪ And it’s being saved and delivered to us in very different ways....
YOU CAN’T STUDY THE 1990S WITHOUT THE WEB.

AND HISTORIANS AREN’T READY.
Putting those 22 years since 1996 into context...

- 1996 was 22 years ago
  - That’s like writing about the Second World War in 1967...
  - Or 1968 in 1990.

- Doesn’t seem so outlandish does it?
WHY AREN’T WE READY?

Access has lagged...
Current State of Access for Most Users who Don’t Hang out at IIPC
Current State of Access for Most Users

- **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 WHAT IF YOU WANT TO WORK WITH THE WARCS THEMSELVES?
If ever an audience has never needed an introduction to a WARC file...
Working with Web Archives

Current state of access is command line, i.e. Altiscale + Pig

```
Archive = LOAD "$I_PARSED_DATA" USING SequenceFileLoader()
  AS (key:chararray, value:chararray);
Archive = FOREACH Archive GENERATE FROMJSON(value) AS m:[];
Archive = FILTER Archive BY m#`errorMessage` is null;
ExtractedCounts = FOREACH Archive GENERATE m#`url` AS
  src:chararray,
  SURTURL(m#`url`) AS surt:chararray,
  REPLACE(m#`digest`,`sha1:`,``) AS checksum:chararray,
  SUBSTRING(m#`date`, 0, 8) AS date:chararray,
  REPLACE(m#`code`, `[^p{Graph}]`, ` `) AS code:chararray,
  REPLACE(m#`title`, `[^p{Graph}]`, ` `) AS title:chararray
  REPLACE(m#`description`, `[^p{Graph}]`, ` `)AS
description:chararray,
  REPLACE(m#`content`, `[^p{Graph}]`, ` `) AS
content:chararray;
UniqueCaptures = FILTER ExtractedCounts BY content MATCHES`.
  *naturals+disaster.*` OR content MATCHES`
  *naturals+disaster.*` OR content MATCHES`.*desertification.*` OR content MATCHES`
  *climates+change.*` OR content MATCHES`.*pollution.*` OR
content MATCHES`.*foods+security.*`;
STORE UniqueCaptures INTO `$O_DATA_DIR` USING PigStorage('\u0001');```
import io.archivesunleashed._
import io.archivesunleashed.matchbox._
RecordLoader.loadArchives("example.arc.gz", sc)
  .keepValidPages()
  .keepDate(List("200804"),
    ExtractDate.DateComponent.YYYYMM)
  .map(r => (r.getCrawlDate,
    r.getDomain, r.getUrl,
    RemoveHTML(r.getContentString)))
  .saveAsTextFile("plain-text-date-filtered-200804/")
WHAT’S WRONG WITH THIS KIND OF APPROACH?
Our Experience #1: If you build it, they won’t come...
YOU NEED TO BUILD COMMUNITY.
Our Experience #2: Humanists and social scientists hate command lines!
(As we’ll have fun seeing in the workshop shortly)
YOU NEED TO BUILD USABLE TOOLS.
SO WHAT’S NEXT FOR INCREASING ANALYTICAL USE OF WEB ARCHIVES?
Enter our vision...

Archives Unleashed
OUR GOAL

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

Ian Milligan
PI (Historian)

Nick Ruest
Co-PI (Librarian)

Jimmy Lin
Co-PI (Computer Science)

Plus staff, students, and postdoctoral fellows from history, computer science, and with MLIS backgrounds.
SO WHAT DO WE DO?

We do three things.
The Archives Unleashed Toolkit

The Archives Unleashed Cloud

The Archives Unleashed Datathons
The Archives Unleashed Toolkit
Wayback Machine

It’s great, but it’s predicated on really knowing what you’re looking for....
Great if you know what you’re looking for...
Working with Web Archives

... but instead we work with the data that powers the Wayback Machine.
Working with Web Archives

THE GOOD STUFF!

And let people extract the info they need from web archives.
SO INSTEAD OF GOING PAGE BY PAGE, SCHOLARS CAN TAKE A “BIG DATA” APPROACH TO 20+ YEARS OF HISTORY
Working with Cultural Heritage at Scale
Working with Cultural Heritage at Scale
Working with Cultural Heritage at Scale

Filter
Analyze
Aggregate
Visualize
Working with Cultural Heritage at Scale
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 “Stephen Harper” or “Donald Trump”)

Visualize

- Output data as a visualization
  - Tables of results
  - External applications (GEXF files for Gephi)
Working with Cultural Heritage at Scale
So why doesn’t everybody use our toolkit/FAAV cycle?
THIS IS SUPER COOL, RIGHT...

Right???
IF THIS LOOKS COOL, YOU’LL BE PLAYING WITH THIS IN THE WORKSHOP SHORTLY.
THE CATCH, OF COURSE, IS WHAT I SPOKE ABOUT EARLIER: IT’S COMMAND LINE BASED....
Our cutting edge interface..
(and remember, people don’t like command lines!)
The Archives Unleashed
Cloud

Provides access to...

The Archives Unleashed
Toolkit

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
The Archives Unleashed Cloud

<table>
<thead>
<tr>
<th>Title</th>
<th>Analyzed</th>
<th>Public</th>
<th>Files</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Summitry Archive</td>
<td>No</td>
<td>Yes</td>
<td>549</td>
<td>425 GB</td>
</tr>
<tr>
<td>University of Toronto Libraries Digital Collections</td>
<td>No</td>
<td>Yes</td>
<td>110</td>
<td>63.6 GB</td>
</tr>
<tr>
<td>Federal Election Candidate Sites 2015</td>
<td>No</td>
<td>Yes</td>
<td>310</td>
<td>206 GB</td>
</tr>
<tr>
<td>Snowden Archive</td>
<td>Yes</td>
<td>Yes</td>
<td>42</td>
<td>7.16 GB</td>
</tr>
<tr>
<td>Toronto 2015 Pan Am &amp; Parapan American Games</td>
<td>Yes</td>
<td>Yes</td>
<td>294</td>
<td>50.4 GB</td>
</tr>
<tr>
<td>Hong Kong Politics</td>
<td>No</td>
<td>Yes</td>
<td>1005</td>
<td>1020 GB</td>
</tr>
<tr>
<td>Toronto Mayoral Election 2014</td>
<td>No</td>
<td>Yes</td>
<td>292</td>
<td>292 GB</td>
</tr>
<tr>
<td>Canadian Government Information</td>
<td>No</td>
<td>Yes</td>
<td>10644</td>
<td>4.42 TB</td>
</tr>
<tr>
<td>Aboriginal Canada Portal</td>
<td>Yes</td>
<td>Yes</td>
<td>10</td>
<td>426 GB</td>
</tr>
<tr>
<td>Canadian Labour Unions</td>
<td>No</td>
<td>Yes</td>
<td>6008</td>
<td>984 MB</td>
</tr>
<tr>
<td>University of Toronto Archives Web Collection</td>
<td>No</td>
<td>Yes</td>
<td>10376</td>
<td>1.3 TB</td>
</tr>
<tr>
<td>Canadian Political Interest Groups</td>
<td>Yes</td>
<td>Yes</td>
<td>100</td>
<td>8.75 GB</td>
</tr>
<tr>
<td>Canadian Political Parties and Political Interest Groups</td>
<td>No</td>
<td>Yes</td>
<td>4047</td>
<td>645 GB</td>
</tr>
</tbody>
</table>

AU Cloud Account
- ian.milligan1@gmail.com
- University of Toronto

Archive-It Account
- ResearcherDL
- ********

Jobs Run: 11
Disk Usage: 2.34 TB
The Archives Unleashed Cloud

Nova Scotia Municipal Governments

Hyperlink Diagram

Download Collection Derivatives

AU Cloud Account
i2milligan@uwaterloo.ca
University of Waterloo
Archive-It Account
ruest
********

Update

Jobs Run
6

Disk Usage
46.9 GB

For more information on our project and sponsors, visit archivesunleashed.org.
The Archives Unleashed Cloud

- Generates a common set of derivatives:
  - Domain frequency;
  - Network diagram (pre-generated layout for networks under 50,00 nodes and 200,000 edges);
  - Raw network diagram (just source -> target -> number of links);
  - Full text of entire collection;
  - Full text of the top ten most frequent domains.
- Interactive network in the browser.
- Click on button, downloads it.
The Archives Unleashed Cloud

- Still in development
  - Focus on stability
  - Other WASAPI partners?
- Should we add more derivatives?
- DataFrame support?
- Where does the platform end and the researcher (digital humanist?) begin?
What We’ve Done So Far (Dashboard)

![Dashboard Screenshot]

- **Users**: 79
- **Collections**: 401
- **Files**: 727,731
- **Active Jobs**: 34
- **Jobs Completed**: 2,093

**Job Time**
- 5,683 hours, 43 minutes, and 39 seconds
- 397 hours, 23 minutes, and 37 seconds

**Longest Job**
- Nick Ruest

**Most Jobs**
- Nick Ruest

**Largest Collection**
- 17.6 TB

**Space Used**
- 14.8 TB

**Space Available**
- 18.4 TB

**Data Analyzed**
- 118 TB

**User** | **Institution** | **Collection** | **Queue** | **Start Time** | **End Time** | **Total Run Time** | **Download**
--- | --- | --- | --- | --- | --- | --- | ---
Nick Ruest | Simon Fraser University | SFU Affiliated Conferences/Institutes | download | 2018-11-06 20:17:09 UTC | | | 20.7 GB
Nick Ruest | Simon Fraser University | Public Knowledge Project | download | 2018-11-06 16:48:32 UTC | | | 270 GB
Nick Ruest | Simon Fraser University | Government of Canada | download | 2018-11-06 16:27:12 UTC | | | 474 GB
Nick Ruest | Simon Fraser University | British Columbia Local Governments | download | 2018-11-06 15:56:58 UTC | 2018-11-06 20:17:09 UTC | 4 Hours, 20 Minutes and 11 Seconds | 121 GB
Nick Ruest | Simon Fraser University | SFU Library website | download | 2018-11-06 15:56:14 UTC | 2018-11-06 16:22:18 UTC | 26 Minutes and 3 Seconds | 957 MB
Nick Ruest | Dalhousie University | Mi’kmaw History and Culture | download | 2018-11-06 15:55:43 UTC | 2018-11-06 15:56:14 UTC | 31 Seconds | 40.3 MB
Nick Ruest | Dalhousie University | Legal History Miscellany | download | 2018-11-06 15:53:06 UTC | 2018-11-06 15:56:43 UTC | 2 Minutes and 36 Seconds | 373 MB
Nick Ruest | Dalhousie University | Jottings | download | 2018-11-06 15:53:46 UTC | 2018-11-06 15:53:46 UTC | 2 Minutes and 8 Seconds | 158 MB
IF YOU BUILD IT, THEY STILL WON’T COME.

Even with shiny GUIs, there’s one more step ...
The Archives Unleashed Datathons

• Need to build community, articulated shared visions of tool development and analysis
• Avoid black boxes of search engines and tools that scholars don’t understand!
• To date:
  • Toronto (x2, with University of Toronto)
  • Washington, DC (with Library of Congress)
  • San Francisco, CA (with Internet Archive)
  • Vancouver, BC (with SFU)
  • London, UK (with IIPC and BL)
The Archives Unleashed Datathons
The Archives Unleashed Datathons

- Helping to lower barriers
- Bringing people together
- Establishing a true community of practice to work with web archives
TO CONCLUDE
HISTORIANS IN THE FUTURE WILL NEED TO UNDERSTAND THE WEB.

We want to make sure we’re ready.
... AND YOU SHOULD BE READY TOO..

To Conclude

- Part of this will be new tools....
- .... part of this will be new cultures in the humanities and social sciences
  - Less geographical focus in history and more focus on methods, for example;
- But a recognition that you can’t study the 1990s without web archives.. And we’ll be studying the 1990s soon.
THANKS VERY MUCH.

You can now pre-order my book!