From Archiving to Analysis:
Current Trends and Future Developments in Web Archive Use

Ian Milligan
Associate Professor
Department of History
Web Archives are Cool, the TL;DR Version

- 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.
WHAT DOES A RESEARCHER NEED TO WORK WITH DATA AT SCALE?
SKILLSET ONE: WORKING WITH DATA AT SCALE

- An understanding of:
  - The Basics of **Natural Language Processing (NLP)**
  - **Basic Statistical Knowledge** to work with Quantitative Data (frequency of websites, terms, normalizing numbers, etc.)
  - Flexible **Data Science** skills (or StackOverflow skills..)
- In other words, being equipped with the skills and capacities to analyze text/data at scale.
SKILLSET TWO: UNDERSTANDING HOW DATA IS CONSTRUCTED

- They also need to have a solid understanding of:
  - How and why the data was collected, i.e. selection criteria;
  - What data wasn’t collected;
  - How the software used to create the dataset has changed over time;
- How to clean or normalize data when necessary (i.e. as a URL changes from http://www.ndp.ca to https://www.ndp.ca to http://ndp.ca in the crawl, recognizing that those are probably the same website! You see that, the computer doesn’t unless you teach it to).
OH YEAH...

PLUS THE NORMAL SKILLS OF THE HISTORIAN TOO..
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
ARE HISTORIANS READY?

(Leading question alert!)
NO...
(Sad trombone)
One example of command-line access is the Archives Unleashed Toolkit, based on working with scholars at seven datathons + numerous workshops.

- It allows scholar to use the FAAV cycle to explore it:
  - Filter
  - Analyze
  - Aggregate
  - Visualize

- Allows for complicated research queries using Apache Spark

- (Other examples: ArchiveSpark, Python libraries, etc.)
PROBLEM SOLVED, RIGHT?

Check out our cutting-edge interface....
THE ARCHIVES UNLEASHED TOOLKIT

- It’s easy to use, as long as you:
  - Know how to use the command line;
  - How to access a server;
  - How to use the Spark Shell;
  - How to code, at least somewhat, in Scala;
  - And, have a lot of patience for open-source documentation!

(OK, it’s not easy to use..)

```scala
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/")
```
... AND HISTORIANS HAVE BEEN TURNING AWAY FROM QUANT TO QUAL

- 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)
WHERE TOOLS END AND USERS BEGIN

Historian with Research Question

Advanced Web Archive Analysis
WHERE TOOLS END AND USERS BEGIN

**OPTION ONE:** Historians become programmers, capable of advanced web archive analysis.
- Writing code;
- Contributing to open-source projects;
- Maybe need a bit of support but in general are self-sufficient in a computational environment.
WHERE TOOLS END AND USERS BEGIN

OPTION TWO: Historians do not develop technical skills, continuing solely as subject-matter experts.
- An approach that sees using a web archive as similar to using PROQUEST or JSTOR – a bit of work but nothing out of the ordinary.
OPTION THREE: The middle ground
- Some computational skills, but platforms designed to resemble conventional research processes as much as possible
- Positioned in a world where you can draw on standard library research support (i.e. not web archiving particularly, but “working with text”)
- No command lines!
WHERE TOOLS END AND USERS BEGIN

- We can draw on a model of successful interdisciplinary cooperation to see how this can be effected.

- If a computer scientist and a historian work together, it isn’t just “historian: now you become a computer scientist;” nor is it “computer scientist: now you become a historian.” Compromise is worked out in terms of:
  - Workflow (Platforms? Collaboration?)
  - Publication venues
  - The shape of the work

- It really is often meeting half way, and being conscious.
THREE PERSONAS

Person A
Computational Humanist

Person B
Digital Humanist

Person C
Conventional Historian
IN OTHER WORDS...

Historian with Research Question

Person C

Person B

Person A

Advanced Web Archive Analysis
HOW CAN WE SUPPORT THIS COMPUTATIONAL TURN FOR ALL THREE PERSONAS?
Enter our project...
FIRST, WE NEED AN INTERDISCIPLINARY TEAM

Ian Milligan
Historian

Nick Ruest
Library/Archives

Jimmy Lin
Computer Science
WE CAN’T GO IT ALONE...!
FROM ARCHIVING TO ANALYSIS
(So why would we expect our users to go it alone??)
Archives Unleashed aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past.
AND A SET OF TOOLS...

Cloud

Toolkit

Notebooks

Datathons
The sort of scholar we would have called a “computational humanist” or “computational social scientist.” Comfortable installing packages, understanding dependencies, fluent on the command line, and can Stack Overflow like a master.
COMPUTATIONAL HUMANITIES

- Needs to have some sort of toolkit to translate WARC files into something they can work with
- Can interpret tested yet still dense documentation
- Can troubleshoot error messages (i.e. knows what to Google – this is a hard-earned skill)
- Can take sample scripts, execute them to see outcome, and then change them to run on their own data with own questions
- Can use the command line!

```java
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/"
)```

FROM ARCHIVING TO ANALYSIS
COMPUTATIONAL HUMANITIES

- The **Archives Unleashed Toolkit** was designed around this persona.

- Allows a user to take WARCs and:
  - Determine elemental statistics about a collection;
  - Extract particular images, domains, URLs, pages with keywords, etc.
  - Do sophisticated Apache Spark-powered network analysis; and
  - Write custom Scala scripts to do almost anything you imagine with our set of custom web archiving User Defined Functions
LET’S SEE THIS IN ACTION...
Welcome to the Archives Unleashed Project

Welcome

Archives Unleashed aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past. Supported by a grant from the Andrew W. Mellon Foundation, we are developing web archive search and data analysis tools to enable scholars, librarians and archivists to access, share, and investigate recent history since the early days of the World Wide Web.

Interested in the project? Subscribe to our newsletter! Or you can follow the links at left for information about the project, the Archives Unleashed Cloud, Archives Unleashed Toolkit, Archives Unleashed Jupyter Notebooks, or our events.

We’re always looking for ways to engage archivists, librarians, researchers, developers, or any others interested in born-digital heritage!

Contact Us

Questions? Comments? Please contact us, either by leaving an issue on one of our GitHub projects or by sending us email. (See our Github Pages for links.):
CAN DO SOME COOL STUFF WITH THE TOOLKIT

FROM ARCHIVING TO ANALYSIS
WE MADE THE MISTAKE OF ASSUMING THESE WERE OUR MAIN USERS...
IN OTHER WORDS...

WE DIDN’T GO FAR ENOUGH... EXPECTED THEM TO COME TO US!
ENTER THE ANDREW W. MELLON FOUNDATION AND OUR GOAL OF MAKING AN ACCESSIBLE TOOL
PERSONA B

The sort of scholar we would have called a “digital humanist.” They’re comfortable with computers, can use some off-the-shelf tools like Voyant or Gephi to work with text/networks, and can use tutorials like those of the Programming Historian to learn some basic Python or R.

THE DIGITAL HISTORIAN
DIGITAL HUMANIST

- Needs to have some sort of toolkit to translate WARC files into something they can work with
- Can think critically about data
- Can use off-the-shelf tools to work with text/data, i.e. Voyant Tools or Gephi or other things based out of the "Programming Historian"
- In general doesn’t want to use the command line or write custom programs
COMPUTATIONAL HUMANITIES

- The Archives Unleashed Cloud was designed around this persona.

- Allows a user to take WARCs and:
  - Use a modern UI to sync their collections from the provider;
  - Run basic analyses in the browser to find major sites of interest;
  - Download derivative file formats that can integrate with standard workflows.

- In other words: let’s get the WARC out of the equation and *translate* it into a standard file format.
LET’S SEE THIS IN ACTION...
THE CLOUD IN ACTION

FROM ARCHIVING TO ANALYSIS
ENABLES THE SAME THINGS AS THE TOOLKIT. MINUS CUSTOMIZATION
WE’RE MOVING TOWARDS OUR USERS... BUT THEY STILL REQUIRE SOMEWHAT SPECIALIZED SKILLS TO INTERPRET DATA.
IN OTHER WORDS...

WE’RE GETTING THERE.. THEY COME PART WAY, WE COME PART WAY.
The sort of scholar who might use computers – like for Word, some light Excel – but otherwise generally just wants to do historical research without earning new technical skills.

THE CONVENTIONAL HISTORIAN
CONVENTIONAL HISTORIAN

- This is the toughest group to serve – if in the Toolkit model we weren’t going out far enough to serve people in a reasonable area; this person might not be coming out far enough to meet us...

- But we will try.
CONVENTIONAL HISTORIAN

- The **Archives Unleashed Jupyter Notebooks** were designed with this in mind.
  - Uses web browser (which we can all use);
  - Slightly confusing layout of pressing play buttons and executing code, but thanks to Markdown integration we can mark it up.
  - Extensively playtested (and playtesting) with PhD students and MA students who have varying levels of computational comfort.

- **Exploring one-click hosting with Google Colab/Google Drive**
STILL NOT THE EASIEST TO USE..

But we’re trying.
LET'S SEE THEM IN ACTION!
Welcome to the Text Analysis Notebook

Welcome to the Archives Unleashed Text Analysis Jupyter Notebook. This demonstration takes the full text derivatives from the Cloud and uses Python to analyze the derivatives for your collection.

Please feel free to create an issue to let us know about any bugs you encounter 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 derivative.

```
In [1]: COLLECTION_ID = '4867' # Change to switch collection.
```

```
In [2]: ASSET_PATH = 'data/' # Change value to full path to your data, including trailing slash.
```
IN OTHER WORDS...

WE'RE GETTING THERE... BUT THEY STILL NEED TO COME OUT A BIT..!
Historians in the future will need to understand the Web.

We need to make sure they're ready.
WANT TO LEARN MORE?

- I am awful at shameless self promotion because I have shame.

THANKS TO OUR FUNDERS!

THE ANDREW W. MELLON FOUNDATION

START SMART LABS

YORK UNIVERSITY

UNIVERSITY OF WATERLOO

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada