

Working with Web Archives

DSAC Presentation, Thursday January 14th

Ian Milligan
Assistant Professor
@ianmilligan1



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Nick Ruest
Digital Assets Librarian
@ruebot



Overview

- 1. Overview of funded projects/overview/current research projects.
- 2. Shape of our data and munging
- 3. Workflow
- 4. What could we do with library data?

Funded Projects

- Three SSHRC Grants
 - **Insight** (2015 - 2020) [Ian Milligan, Nick Ruest & William Turkel]
 - **Insight** Development Grant (2013 - 2016) [Ian Milligan, William Turkel]
 - **Connection Grant** (2015 - 2016) [Ian Milligan, Jimmy Lin, Matthew Weber, Nathalie Casemajor]
- Ontario Ministry of Research and Innovation **Early Researcher Award** (2015 -2020) [Ian Milligan]
- Compute Canada **Research Portals and Platforms** (2016 - 2018) [Ian Milligan, Nick Ruest + University of Alberta]

Why?

The Web as a Primary Source

- **Web archives will fundamentally affect the way historians write history**
 - We will have easier access to information on a previously-unknown scale, as well as improved capability to parse it;
 - Yet historians need to reflect on the shape that Web-based primary sources will take, and **how we will be able to access them**

199

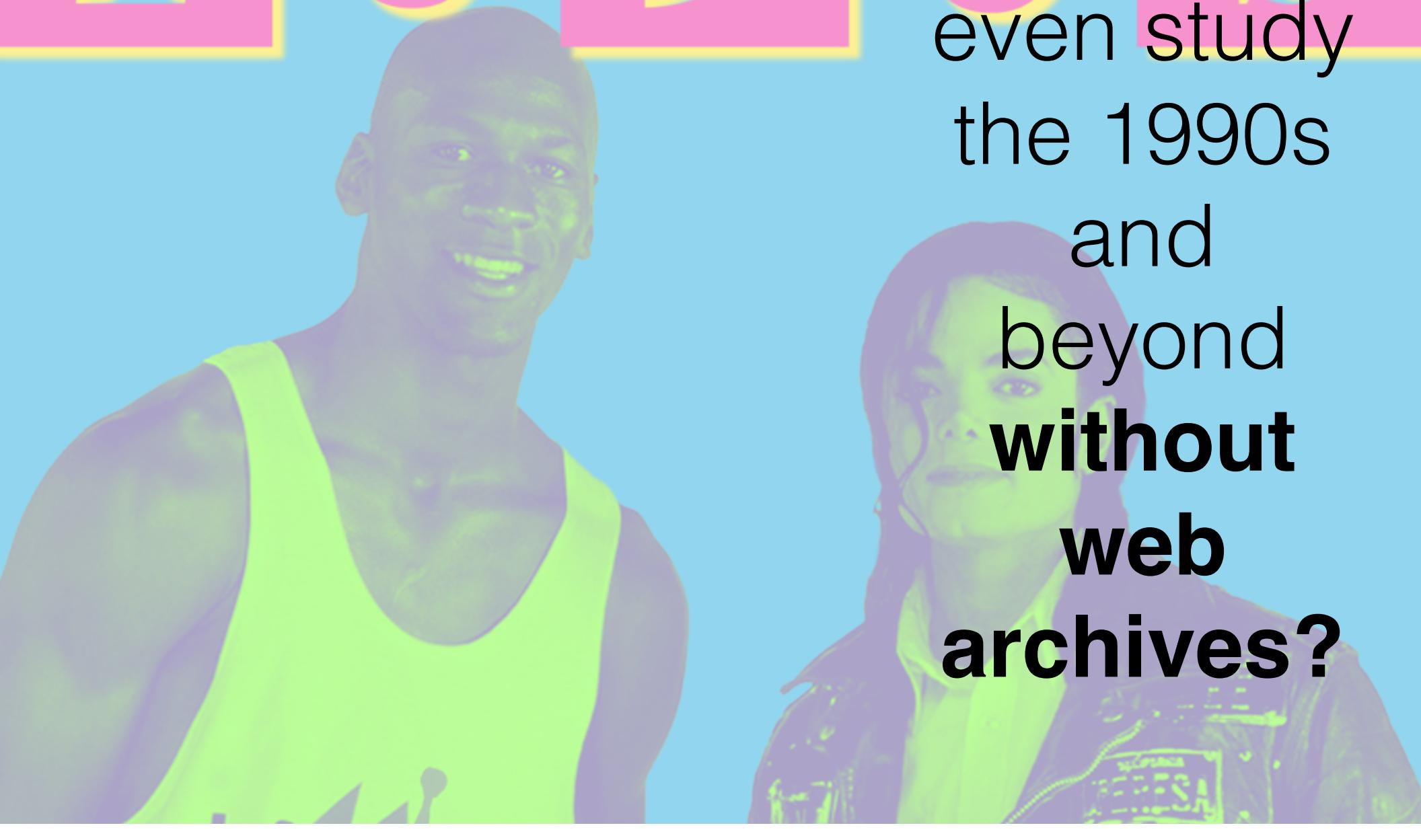
99

99

0

S

Could one
even study
the 1990s
and
beyond
without
web
archives?



Early Outputs

Trend results 2005-2015 x Ian

← → C Home webarchives.ca/graph

Web Archives for Historical Research - Canadian Politics Search Trends About Datasets

Welcome to the Web Archives for Historical Research political parties portal. Before diving in, we encourage you to visit our [about](#) page. X

The graph shows the relative frequency of a term in each year - i.e. how often did "recession" appear in 2008 as a percentage of all keywords? Enter your search terms and click "Update Graph." To see actual results, click on the graph's line where it intersects a year.

You can also use this graph to run keyword searches on particular top-level domains. To see how often the word "global warming" appeared in the Conservative Party of Canada's domain versus the New Democratic Party of Canada, search with the query: "global warming" domain:conservative.ca, "global warming" domain:ndp.ca [\[view this example\]](#).

For a full list of domains, consult the [about page](#).

Colours are assigned to the graph below depending on the order they are inputted (i.e. the Conservative Party of Canada will not always be blue).

recession,depression 2005 2015 Update Graph Reset All Download

Percentage

2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 Date

Some Sample Queries to Get Started With:

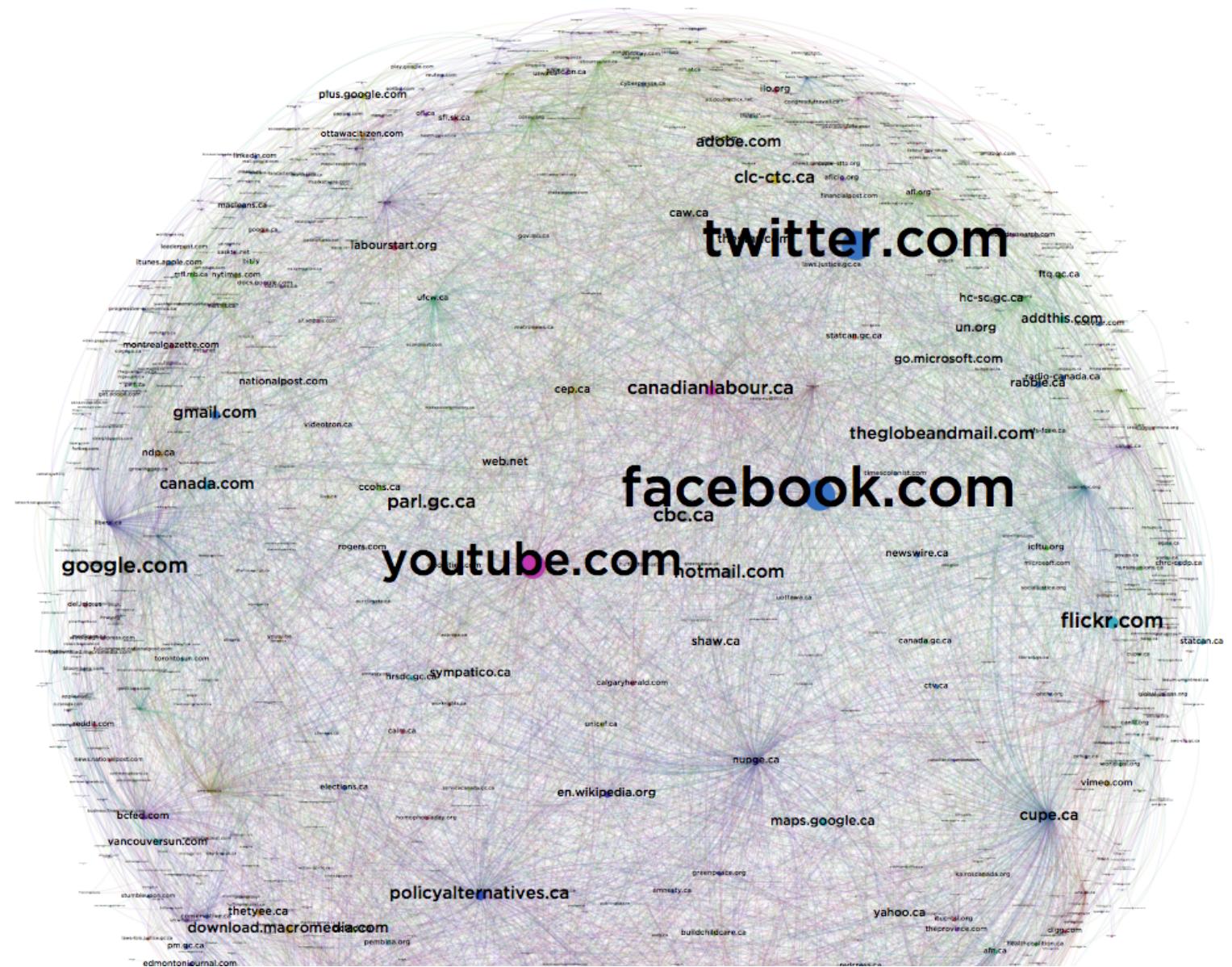
- "[Public Transit](#)": Liberals/NDP/Conservatives. We use quotation marks to search on the exact phrase.
- "[Tar Sands](#)": Amongst the three major parties, who uses and subsequently stops using this phrase for Canada's Oil Sands?
- "[Climate Change](#)": Similar to the above two queries, we can see this issue rise and possibly wane in public interest.

Feel free to take these examples and adapt them to your own purposes. If you find something interesting, [please let us know!](#)

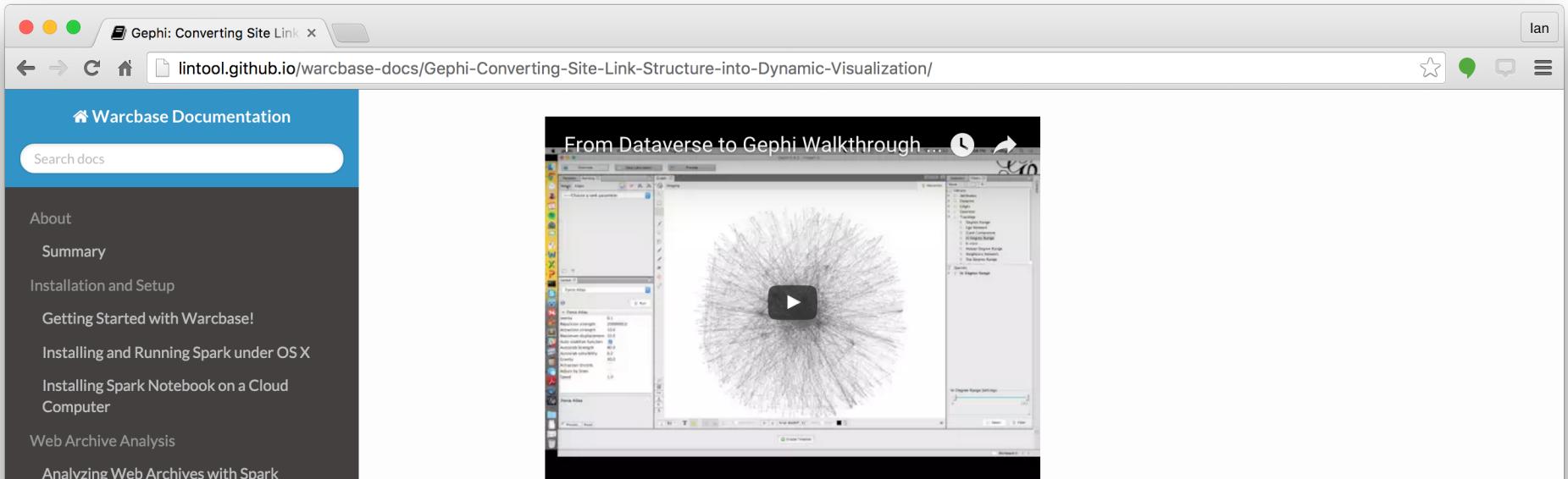
Click on a point on the graph to show a random sample of the matching records from that year...

Date	recession (%)	depression (%)
2005	0.0	0.4
2006	0.7	0.4
2007	0.8	0.3
2008	1.0	0.5
2009	1.6	0.8
2010	2.0	2.8
2011	2.0	2.3
2012	1.0	0.5
2013	0.7	0.5
2014	0.6	0.5
2015	0.5	0.6

Early Outputs



Early Outputs



Step One: Generate GDF Format Output

You can write directly to a Gephi-readable format by using our [WriteGDF](#) function. Here is an example script:

```
import org.warcbase.spark.matchbox.RecordTransformers._  
import org.warcbase.spark.matchbox.{ExtractTopLevelDomain, ExtractLinks, RecordLoader, WriteGDF}  
import org.warcbase.spark.rdd.RecordRDD._  
  
val links = RecordLoader.loadArc("/collections/wearchives/CanadianPoliticalParties/arc/", sc)  
.keepValidPages()  
.map(r => (r.getCreateDate, ExtractLinks(r.getUrl, r.getContentString)))  
.flatMap(r => r._2.map(f => (r._1, ExtractTopLevelDomain(f._1.replaceAll("^\\s*www\\.", ""), ExtractTopLevelDomain(  
.filter(r => r._2 != "" & r._3 != "")  
.countItems()  
.filter(r => r._2 > 5)  
WriteGDF(links, "all-links.gdf")
```

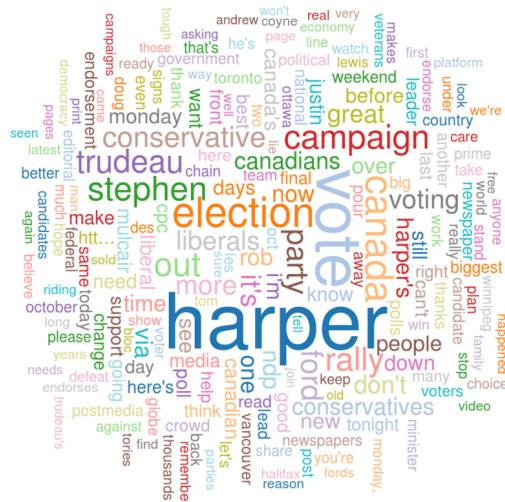
The ensuing [all-links.gdf](#) can be natively imported into Gephi.

Step Two: Import into Gephi

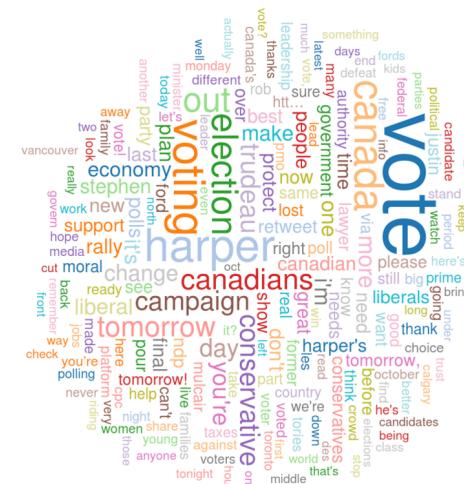
You now want to take it into Gephi. [Install Gephi](#) – make sure you're running Gephi 0.9, as it offers widest compatabilities with systems.

Early Outputs

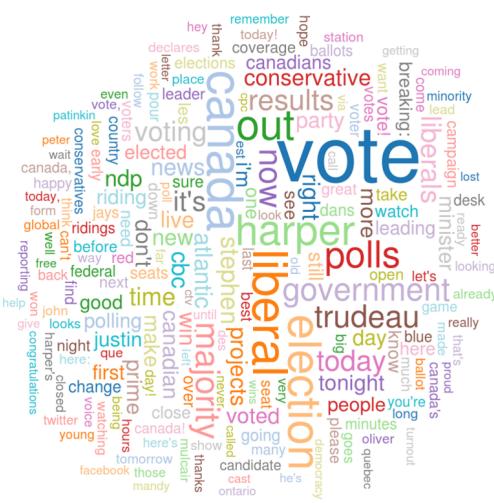
October 17, 2015



October 18, 2015



October 19, 2015



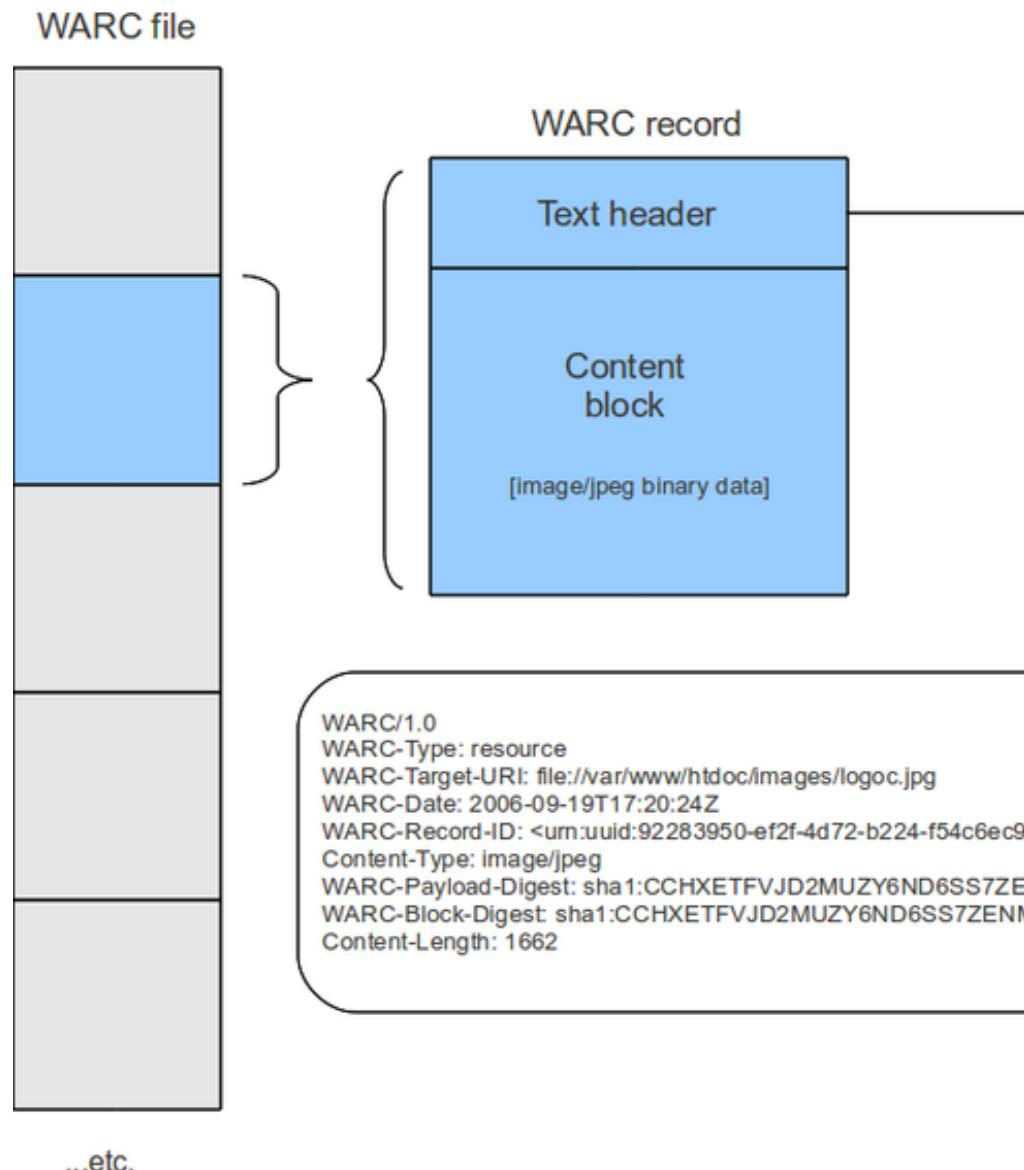
October 20, 2015



Shape of our data?

WARC File

- WebARChive Container Files (WARC), 28500:2009
- Concatenated web objects



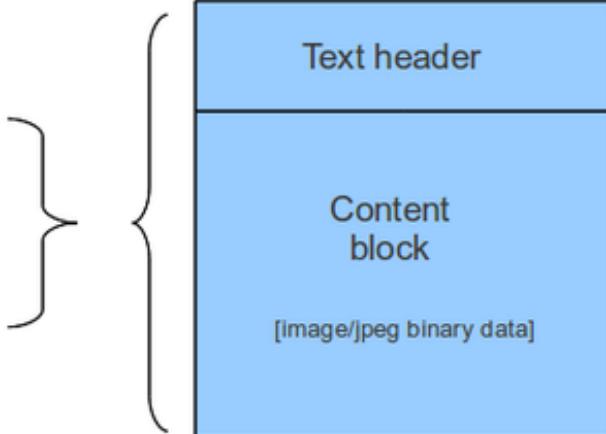
WAT File

- WARC minus content

WARC file



WARC record

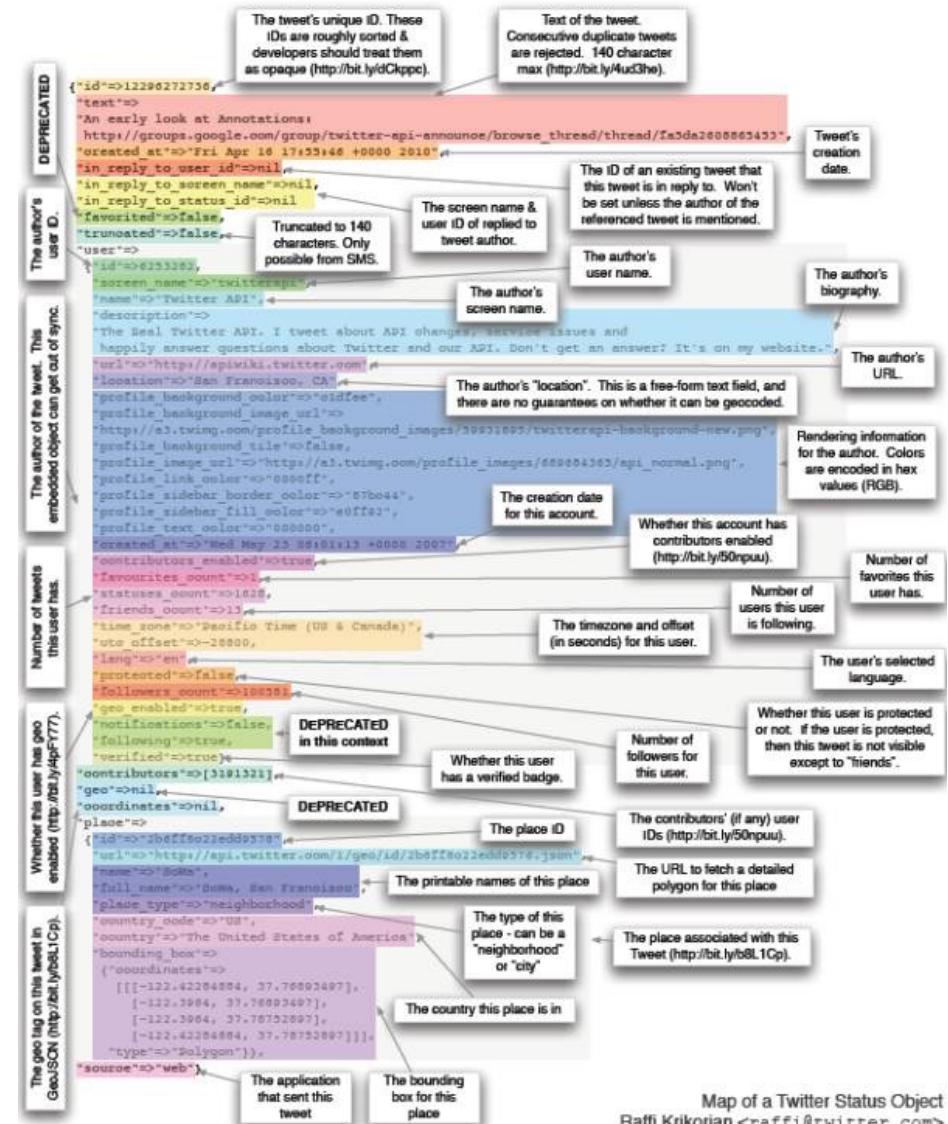


```
WARC/1.0
WARC-Type: resource
WARC-Target-URI: file:///var/www/htdoc/images/logoc.jpg
WARC-Date: 2006-09-19T17:20:24Z
WARC-Record-ID: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec9
Content-Type: image/jpeg
WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZE
WARC-Block-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENM
Content-Length: 1662
```

...etc.

Twitter Data

- JSON
 - Collected using Twitter's Streaming API (or Search API)
 - Generated ourselves or with fellow travellers like Library and Archives Canada



Data Munging

- WARC Files
 - Warcbase – from ingest, indexing, text analysis, entity extraction, etc. Based on the Spark platform.
- Twitter JSON
 - twarc (<https://github.com/edsu/twarc>)
 - bash scripting
 - jq (cat elxn42-tweets.json | jq -c '.text' | cat > elxn42-tweets-text.txt)
- Mathematica

Workflow

- **Computing provided by York University Libraries**
- 1. Downloading or scraping data (Internet Archive via wget or sneakernet)
- 2. Ingesting or indexing into visualization platform
- 3. Analytics run
- 4. Exporting in various datasets (GEXF for Gephi, TXT or CSV for textual analysis, other formats for various visualizations - i.e. NER)

Library-Provided Data?

- Internet Archive Data
- **Archive-It Data from OCUL Partners** - we would love to use this to expand our coverage in <http://webarchives.ca>
- **Other contemporary data** - would be interesting to link up with our web archives, find connections between web archive corpora and other datasets

Propose one or two ways that library-provided data (including Internet Archive data) could advance your work if you had a different kind of access to it or could use it with different tools than ones currently available?

One Way

- More **WARCs** from the Internet Archive
 - GeoCities
 - .ca Top-Level Domain (moon shot)
- More **WARCs** from Archive-It Partners
 - Canadian-based collections
 - Even a comprehensive list of what other non-University of Toronto partners have would be useful, currently no good search engine!

Another Way

- **Support in making datasets interoperable**
 - Could we get the index that powers [webarchives.ca](#) to speak to other databases that the library might have to offer?



Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada



Ontario



compute • calcul
CANADA



UNIVERSITY OF
WATERLOO

Generous Acknowledgements and Thanks

Ian Milligan
Assistant Professor
@ianmilligan1

Nick Ruest
Digital Assets Librarian
@ruebot



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

YORK
UNIVERSITÉ
UNIVERSITY