Between Metadata and Content

Exploring Canadian Political History with Archive-It's Research Services

---

Ian Milligan
Assistant Professor
@ianmilligan1

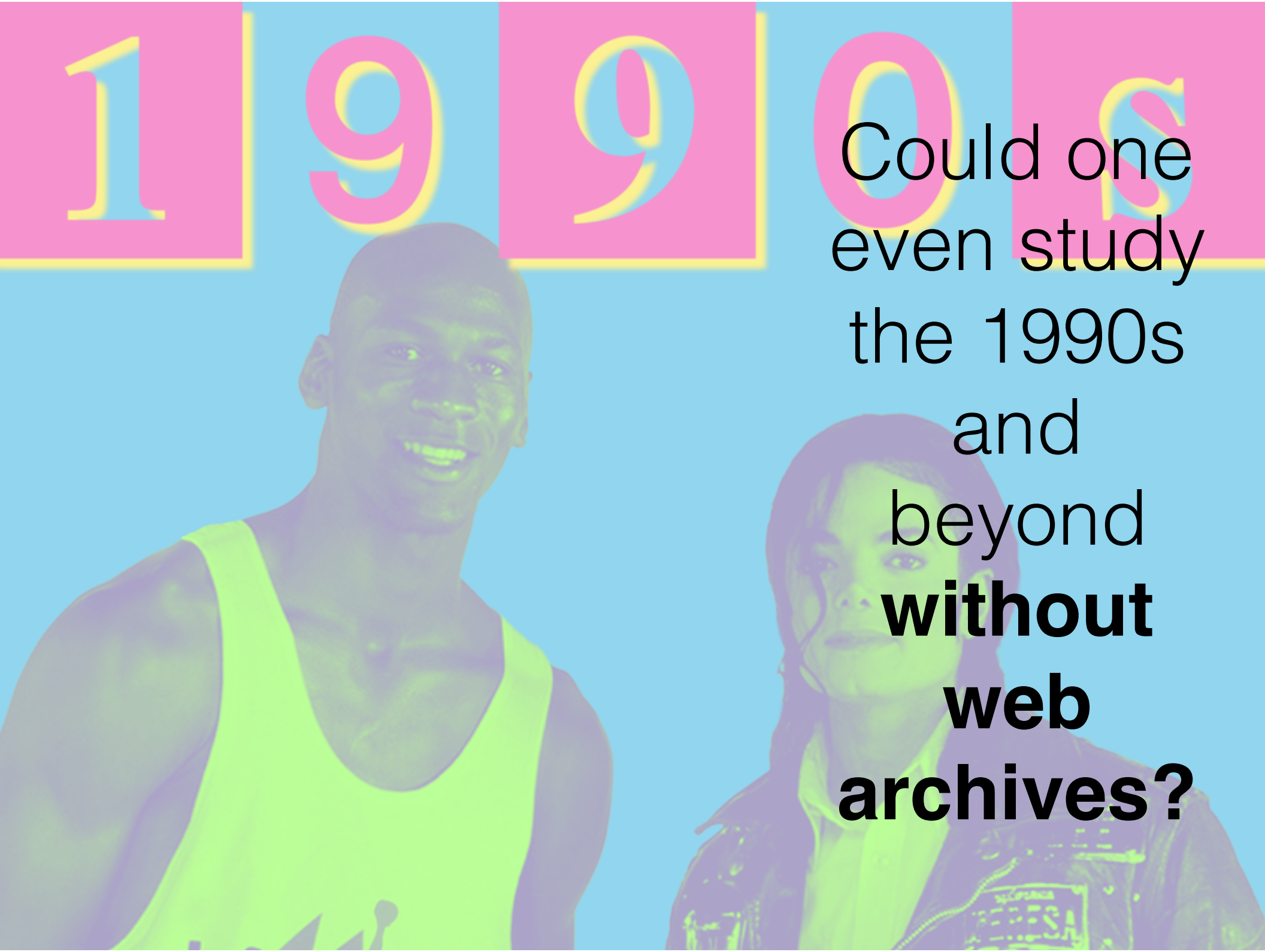
UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Jimmy Lin
Professor and David R. Cheriton Chair
@lintool

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
David R. Cheriton School
of Computer Science

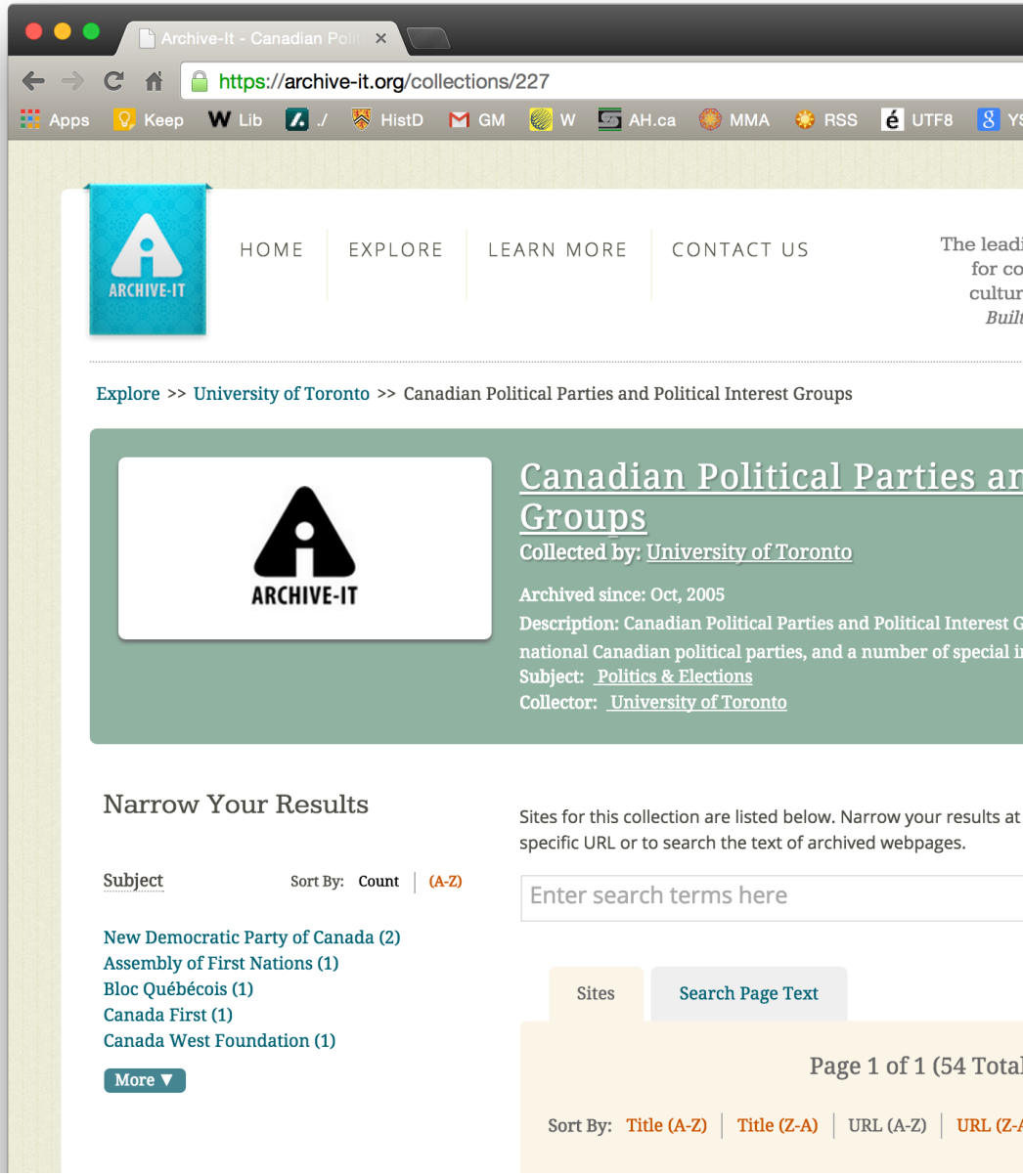
Jeremy Wiebe
PhD Candidate
@jeremyw

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History


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

The Case Study

- **Archive-It Research Services**: "Canadian Political Parties and Political Interest Groups"
- 2005 - 2015
- WAT & WARC files

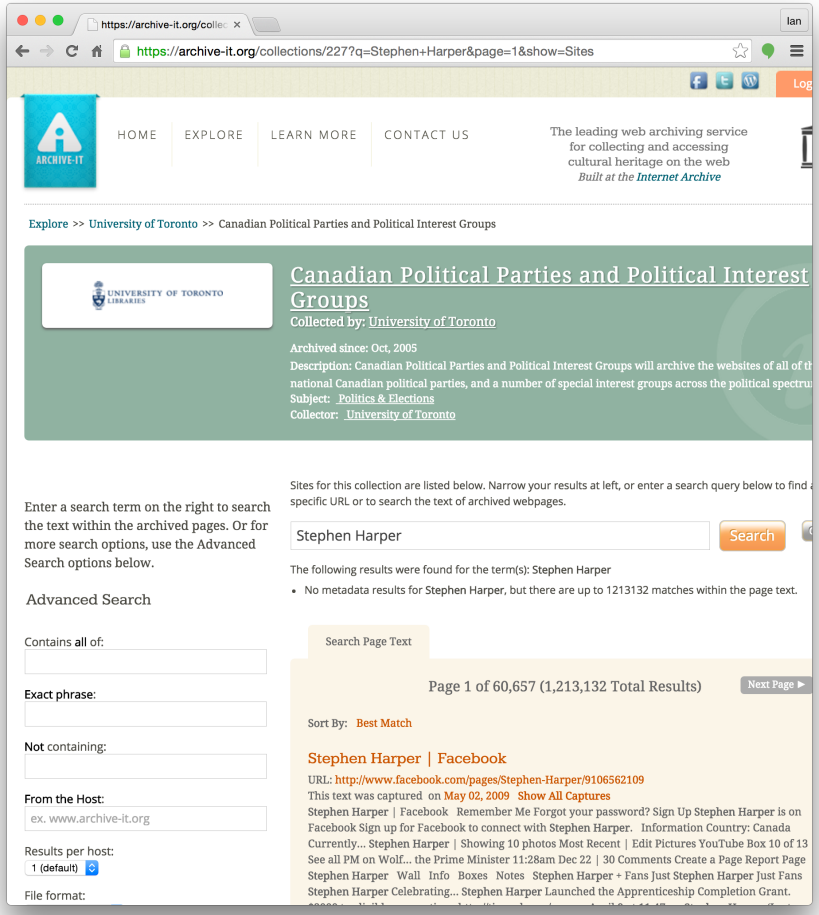


Pivotal Changes in Canadian Politics, 2005-2015

- Militarization of Canadian society?
- Change from ‘natural governing party’ of Liberals to Conservatives
- Major policy changes on foreign policy, environment, etc.
- How to measure?

Current Interface

- Very limited - simple search engine, some advanced options; no facets
- Great collections.. but nobody uses them!
- https://archive-it.org/collections/227

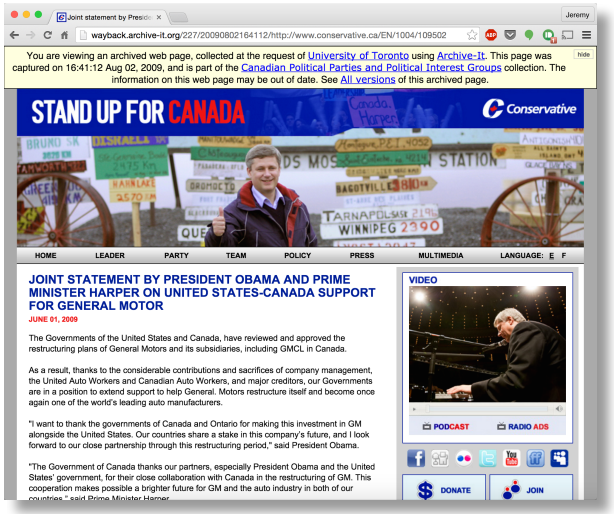


How to provide access?

Warcbase

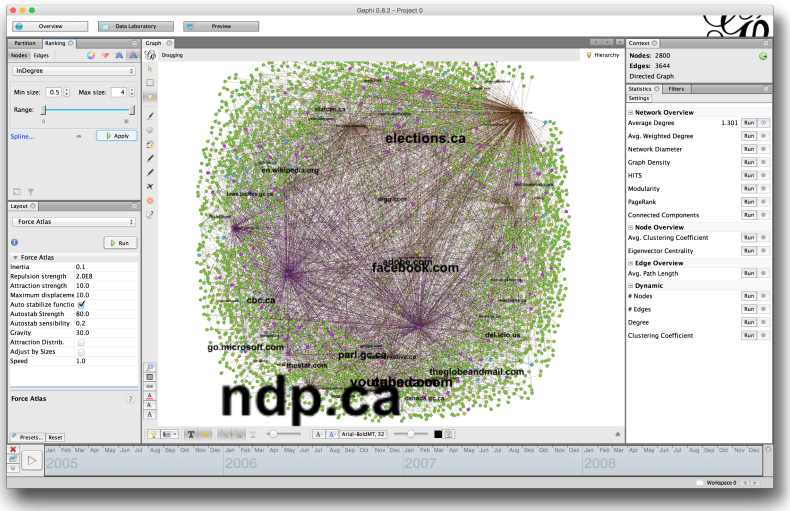
An open-source platform for managing web archives

http://warcbase.org

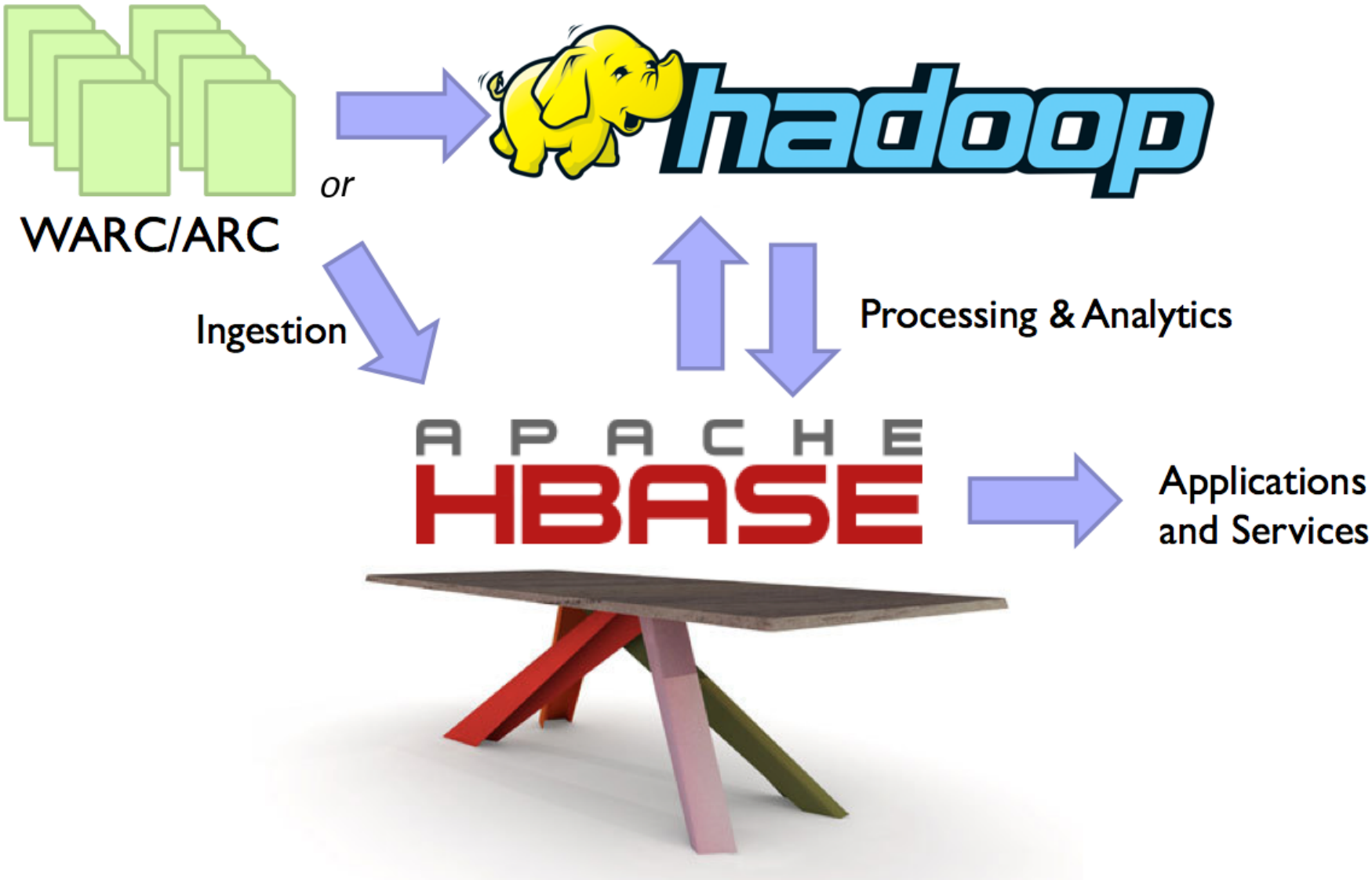


Two main facets

- A flexible data store: your own Wayback Machine
- Scriptable analytics and data processing



Warcbase


WARC/ARC
or
hadoop
Ingestion
Processing & Analytics
APACHE
HBASE
Applications
and Services

Warcbase

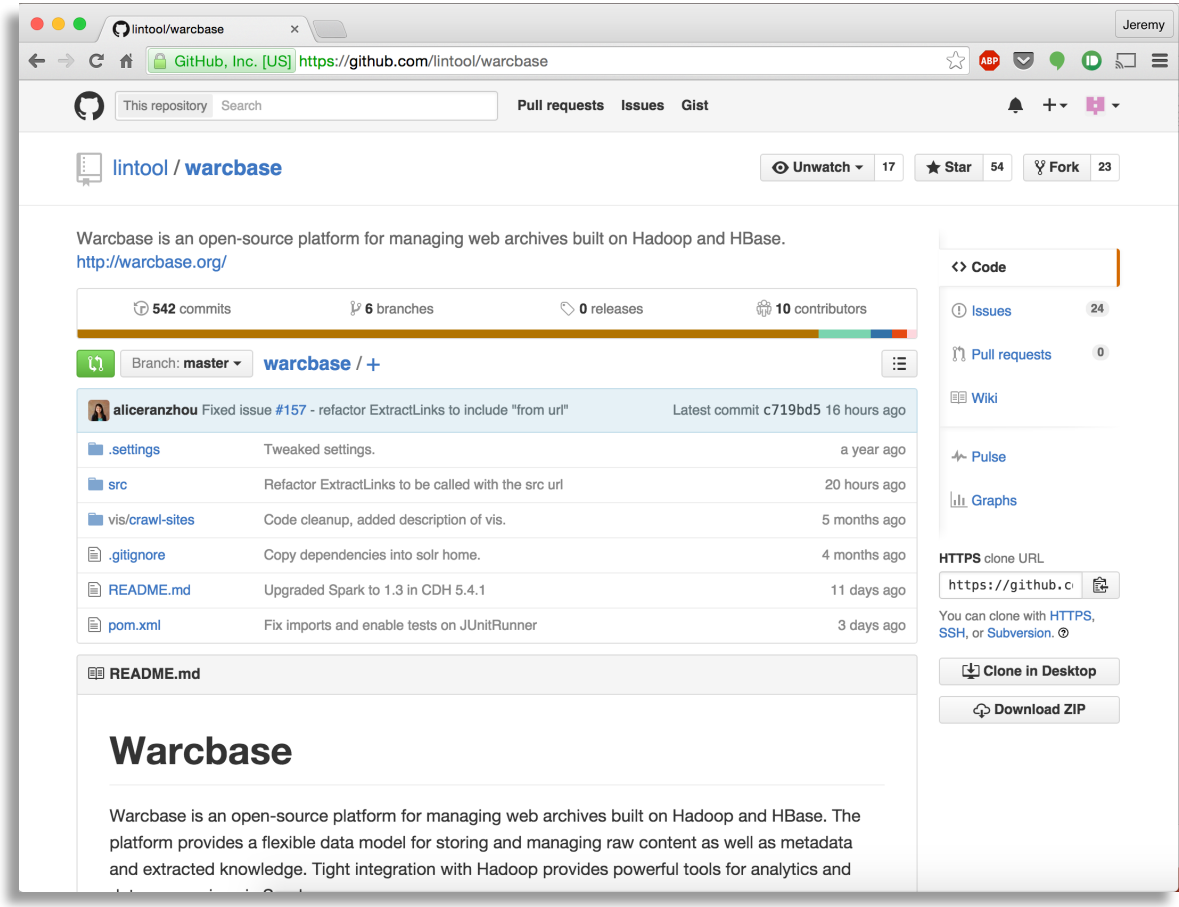
- Framework for distributed storage and distributed processing of very big data
- Scalable
- Potentially very powerful
  - *Trantor*: 1.2PB of disk, 25 compute nodes (each w/ 128GB memory, 2×6-core Intel Xeon E5 v3 = 3.2TB memory and 300 current-generation Intel cores)

- In active development, led by Jimmy Lin, collaborator with Web Archives Historical Research Group

You Can Warcbase Too!

http://warcbase.org

- Flexible requirements: designed for powerful distributed systems, but will also run on your laptop
  - Linux or OS X
- Comprehensive documentation, with script examples



Analysis: Scripting

- Previously, Pig scripts ran Hadoop MapReduce jobs
  - Deprecation imminent
- Currently transitioning to Spark, scripts written in Scala
  - Better performance
  - API and libraries for Python (*PySpark*) – hope to facilitate adoption

A Simple Script

Domain level plain text extraction (Spark)

```
import org.warcbase.spark.matchbox.ArcRecords
import org.warcbase.spark.matchbox.ArcRecords._

val r = ArcRecords.load("src/test/resources/arc/example.arc.gz", sc)
  .keepMimeTypes(Set("text/html”))
  .discardDate(null)
  .keepDomains(Set("greenparty.ca"))
  .extractDomainUrlBody()

r.saveAsTextFile("out/")
```

Named Entity Recognition

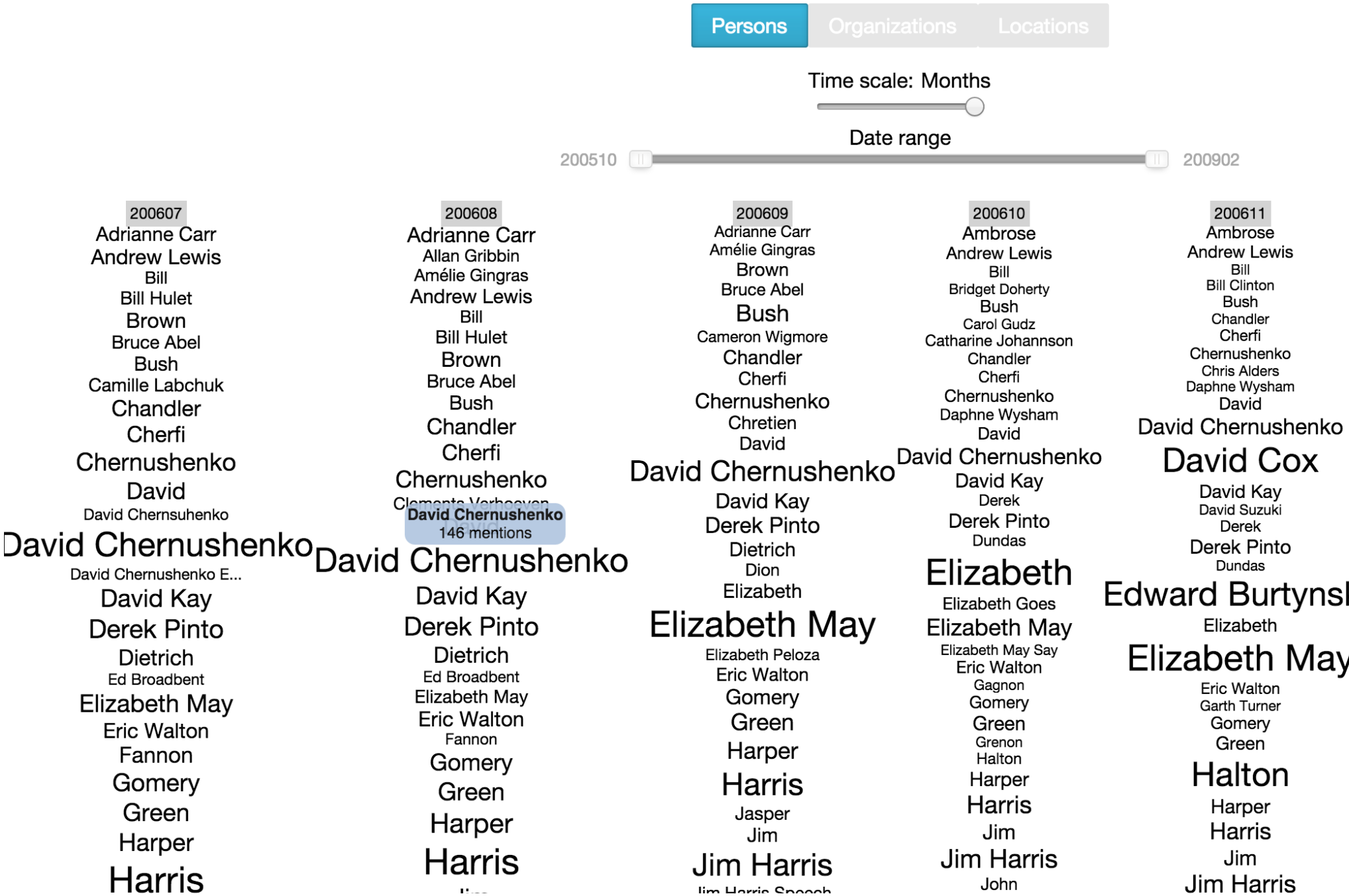
Some interesting results

- Stanford Named Entity Recognition library
- Call by script
- A useful discovery tool

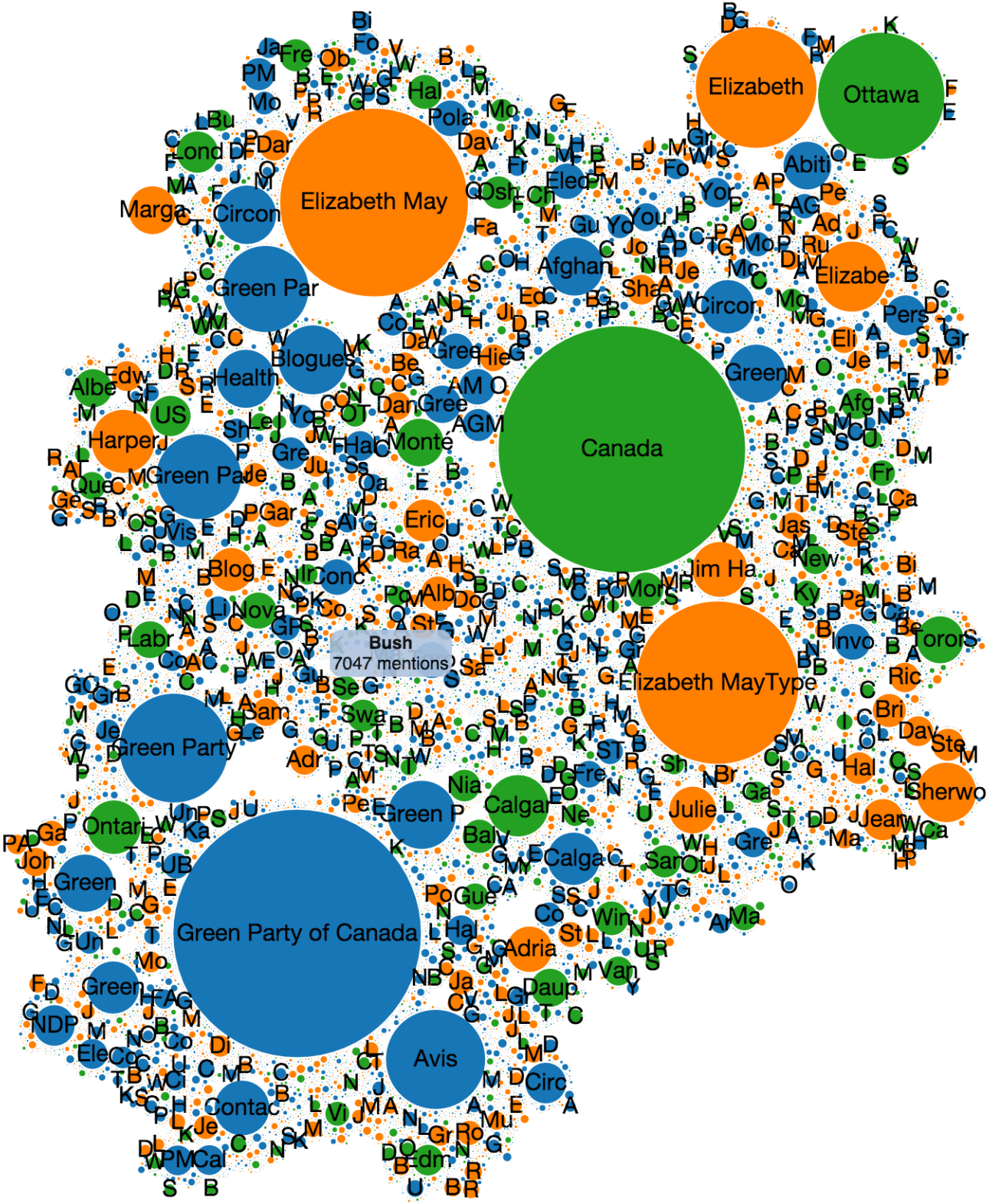
```
%default I_PARSED_DATA_DIR '/user/jrwiebe/cpp.text-greenparty/*.txt';
%default O_ENTITIES_DIR '/user/jrwiebe/cpp.text-greenparty/entities.gz/';
%default I_NER_CLASSIFIER_FILE 'english.all.3class.distsim.crf.ser.gz';

SET mapred.max.map.failures.percent 10;
SET mapred.reduce.slowstart.completed.maps 0.9

REGISTER 'target/warcbase-0.1.0-SNAPSHOT-fatjar.jar';

DEFINE NER3CLASS org.warcbase.pig.piggybank.NER3ClassUDF('$I_NER_CLASSIFIER_FILE');

Scrapes = LOAD '$I_PARSED_DATA_DIR' AS (date:chararray, url:chararray, content:chararray);

Entities = FOREACH Scrapes GENERATE date, url, NER3CLASS(content) AS entityString;

STORE Entities into '$O_ENTITIES_DIR';
```

Named Entity Visualization

Data source: greenparty.csv




Manley
22 mentions

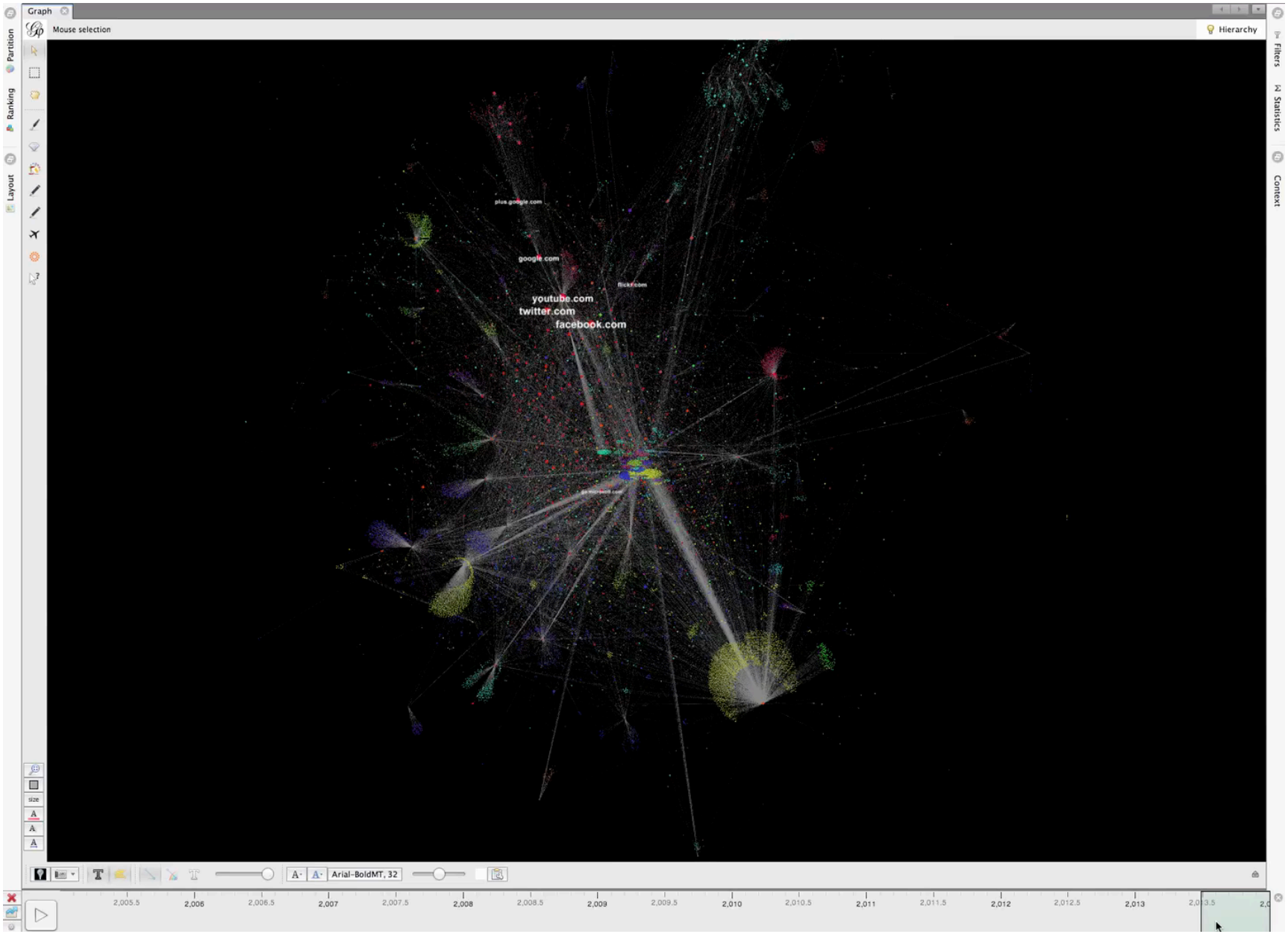

Elizabeth
Ottawa
Elizabeth May
Canada
Green Party
Green Party of Canada
Elizabeth MayType
Bush
7047 mentions

Historians want content

- but metadata might actually be more useful!

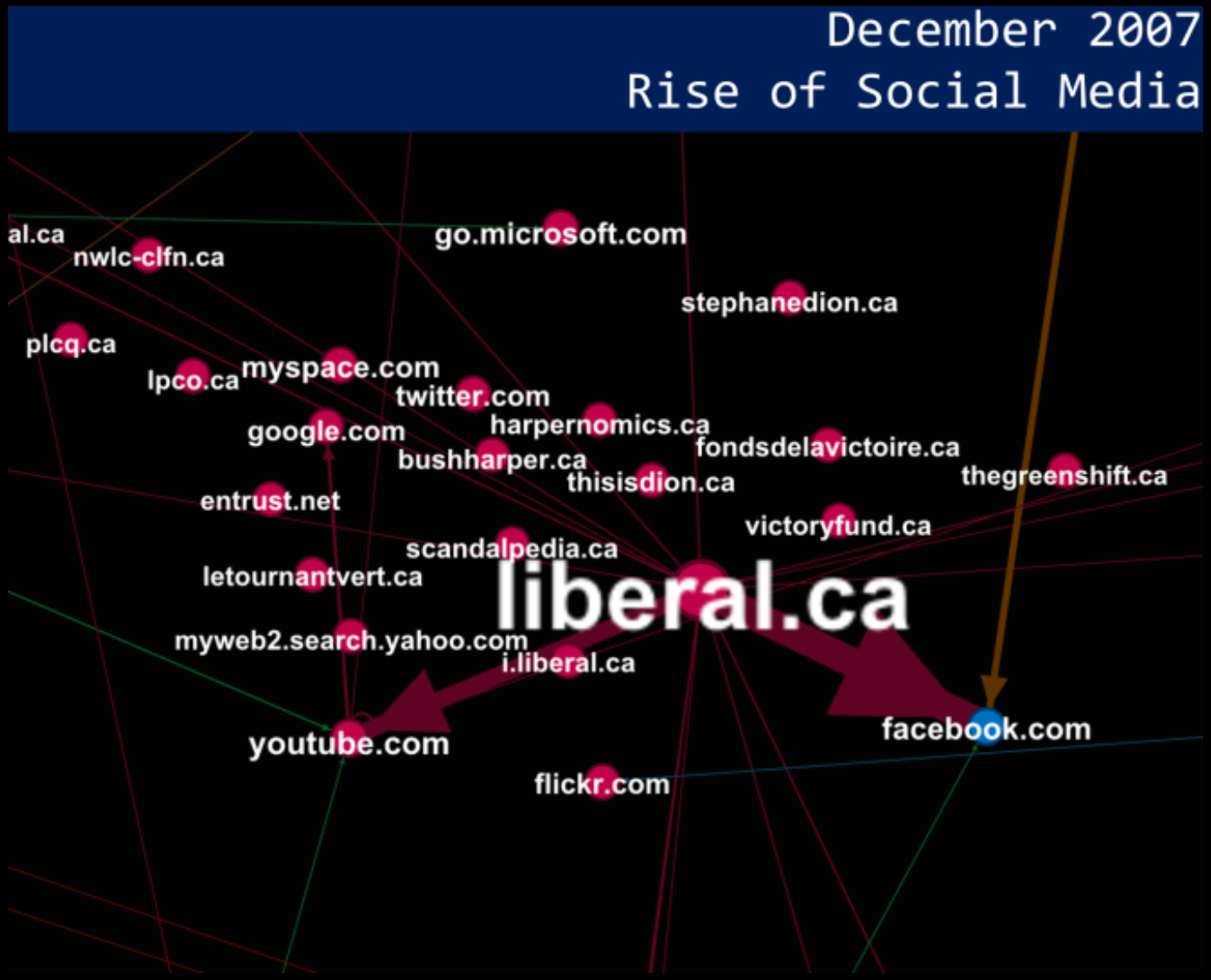
So here’s some cool stuff we can do with warcbase.

Metadata Extraction

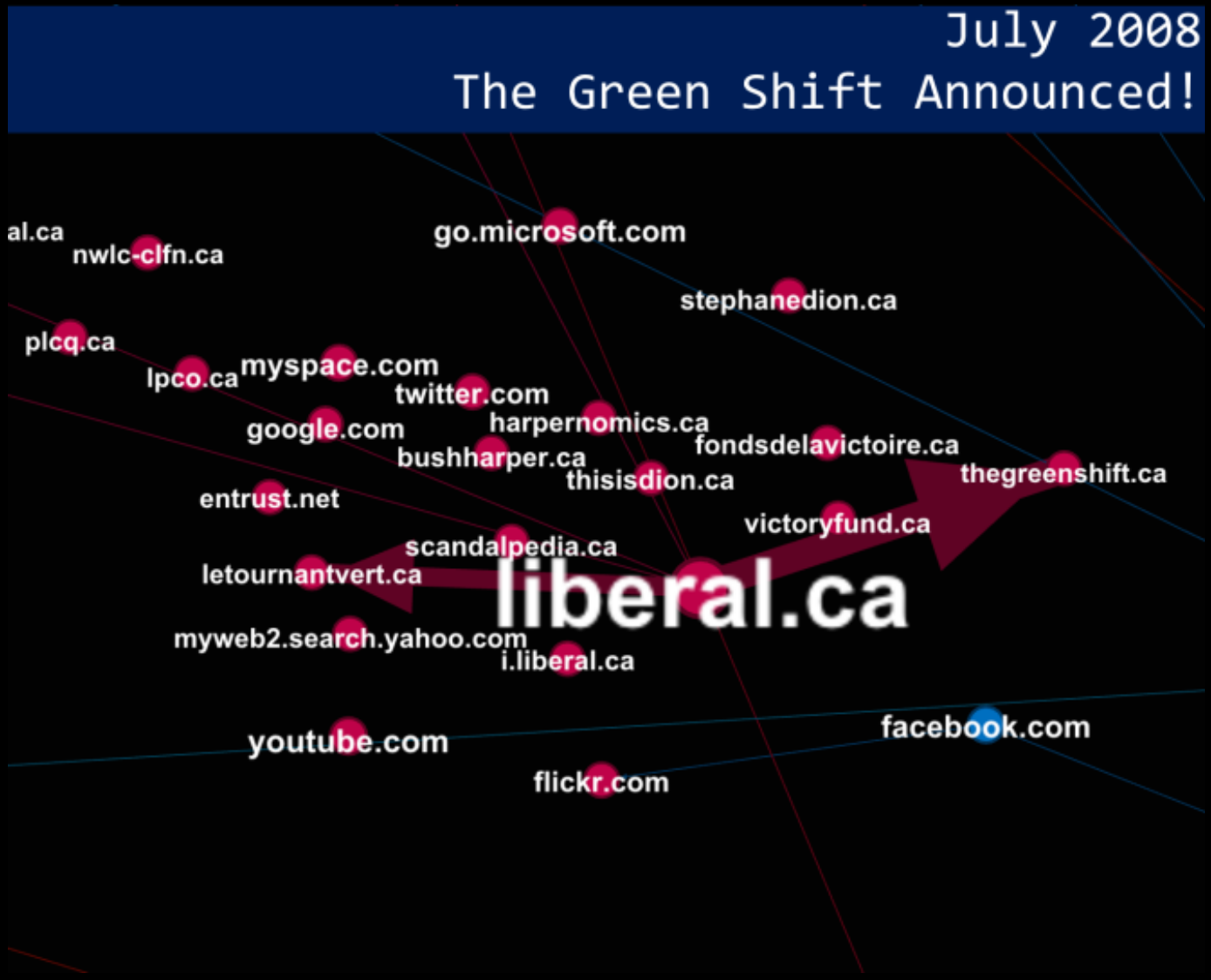


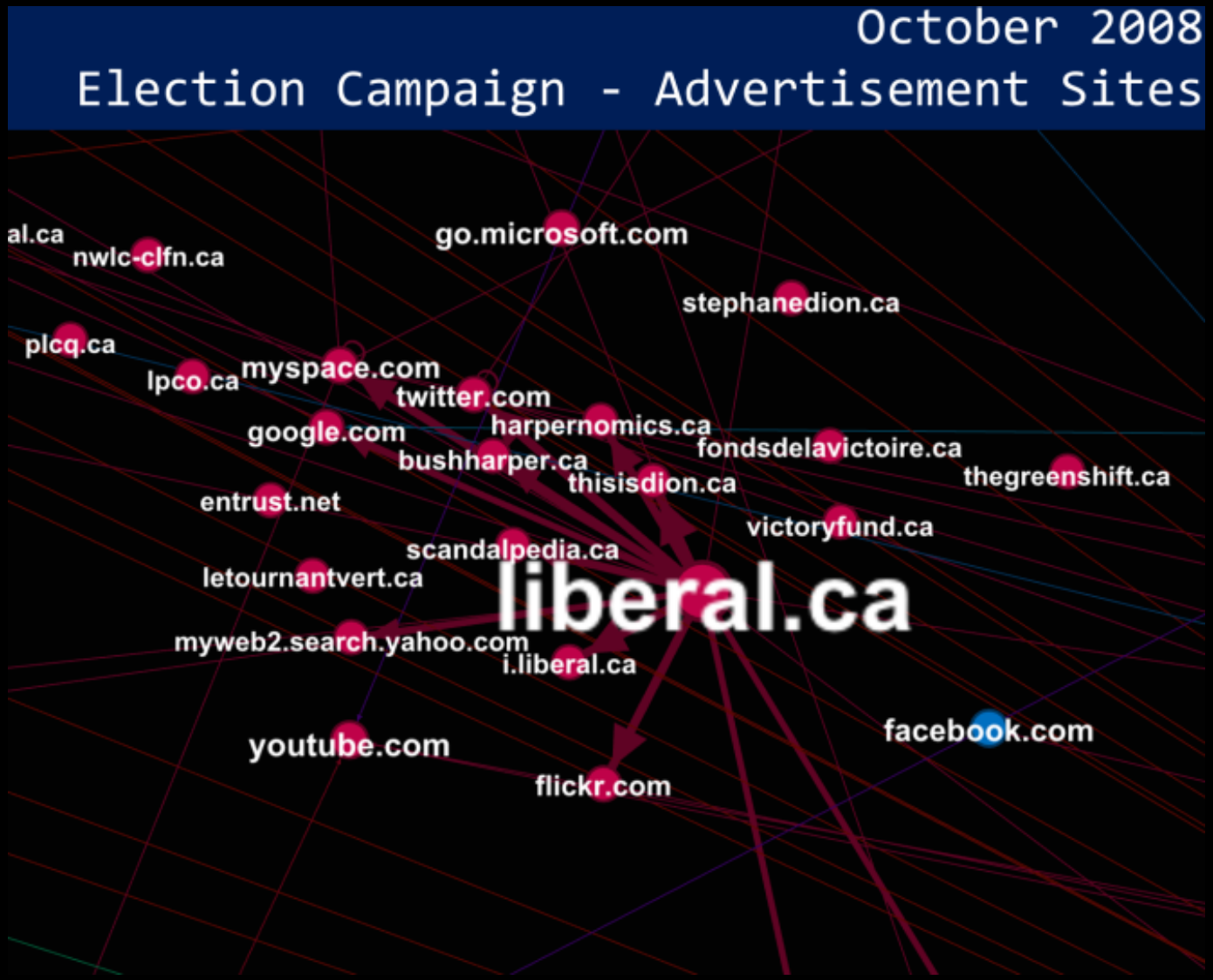
Short stories from metadata as well


December 2006
Stephane Dion Elected Leader of Party
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
facebook.com
youtube.com
flickr.com

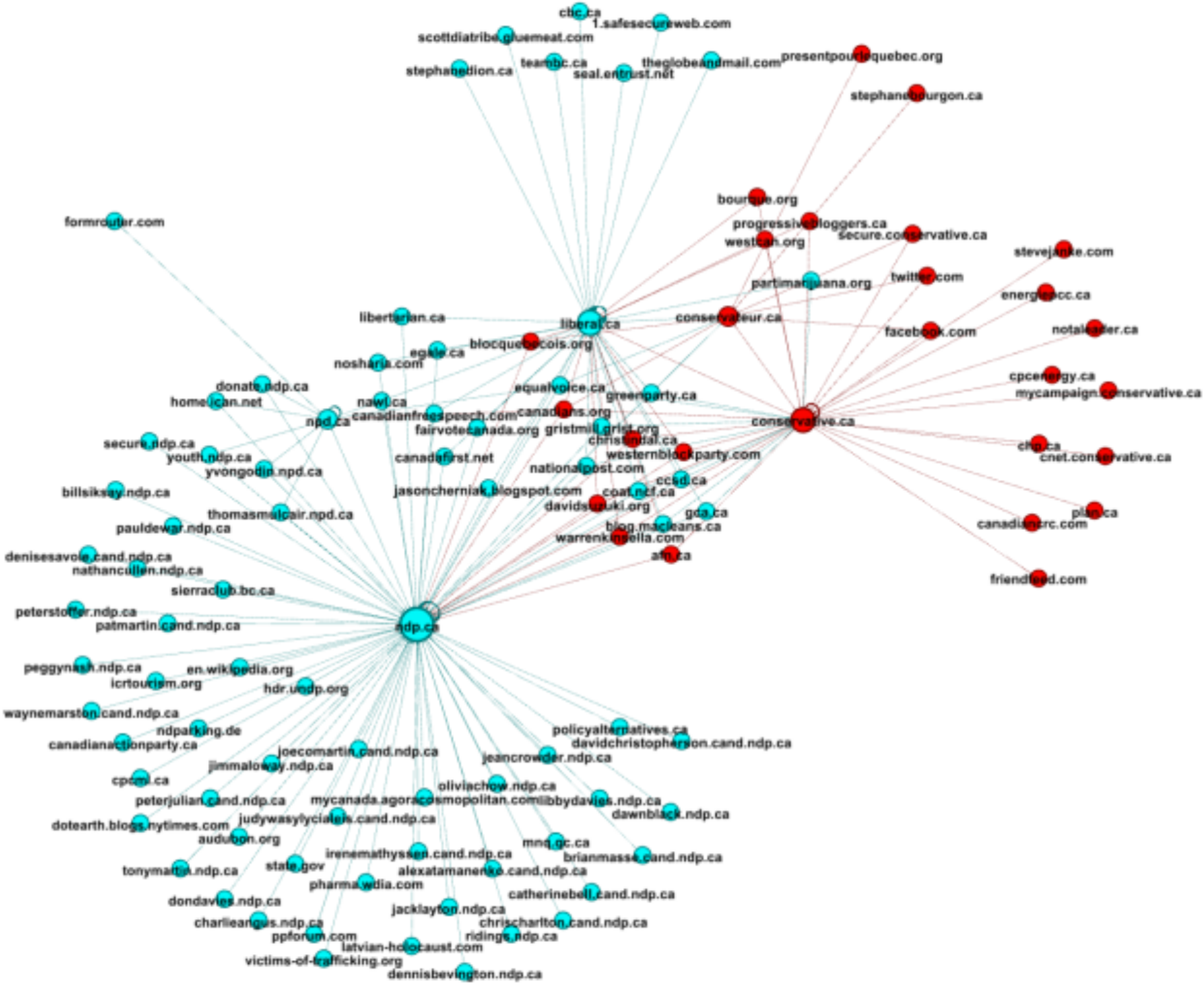

December 2007
Rise of Social Media
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
youtube.com
facebook.com
flickr.com


April 2008
Fundraising with the Victory Fund/
Fonds de la Victoire
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
facebook.com
youtube.com
flickr.com

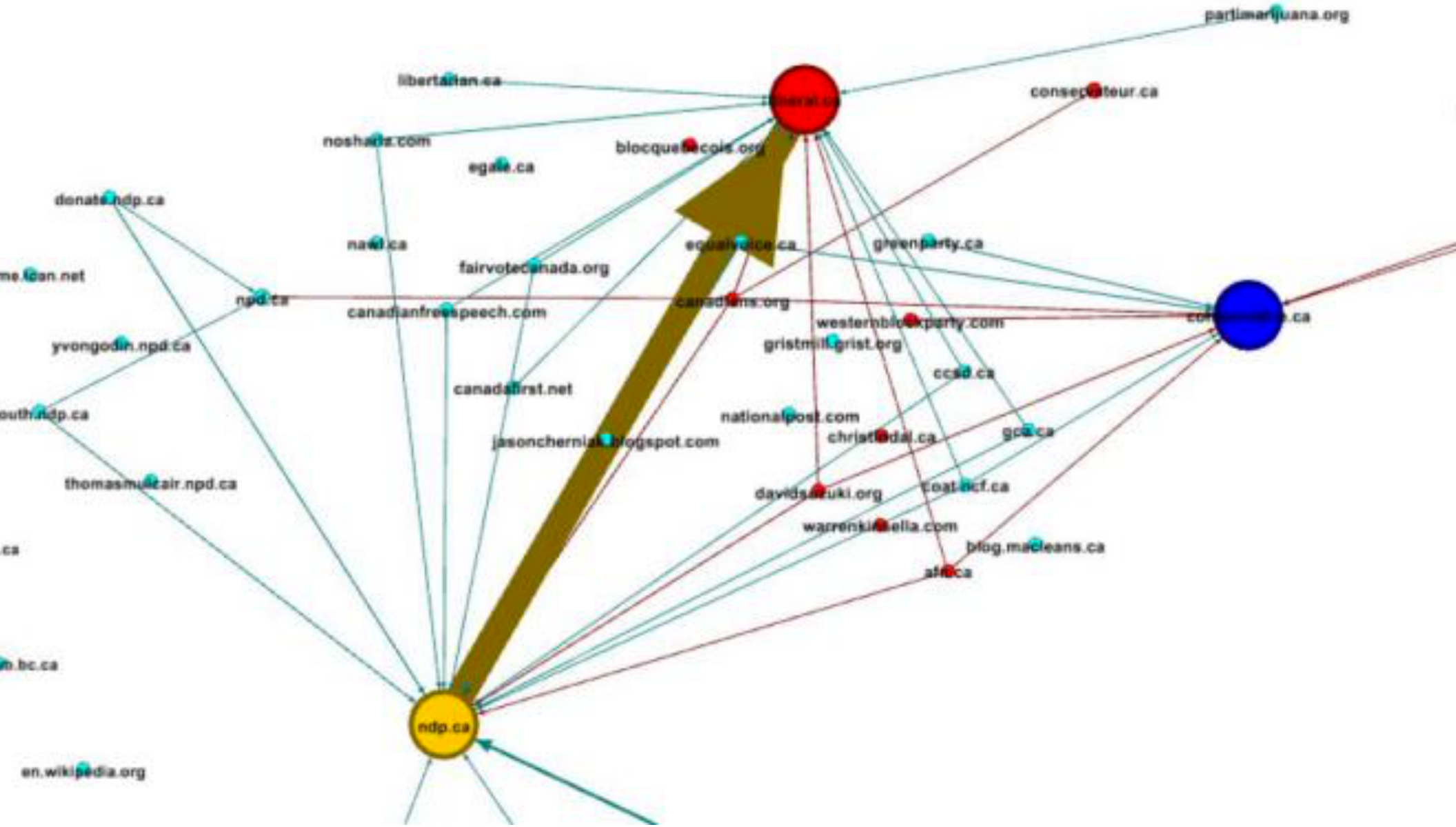

July 2008
The Green Shift Announced!
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
youtube.com
facebook.com
flickr.com


October 2008
Election Campaign - Advertisement Sites
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
youtube.com
facebook.com
flickr.com


December 2008
Election campaign Ends; Attacking Harper
on Anti-American Grounds (bushharper)
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com
i.liberal.ca
facebook.com
youtube.com
flickr.com


cbc.ca
1.safesecureweb.com
scottdiatribe.gluemeat.com
stephanedion.ca
teambc.ca
seal.entrust.net
theglobeandmail.com
presentpourlequebec.org
stephanebourgon.ca
bourque.org
progressivebloggers.ca
secure.conservative.ca
westcoh.org
stevejanke.com
twitter.com
partimarijuana.org
energiepcc.ca
formrouter.com
liberty.ca
liberal.ca
conservateur.ca
facebook.com
notaleader.ca
blocquebecois.org
egale.ca
nosharia.com
cpcenergy.ca
mycampaign.conservative.ca
donate.ndp.ca
equalvoice.ca
greenparty.ca
home.ican.net
nawl.ca
canadians.org
canadianfreespeech.com
npd.ca
fairvotecanada.org
gristmill.grist.org
conservative.ca
christendal.ca
secure.ndp.ca
youth.ndp.ca
canadafirst.net
westernblockparty.com
chp.ca
cnet.conservative.ca
yvongodin.npd.ca
nationalpost.com
billsiksay.ndp.ca
ccsd.ca
jasoncherniak.blogspot.com
coat.ncf.ca
davidsuzuki.org
gda.ca
thomasmulcair.npd.ca
blog.macleans.ca
pauldewar.ndp.ca
warrenkinsella.com
plan.ca
canadiancrc.com
denisesavoie.cand.ndp.ca
atn.ca
nathancullen.ndp.ca
friendfeed.com
sierraclub.bc.ca
peterstoffer.ndp.ca
patmartin.cand.ndp.ca
ndp.ca
peggynash.ndp.ca
en.wikipedia.org
icrtourism.org
hdr.undp.org
waynemarston.cand.ndp.ca
policyalternatives.ca
ndparking.de
davidchristopherson.cand.ndp.ca
canadianactionparty.ca
joecomartin.cand.ndp.ca
jeancrowder.ndp.ca
jimmaloway.ndp.ca
cpcml.ca
oliviachow.ndp.ca
peterjulian.cand.ndp.ca
mycanada.agoracosmopolitan.com
libbydavies.ndp.ca
dawnblack.ndp.ca
dotearth.blogs.nytimes.com
judywasylycia-leis.cand.ndp.ca
audubon.org
mnq.gc.ca
irenemathyssen.cand.ndp.ca
brianmasse.cand.ndp.ca
tonymartin.ndp.ca
state.gov
alexatamanenko.cand.ndp.ca
pharmawtdia.com
catherinebell.cand.ndp.ca
dondavies.ndp.ca
jacklayton.ndp.ca
charlieangus.ndp.ca
chrischarlton.cand.ndp.ca
ppforum.com
ridings.ndp.ca
latvian-holocaust.com
victims-of-trafficking.org
dennisbevington.ndp.ca

2005 Canadian Federal Election



So metadata, of course, helps historians find the content they need!

But content is always easier to sell to the public..

Shine/WebArchives.ca

- UK Web Archive's Shine (https://github.com/ukwa/shine)
- Indexing as bottleneck; warcbase includes a walkthrough for generating Lucene indexes - ~ 5 days locally, few hours with a cluster

Demo

Five Things I've Learned

- Political parties delete content
- User-generated comments were more common in political parties
- Absences can be more informative than presences
- We can see the rise/fall of prominent people
- Enabling user access is truly transformative

Bringing it all together?
(stay tuned for tomorrow)

Thank You!

Any Questions?

Ian Milligan
Assistant Professor
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Jimmy Lin
Professor and David R. Cheriton Chair
@lintool

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
David R. Cheriton School
of Computer Science

Jeremy Wiebe
PhD Candidate
@jeremyw

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History