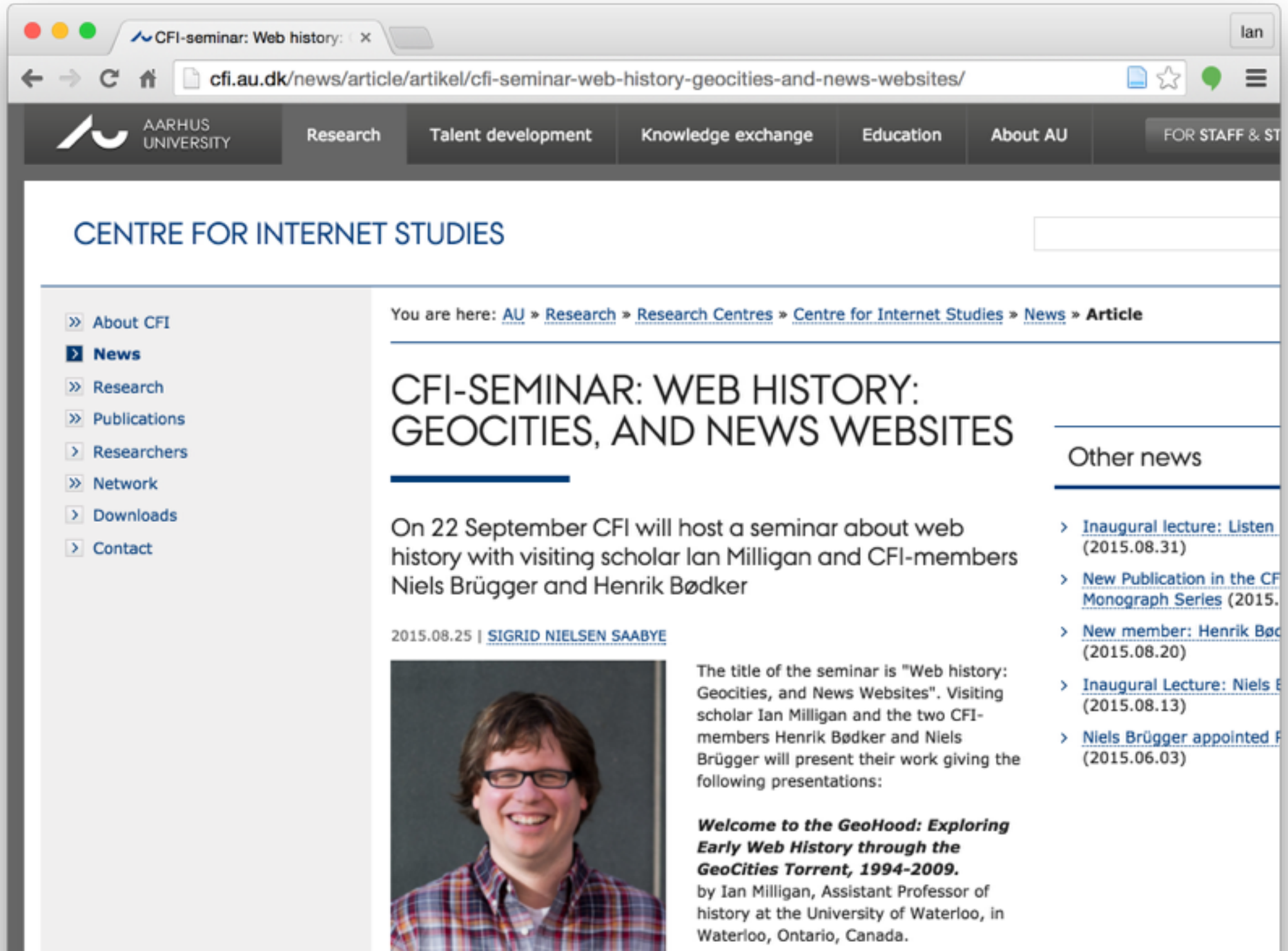
The Record of our Recent Past

Enabling Large-Scale Web Archive Data Mining

Ian Milligan
Assistant Professor

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Later Today


CFI-seminar: Web history:
Ian
cfi.au.dk/news/article/artikel/cfi-seminar-web-history-geocities-and-news-websites/
AARHUS UNIVERSITY
Research
Talent development
Knowledge exchange
Education
About AU
FOR STAFF & ST
CENTRE FOR INTERNET STUDIES
About CFI
News
Research
Publications
Researchers
Network
Downloads
Contact
You are here: AU » Research » Research Centres » Centre for Internet Studies » News » Article
CFI-SEMINAR: WEB HISTORY: GEOCITIES, AND NEWS WEBSITES
On 22 September CFI will host a seminar about web history with visiting scholar Ian Milligan and CFI-members Niels Brügger and Henrik Bødker
2015.08.25 | SIGRID NIELSEN SAABYE
The title of the seminar is "Web history: Geocities, and News Websites". Visiting scholar Ian Milligan and the two CFI-members Henrik Bødker and Niels Brügger will present their work giving the following presentations:
Welcome to the GeoHood: Exploring Early Web History through the GeoCities Torrent, 1994-2009.
by Ian Milligan, Assistant Professor of history at the University of Waterloo, in Waterloo, Ontario, Canada.
Other news
Inaugural lecture: Listen (2015.08.31)
New Publication in the CF Monograph Series (2015.
New member: Henrik Bøc (2015.08.20)
Inaugural Lecture: Niels E (2015.08.13)
Niels Brügger appointed F (2015.06.03)

Some links/resources collected at:

https://github.com/ianmilligan1/Aarhus-Netlab

Why?

The sheer amount of social, cultural, and political information generated every day presents new opportunities for historians.

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

No.

Historians need to do this now, or we're going to be left behind.

Nightmare Scenario

- Wayback Machine won't be enough. We won't use that.
- Historians rely uncritically on **date-ordered keyword search results**, putting them at mercy of search algorithms they do not understand;
- Historians are completely left out of post-1996 research, letting everybody else do the work (a la Culturomics project/*Nature* magazine article);
- Our profession gets left behind…

But what will web archives look like?

- Three Distinct Case Studies

  - **Wide Web Scrape**, March - December 2011 (Internet Archive) (sample of 80TB WARC collection);

  - **GeoCities End-of-Life Torrent**, 2009 (Archive Team);

  - **Archive-It Longitudinal Collections, Canadian Political Parties & Labour Organizations**, 2005-2014 (Archive-It/University of Toronto)

Example Dataset

- Archive-It Collection 227, **Canadian Political Parties and Political Interest Groups (University of Toronto)**

- October 2005 - Present
  - All major and minor political parties, as well as organized political interest groups (Council of Canadians, Coalition to Oppose the Arms Trade Assembly of First Nations, etc.)
- Started by now-retired librarian, hard to get details on seed list

Pivotal Changes in Canadian Politics, 2005-2015

- Militarization of Canadian society?

- Change from ‘natural governing party’ of Liberals to Conservatives

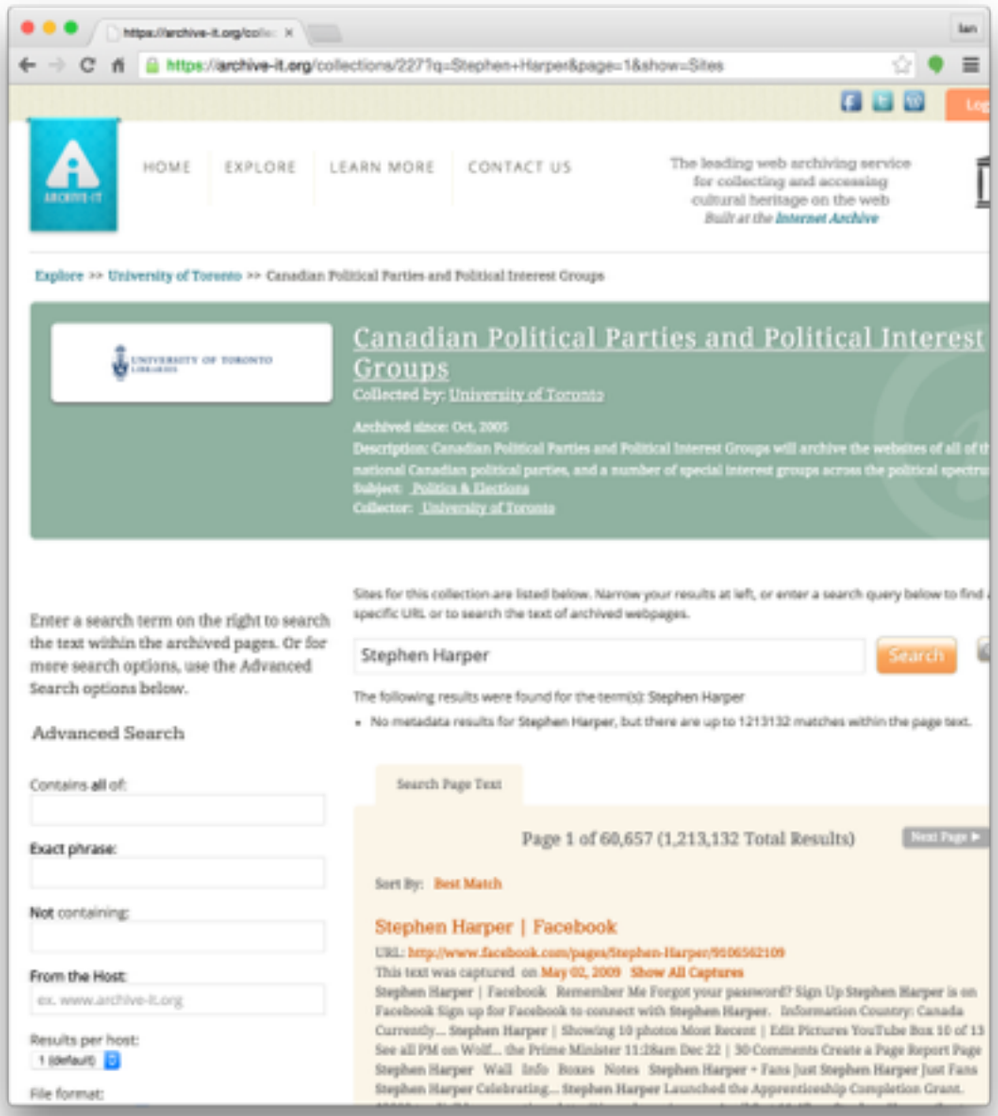
- Major policy changes on foreign policy, environment, etc.

- How to measure?

Current Interface

- Very limited - simple search engine, some advanced options; no facets

- Great collections.. but nobody uses them!



How to provide access?

Two Main Approaches

- Warcbase
  - Link extraction and analytics
  - Full-text extraction and analytics
- Full-text faceted search
  - UK Web Archive's **Shine** solr front end

WAT files

vs.

ARC/WARC files

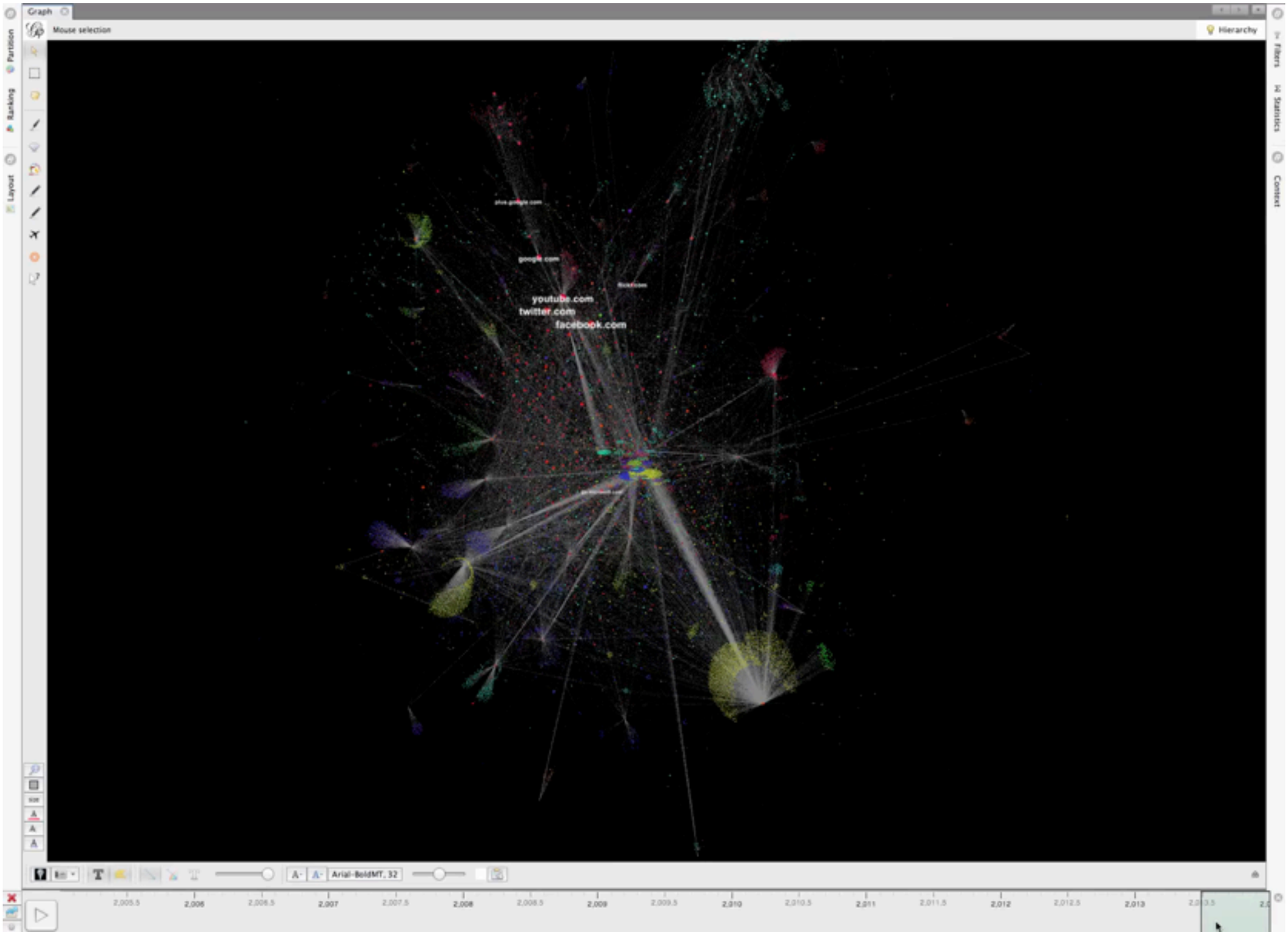
Problem One:

Historians want content, but we can only locally work with metadata

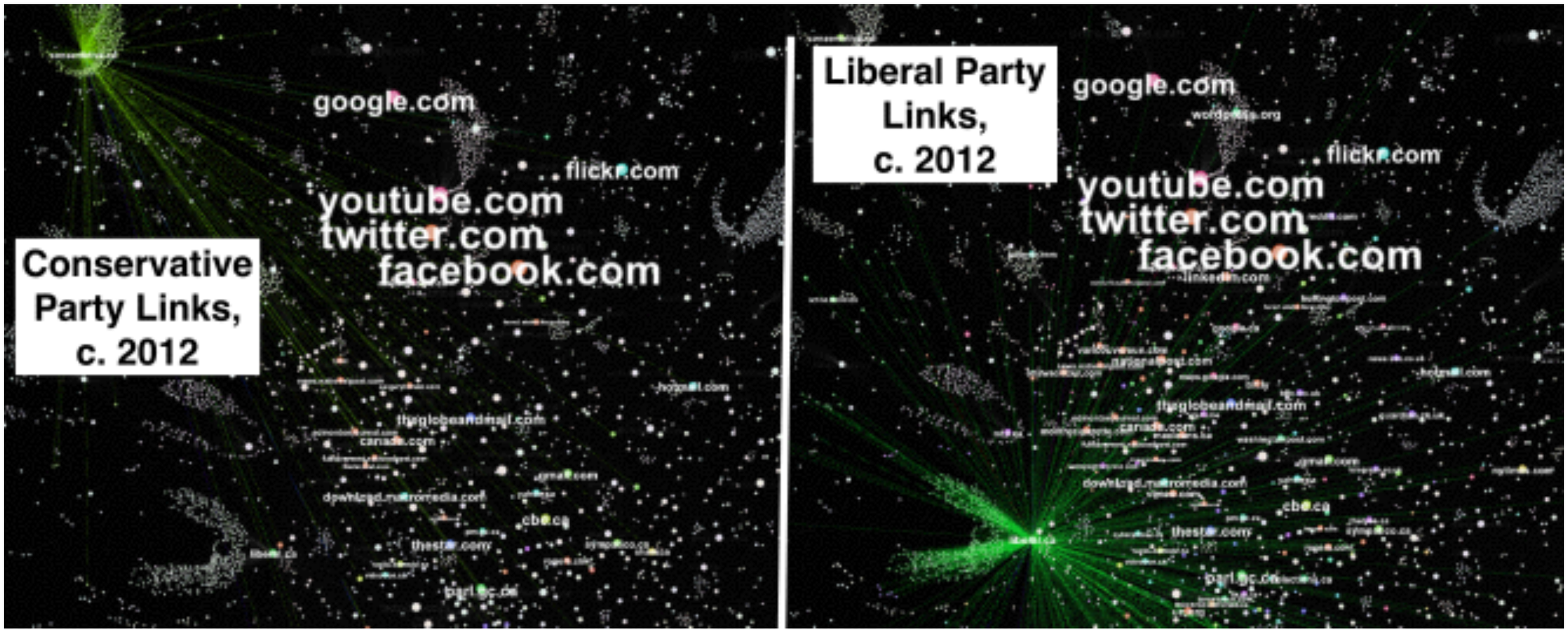
Do we want **metadata** or **content analysis**?

Historians NEED content,
but metadata can help us
find and **contextualize** it

Metadata Extraction



Metadata Extraction



Metadata Extraction

liberal.ca	27
liberal.ola.org	27
liberal.us1.list-manage.com	27
liberal.us1.list-manage1.com	27
liberal.us1.list-manage2.com	27
liberaluniversity.liberal.ca	27
license.icopyright.net	27
live.cbc.ca	27
lpc.ca	27
macleans.ca	27
masses.tao.ca	27
mcss.gov.on.ca	27
mediaignite.com	27
mediasales.cbc.ca	27
membercentre.cbc.ca	27
mentalhealthcommission.ca	27
metrics.mmailhost.com	27
mondesdesfemmes.ca	27
music.cbc.ca	27
nawl.ca	27
newswire.ca	27
nowtoronto.com	27
npd.ca	27

colincarriemp.ca	12
colincarriemp.ca&lang=fr	12
colinmayes.ca	12
colinmayes.ca&lang=fr	12
congrespcc.ca	12
conservateur.ca	12
conservateur.us5.list-manage.com	12
conservative.ca	12
conservative.us5.list-manage.com	12
consumersfirst.ca	12
corneliuchisu.ca	12
corneliuchisu.ca&lang=fr	12
costasmenegakis.ca	12
costasmenegakis.ca&lang=fr	12
cpcconvention.ca	12

Metadata Extraction

- Results @ http://ianmilligan.ca/2015/02/05/topic-modeling-web-archive-modularity-classes/

Metadata Extraction

- Conservative themes (2014): economic development, family, immigration, legislation, women's issues, senior issues, Ukrainians, constituency offices, some prominent (and not-so-prominent) MPs, and of course, our economic action plan.

- Liberal themes (2014): Justin Trudeau (the new leader), cuts to social programs, child poverty, mental health, municipal issues, labour, workers, Stop the Cuts, and housing.

Metadata Extraction

- Conservative themes (2006): education, university, but **tons of information on Aboriginal issues**;

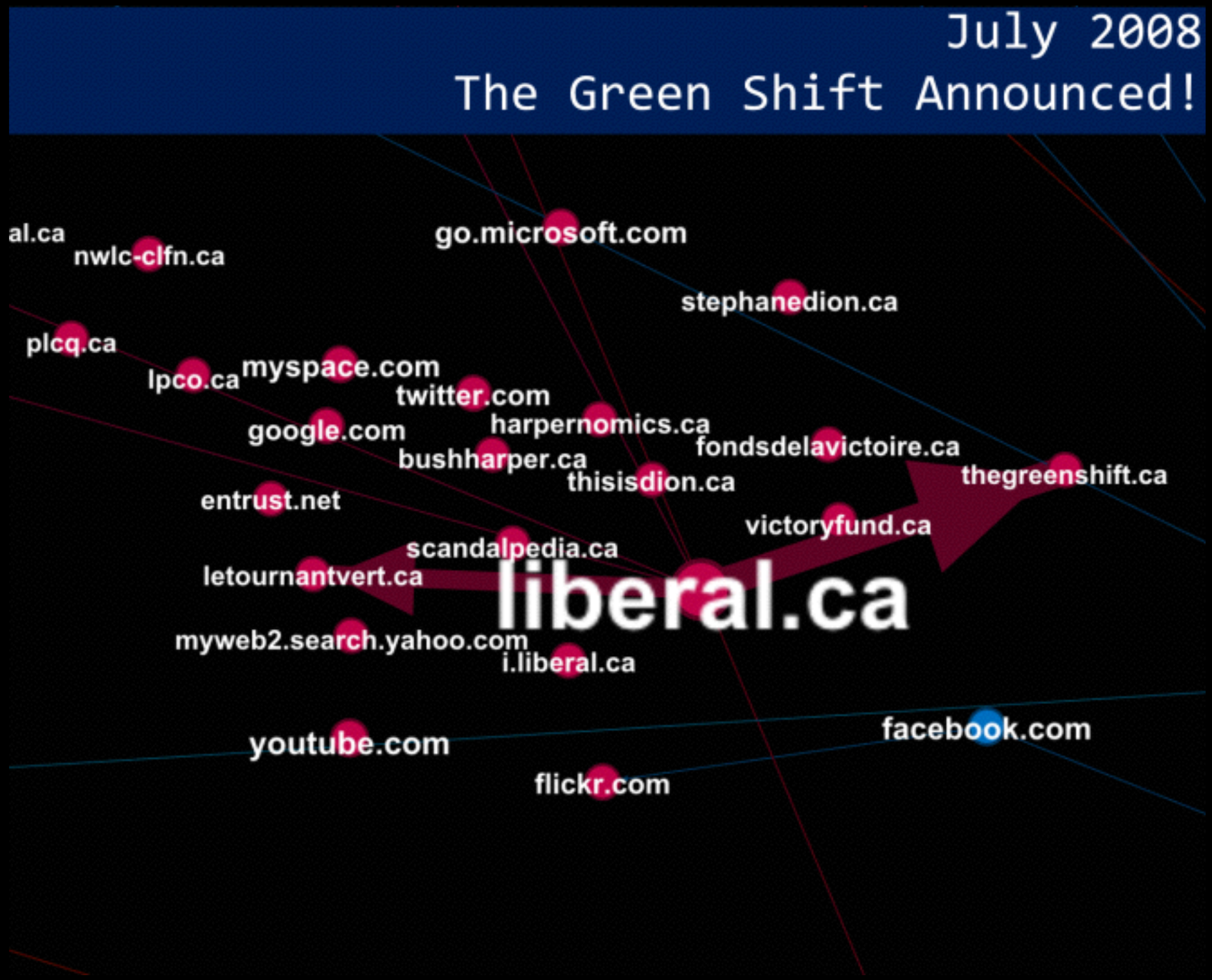
- Liberal themes (2006): community questions, electoral topics, universities, human rights, child care support.

As well as short stories...

December 2006
Stephane Dion Elected Leader of Party

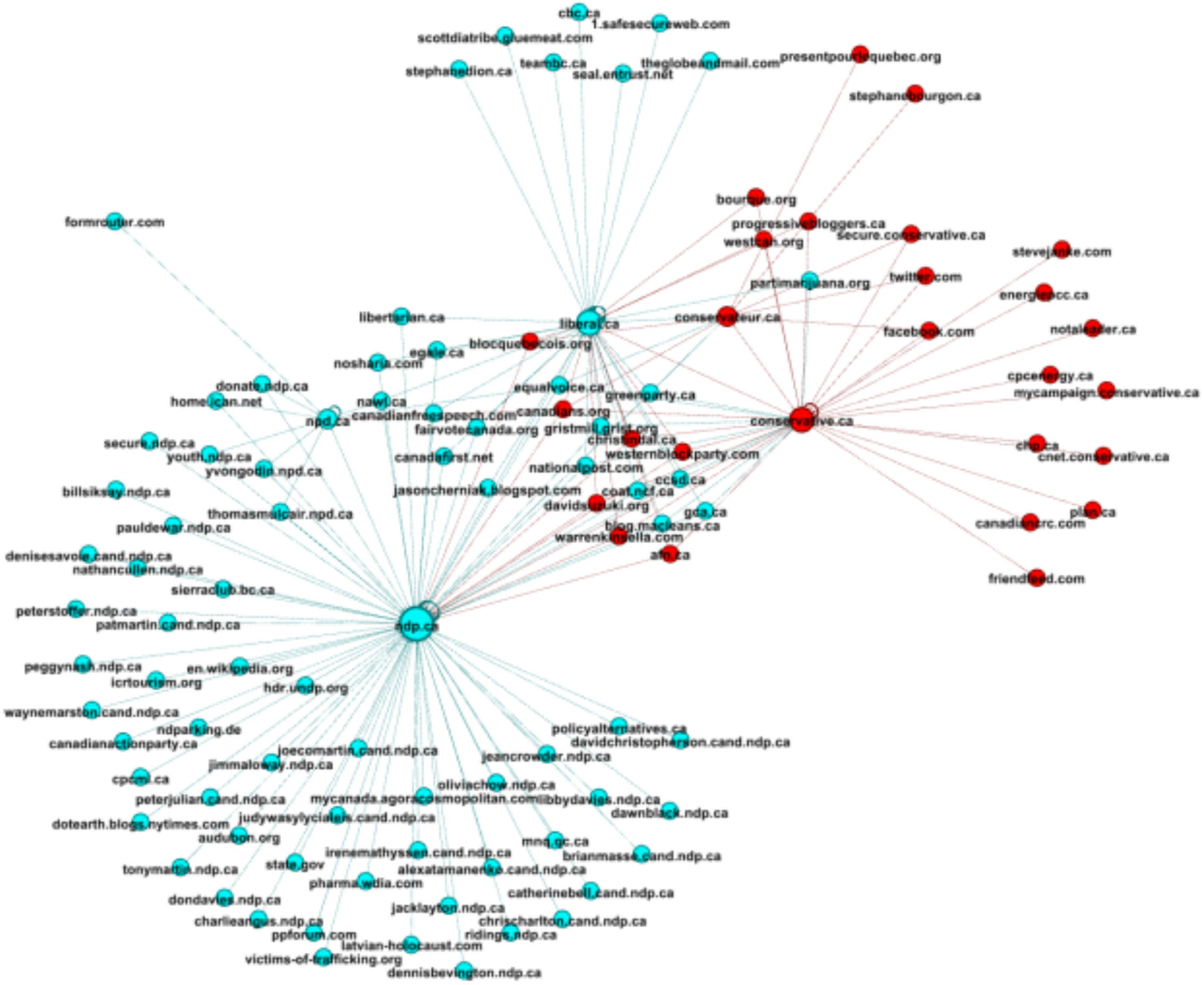
December 2007
Rise of Social Media

April 2008
Fundraising with the Victory Fund/ Fonds de la Victoire

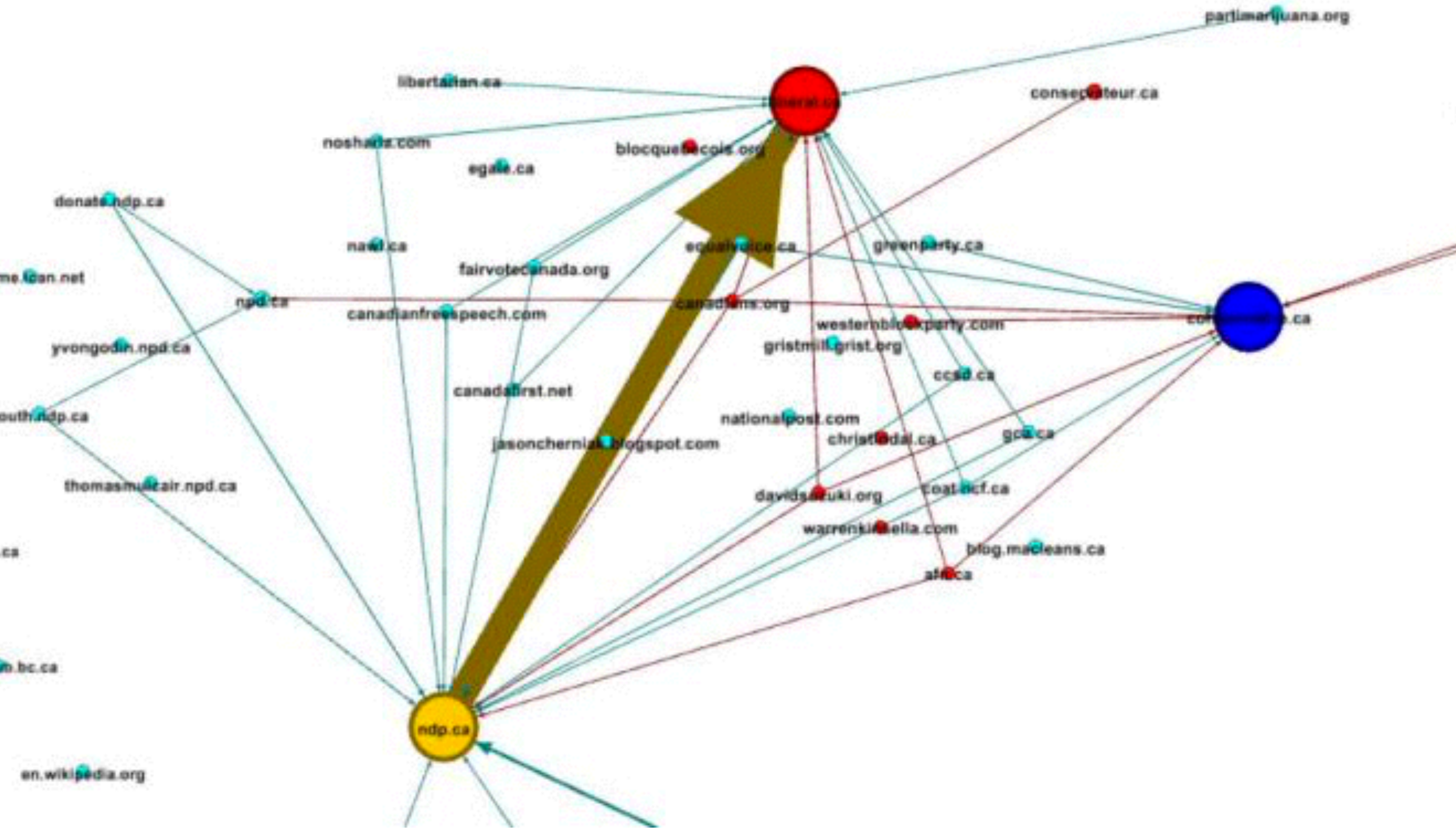

July 2008
The Green Shift Announced!
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com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

October 2008
Election Campaign - Advertisement Sites

December 2008
Election campaign Ends; Attacking Harper on Anti-American Grounds (bushharper)


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
westcan.org
secure.conservative.ca
formrouter.com
partimarijuana.org
twitter.com
stevejanke.com
energyprcc.ca
libertarian.ca
liberal.ca
conservateur.ca
facebook.com
notaleader.ca
egale.ca
blocquebecois.org
noshame.com
donate.ndp.ca
equalvoice.ca
greenparty.ca
cpcenergy.ca
homelcan.net
nawl.ca
canadians.org
mycampaign.conservative.ca
canadianfreespeech.com
npd.ca
fairvotecanada.org
gristmill.grist.org
conservative.ca
christendial.ca
secure.ndp.ca
youth.ndp.ca
canadafirst.net
westernblockparty.com
cbp.ca
cnet.conservative.ca
yvongodin.npd.ca
nationalpost.com
ccsd.ca
billsiksay.ndp.ca
jasoncherniak.blogspot.com
coat.ncf.ca
davidsuzuki.org
gda.ca
plan.ca
thomasmulcair.npd.ca
blog.macleans.ca
canadiancrc.com
pauldewar.ndp.ca
warrenkinsella.com
afn.ca
denisesavoie.cand.ndp.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
ndparking.de
policyalternatives.ca
canadianactionparty.ca
davidchristopherson.cand.ndp.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
judywasylycia-leis.cand.ndp.ca
dotearth.blogs.nytimes.com
audubon.org
mnq.qc.ca
irenemathyssen.cand.ndp.ca
brianmasse.cand.ndp.ca
tonymartin.ndp.ca
state.gov
alexatamanenko.cand.ndp.ca
pharmawatch.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

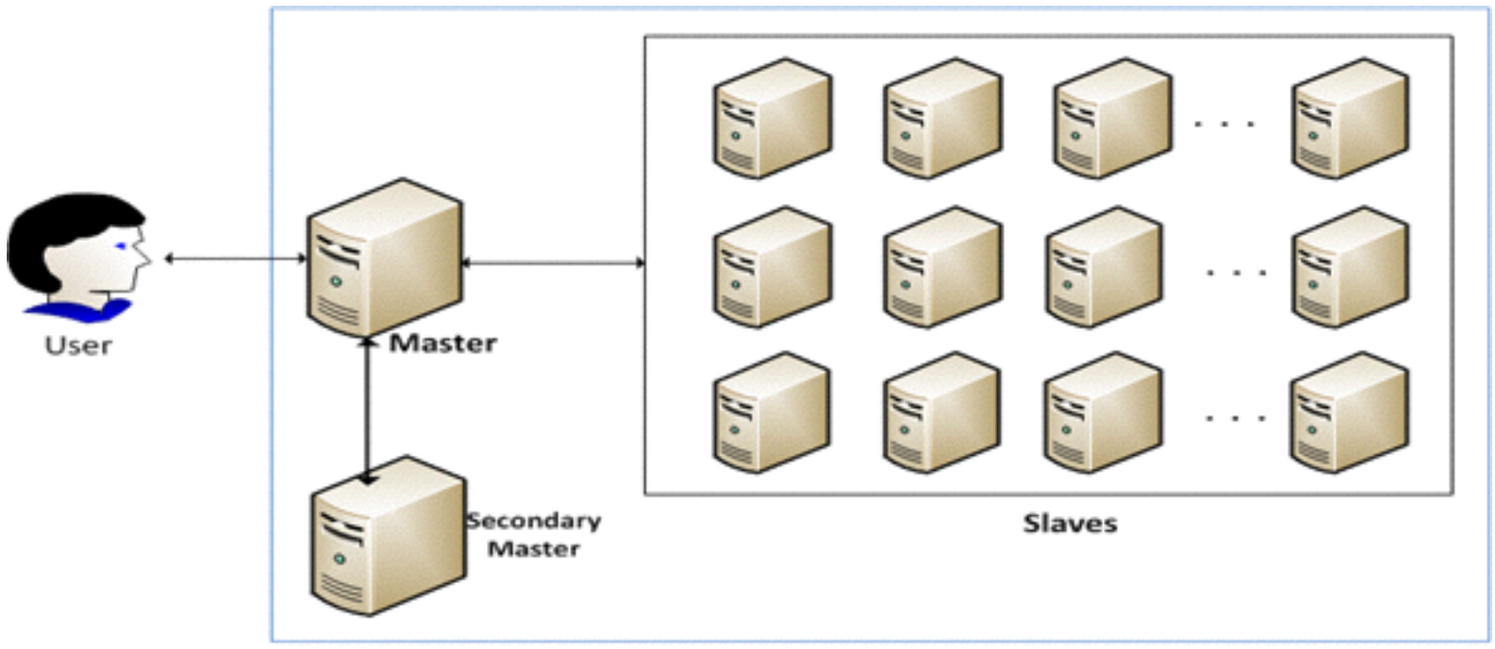
2005 Canadian Federal Election



WATs help us find the files we need to use - and to **contextualize them**

Problem Two:

You can do amazing things with the content (WARCs), but you need a cluster or powerful computer.



Hadoop Cluster

WARC Analysis

- **2005-2009**: 244 GB of content; 2.9 GB of plain text

- 10,606,822 websites

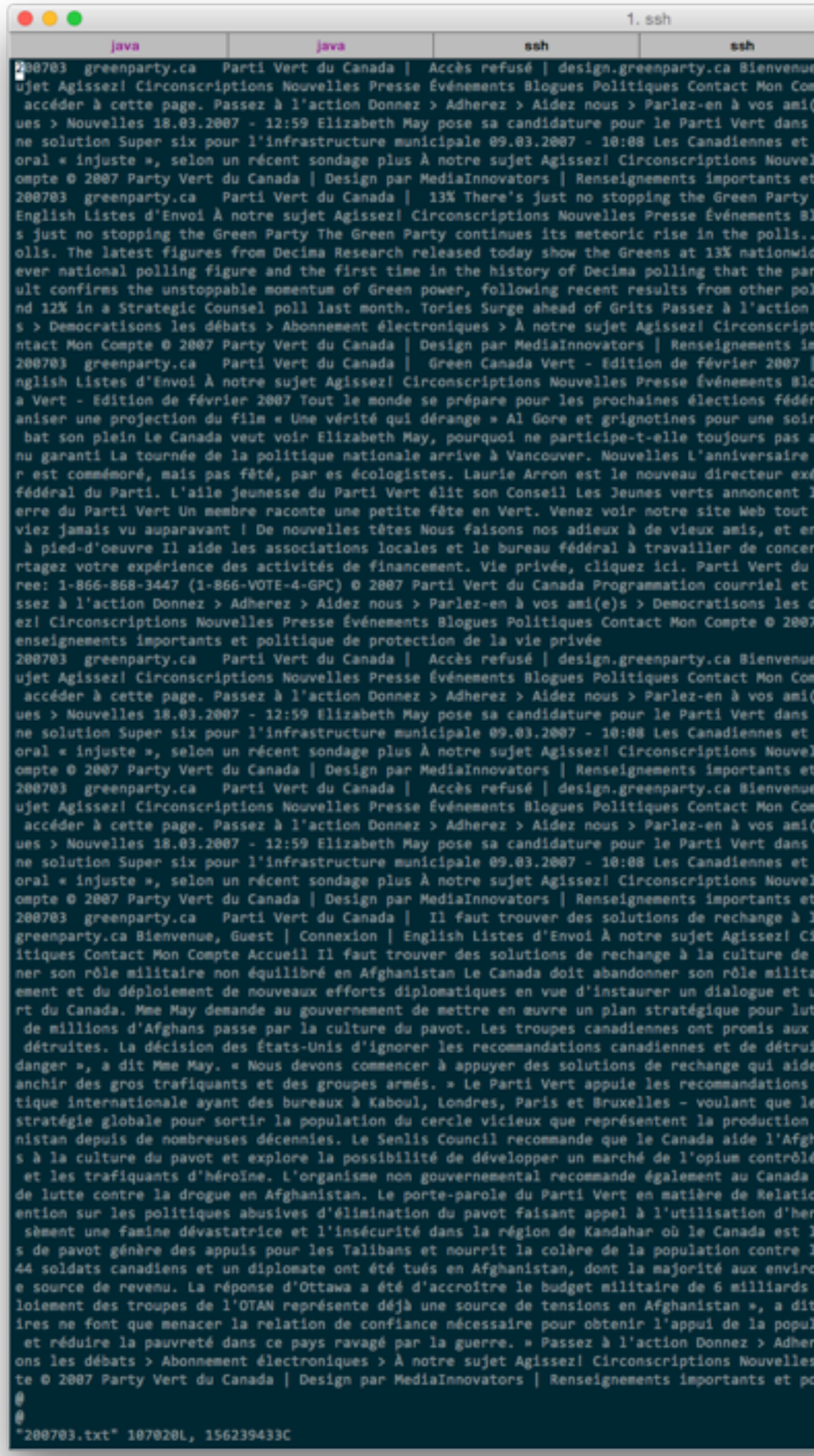
- On a local powerful node (3 Ghz 8-Core Intel Xeon E5/64 GB RAM, data on SSD), about three to four hours per query

- On a cluster, about ~10-20 minutes per query, depending on traffic

Large-Scale Text Analysis

- With Hadoop about 15-20 minutes to extract all plain-text from any specified queries: i.e. all pages belonging to Green Party, Liberal Party, Conservative Party, Council of Canadians, etc.

- Compared to “out of memory”/ go home for an extended weekend on a local node



Large-Scale Text Analysis

- **NER/LDA/Keyword Frequency broken down by scrape date**: i.e. scrape carried out 2005-10, see change over time;

- Downside: not everything is optimized for parallel environment; if not, it crawls (there goes a day)

- Downside: scrape date != creation date, requiring temporal analysis

Recipe Book Idea

Recipe book:

https://github.com/lintool/warcbase/wiki

NER

October 2005

62476 Stephen Harper

30234 Michael Chong

30109 Gwynne Dyer

28011 ami Entrez

26238 Paul Martin

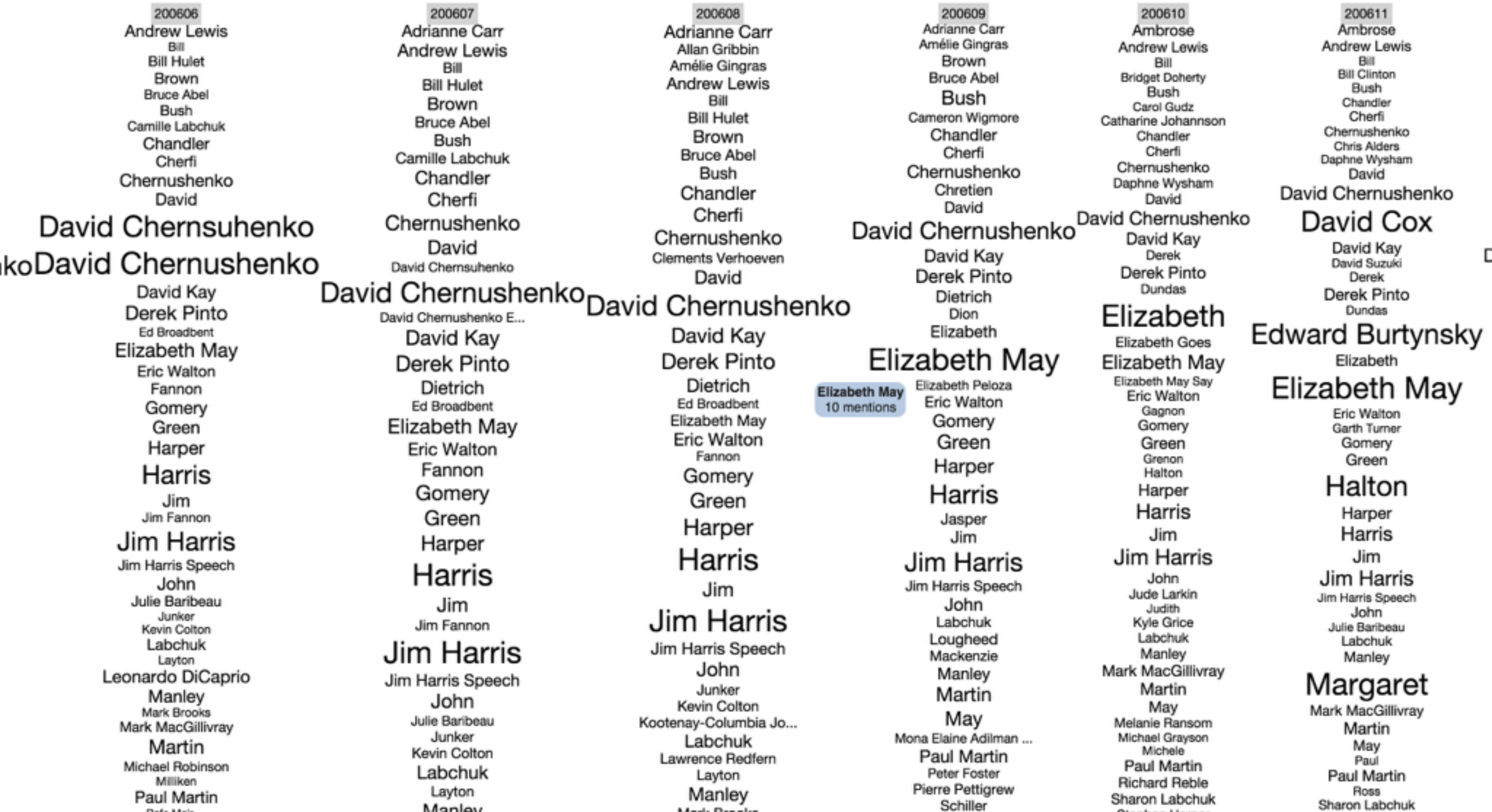
22303 Harper

NER

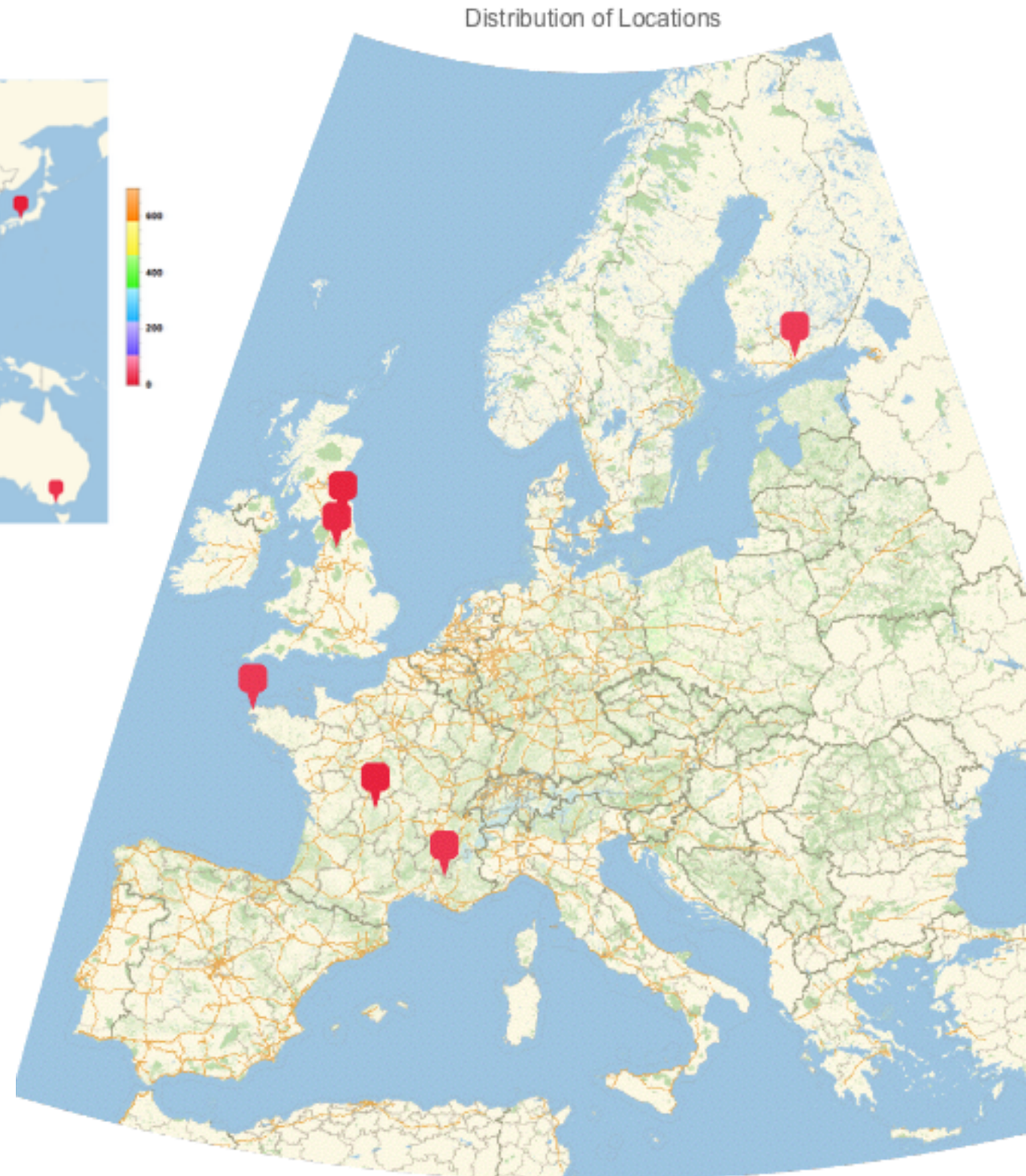
November 2008

3188 Stéphane Dion
2557 Stephen Harper
2471 Stephen HarperLaureen
2410 Dion
2356 Harper

Visualizing Interface Next Step?



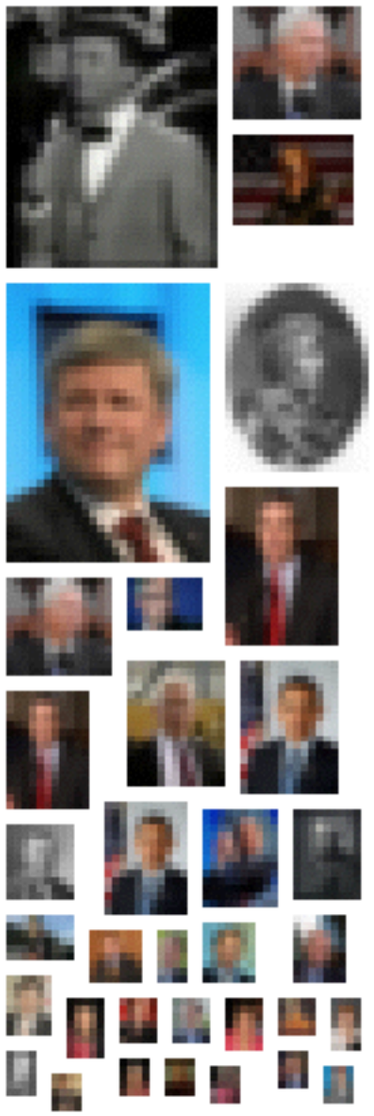
Integration with Wolfram| Alpha

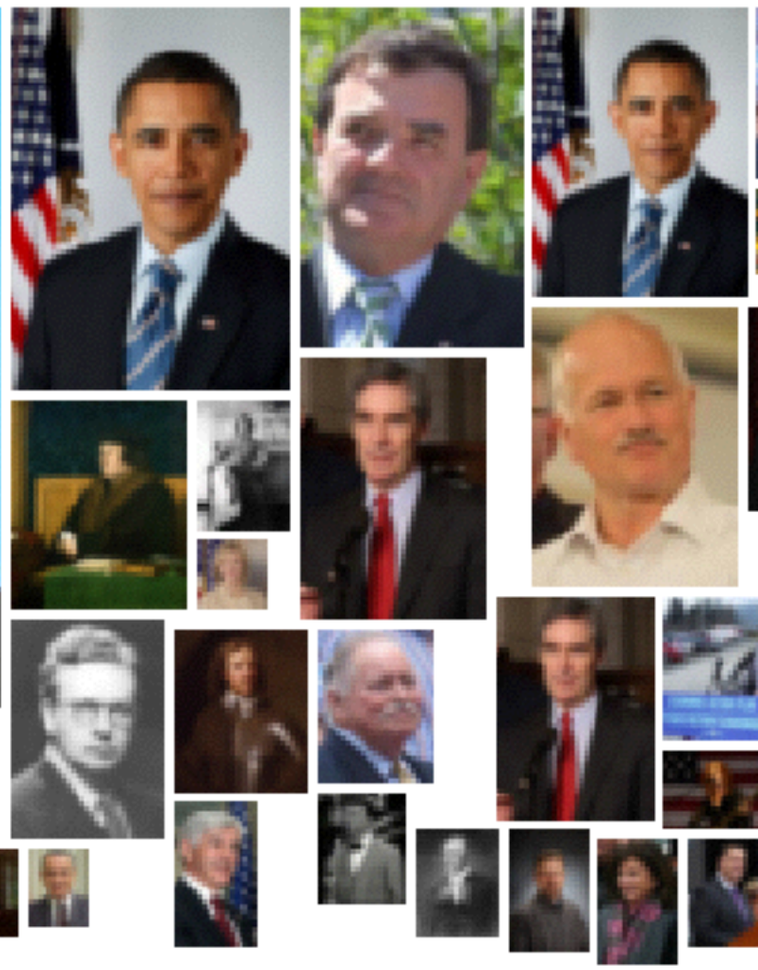



```
In[95]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]

Out[95]= {Canada, Calgary, Colombia, Montreal, $Failed, Ontario, Canada, Afghanistan,
  Ottawa, Manitoba, Canada, British Columbia, Canada, Toronto, Nova Scotia, Canada,
  Saskatchewan, Canada, Quebec, Canada, Alberta, Canada, Winnipeg, Washington,
  Canada, Nunavut, Canada, White Horse, United States, Petawawa, Mumbai,
  Lima, The Americas, Saskatoon, Peru, Edmonton, Mexico, Chicoutimi-Jonquiere,
  New Brunswick, Canada, United States, Newfoundland and Labrador, Canada, Qandahar,
  $Failed, Asia-Pacific, Newfoundland and Labrador, Canada, Victoria, United States,
  Quebec City, India, $Failed, Atlantic Ocean, St. Germain, Durham, Battle of Waterloo,
```

Integration with Wolfram| Alpha





Shine/WebArchives.ca

- UK Web Archive's Shine (https://github.com/ukwa/shine)

- Indexing as bottleneck
  - ~ 250GB of WARCs takes ~ 5 days on a single machine
  - Hadoop indexer available if data in HFDS
- ~ 90GB index size

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

Shine

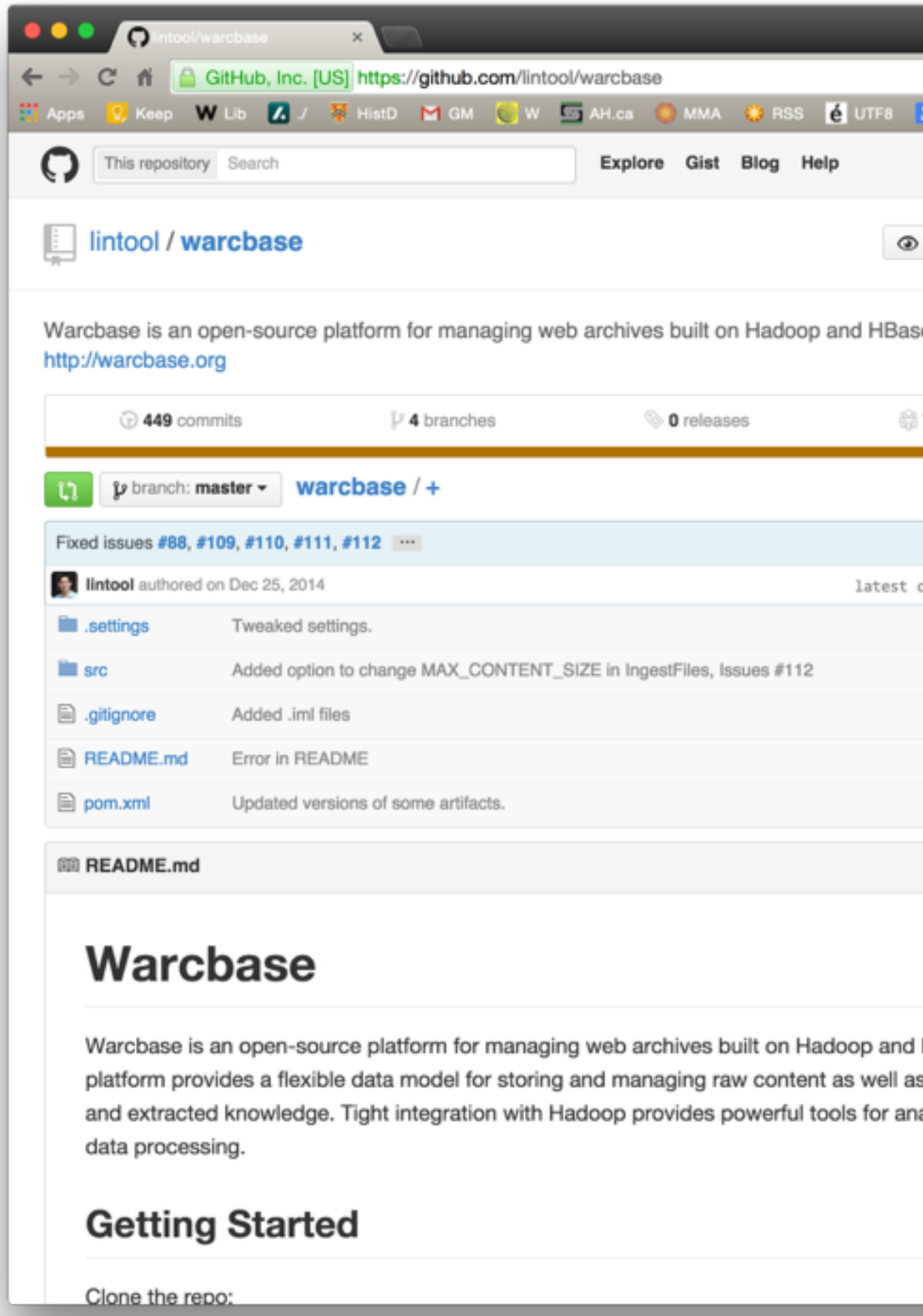
- **Advantages**: accessible to the general public, easy to use, interactive trend diagram allows digging down for context, can move down to level of document itself.

- **Disadvantage**: keyword searching requires you know what to look for; random sampling misleading when tens of thousands of records; etc.

- Doesn’t take advantage of what makes web sources so powerful: hyperlinks

Building connections between Warcbase and Shine

End-user tools and co-operation with CS colleagues is key.



But the shared promise…

More voices, more people, the promise of social history achieved.

Thank you!

@ianmilligan1
ianmilligan1@gmail.com

Ian Milligan
Assistant Professor

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History