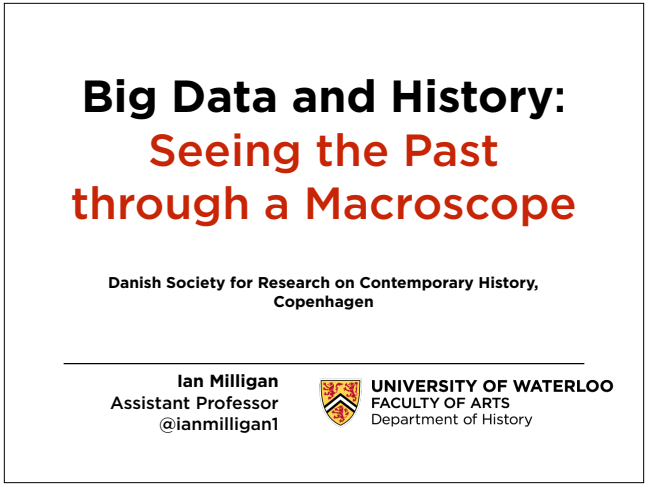


Hi everybody and thanks so much for having me here today - to Professor Sorensen for the gracious invitation, and to the Danish Society for Research on Contemporary History for making my attendance here possible.

What I want to do here today is talk about why I think Big Data matters, and walk you through several major points:

(1) That Big Data is already reshaping our profession - and it is something that will continue to accelerate thanks to the advent of web archives, the area I work on with many others including Professor Niels Brugger;
(2) That the shift from historical scarcity to abundance is a fundamental shift;
(3) And that we need new tools, the Macroscope, to access this information. I'll give examples from my work, with links and notes towards different tools and walkthroughs that I've used to access it.

And then we will have time for questions!

Historians are largely unprepared to engage with the quantity of digital sources that will fundamentally transform their trade.

So let's begin with my thesis:

Because, as I've explained in other venues as well as here, I believe that historians are unprepared to engage with the quantity of digital sources that will fundamentally transform their trade. The advent of the World Wide Web, both with born-digital sources that can shed so much light on human culture and activity after the advent of web archived material in approximately 1996, as well as digitized traces of the past, presents a challenge to us as professionals.

… we need to think about data …

It means that we need to start thinking about data. How to preserve it, how to access it, how to interpret it, and how to facilitate its reuse.

Today's Talk

- **1. Prologue:** Big Data is everywhere
- **2. The Web Age:** Will accelerate this process
- **3. What can we do with big data?**

So what I want to do today, in the time we have together, is to explore a few different dimensions of historical research and how data is transforming it. Let's start with the big point about why thinking about data matters for us, go into the particulars of the Big Data of the Web Age - a plug for my own research - and then conclude around some of the problems that I think we're going to have to tackle.

A Prologue: Big Data is Everywhere

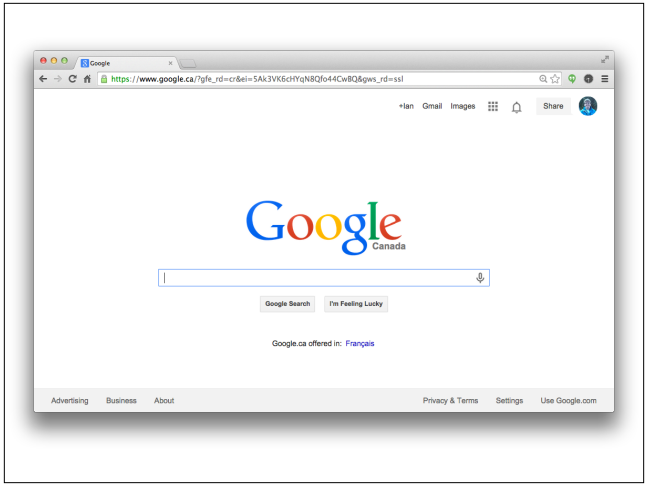
Let's start with a prequel to the Web Age.

What do we mean by Big Data?

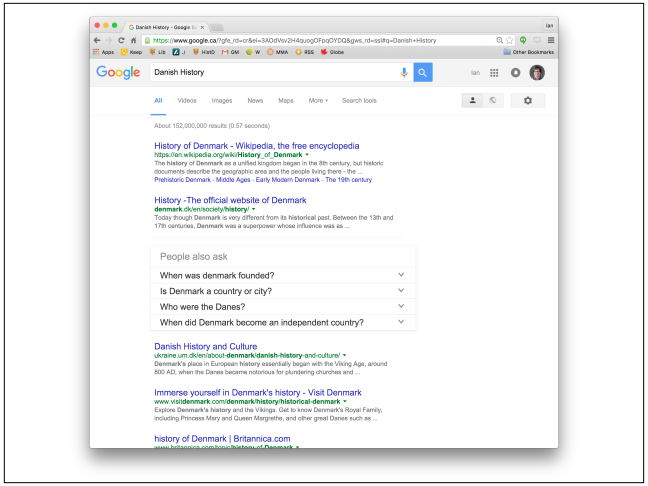
- Computational definition: the 5 Vs (Volume, Velocity, Variety, Veracity, and Value)
- "For us, as humanists, big is in the eye of the beholder. **If it's more data that you could conceivably read yourself in a reasonable amount of time, or that requires computational intervention to make new sense of it, it's big enough**!" (Shawn Graham, Ian Milligan, Scott Weingart, *Exploring Big Historical Data*)

two kinds of definitions

Why is it everywhere?

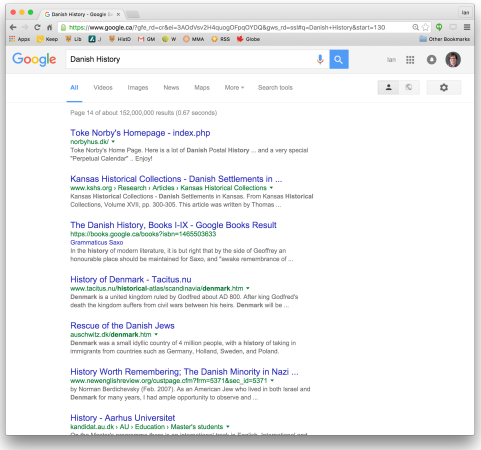


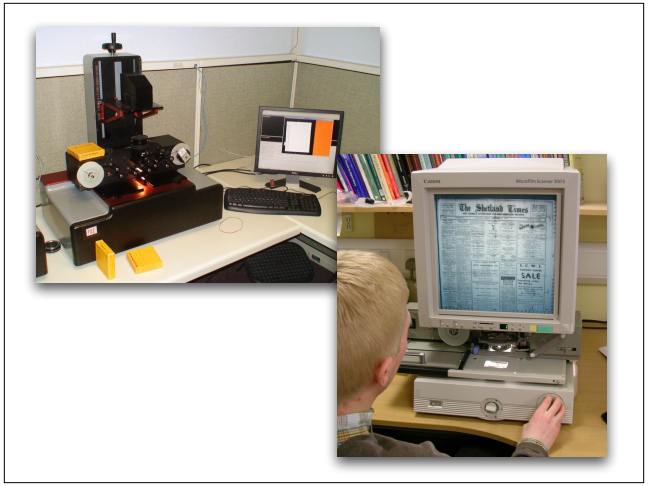
The first point I want to make is that no matter the kind of research historians are doing, chances are they're using data - and at the mercy of those who interpret it for them. Every time you run a search result…



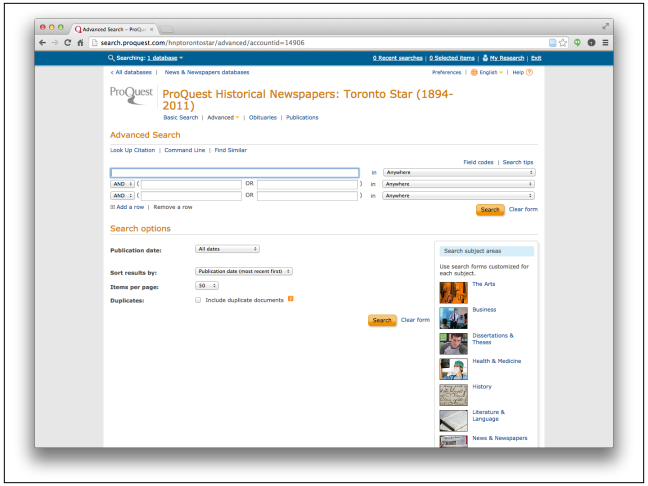
There's a reason that the wikipedia article is at the top of this search on Google Canada for Danish History, that the 100th result is a home page of a Danish researcher, and that beyond that we have resources that we never know about - who here actually visits the 1000th most popular site on a Google search?

Most humanists probably don't think about it, but Google's skewing their research. ITHAKA S+R did a survey, and at least in North America, this is our new **research assistant**.


Danish History - Google Se
https://www.google.ca/?gfe_rd=cr&ei=3AOdVovZH4quogOFpqOYDQ&gws_rd=ssl#q=Danish+History&start=130
Apps Keep LIb HistD GM W MMA RSS Globe Other Bookmarks
Google
Danish History
Ian
All Videos Images News Maps More Search tools
Page 14 of about 152,000,000 results (0.67 seconds)
Toke Norby's Homepage - index.php
norbyhus.dk/
Toke Norby's Home Page. Here is a lot of Danish Postal History ... and a very special "Perpetual Calendar" .. Enjoy!
Kansas Historical Collections - Danish Settlements in ...
www.kshs.org › Research › Articles › Kansas Historical Collections
Kansas Historical Collections - Danish Settlements in Kansas. From Kansas Historical Collections, Volume XVII, pp. 300-305. This article was written by Thomas ...
The Danish History, Books I-IX - Google Books Result
https://books.google.ca/books?isbn=1465503633
Grammaticus Saxo
In the history of modern literature, it is but right that by the side of Geoffrey an honourable place should be maintained for Saxo, and "awake remembrance of ...
History of Denmark - Tacitus.nu
www.tacitus.nu/historical-atlas/scandinavia/denmark.htm
Denmark is a united kingdom ruled by Godfred about AD 800. After king Godfred's death the kingdom suffers from civil wars between his heirs. Denmark will be ...
Rescue of the Danish Jews
auschwitz.dk/denmark.htm
Denmark was a small idyllic country of 4 million people, with a history of taking in immigrants from countries such as Germany, Holland, Sweden, and Poland.
History Worth Remembering; The Danish Minority in Nazi ...
www.newenglishreview.org/custpage.cfm?frm=5371&sec_id=5371
by Norman Berdichevsky (Feb. 2007). As an American Jew who lived in both Israel and Denmark for many years, I had ample opportunity to observe and ...
History - Aarhus Universitet
kandidat.au.dk › AU › Education › Master's students

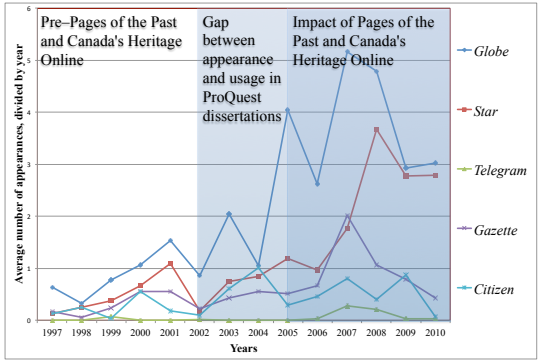


But in the Web age, historians are accessing most of their sources through mediated search portals, which necessitates computational thinking. For decades, historians have used microfilms, and now in literally the last decade, we've been using Web-based primary sources.



On a whim, I was wondering what Web-based primary sources might mean - mainly because I was reading somebody's dissertation, they were talking about events in Oshawa, Ontario, and they were using the Toronto Star - not the Oshawa Daily Times. Or, to be honest, the first time I'd started researching a strike in Hamilton, I started using the Globe and Mail until my supervisor pointed out that the Hamilton Spectator would be better.

Well, luckily I wasn't alone.



Ian Milligan, "Online Databases, Optical Character Recognition, and Canadian History, 1997–2010," *Canadian Historical Review*, 94.4 (December 2013): 540-569.

As things are digitized, people use them more. And the things that aren't digitized, are used less. That's fine, except:

Digitized sources don't reflect the geographic diversity of Canada or the GLOBAL NORTH - i.e.

* Hamilton not digitized, but Toronto is. That makes decisions made in 2001-era Canada impacting our histories of 1890s Canada.
* The OCR was generated in 2001 (!) based on digitized microfilms
* The original OCR is inaccessible to the creators
* The search engine is dreadful
* Our citation methods don't differentiate generally between whether we got it in a collection file, a microfilm, an original copy, or an online database - and all these things are different

In short, **digital sources are mediated in different ways than the original sources**, and arguably for a decade Canadian historians were not recognizing the role that algorithms and data play in the mediation of our research.

Digital resources in the Web age are useful, but we're a profession not terribly equipped to deal with this.

... this is our long-term track record w/ digital resources ...

I mention this by way of introduction because it sort of sets out the track record that historians have when we're dealing with this material. **Digital resources came, and we uncritically used them for years**, not just in Canada but across the historical profession. I've made this point in a Canadian Historical Review article and it's not immodest of me to say that it's appearing on syllabi and being discussed in France, Australia, and the United States because it's a relatively novel point for historians.

Our history with digital sources is the **unreflective use of technology**.

… we've become, in some ways, a discipline defined by the keyword …

We've become a keyword-based discipline almost overnight, from finding aids to digitized newspapers.

A process that is only now beginning to accelerate.

And this becomes incredibly pressing, because soon historians are going to have a TON of data to draw on. This brings me into the main topic of my presentation: the web age. If we can’t handle newspapers critically once they're digitized and presented as DATA, how can we handle the 1990s?

First - more data than ever before being preserved;

Second - it'll be saved/delivered to us in very different ways

Since the 1890s, this has been the bread-and-butter of the historical profession. An archival box. Good, old fashioned, 'analog' to use an imprecise word, technology. Historians go to archives, we flip these open, we see a bunch of documents, and from these traces of the past we attempt to reconstruct narratives that we call history.

While historians have challenged some of the preconceptions of this since the 1960s, the old gold standard was government documents. Even since the 1960s, we largely - not wholly, as oral historians and others would testify - use archives deposited with organizations: Library and Archives Canada, the University of Waterloo's Special Collections, etc.

This all means that historians are working with scant traces of the past. 99.9% of things that happen are never recorded (something which philosophically really isn't changing with new sources). Events are transitory, they happen, and unless it ends up in an archive box, or a newspaper, they are generally lost.

The problem is, with these, in a word, that of scarcity, as the late great American historian Roy Rosensweig noted.

So if that has been the landscape for well over a hundred years of professional development - - I want to hold to you today that the arrival of something new portends to change and alter the work that we do. And it is this:

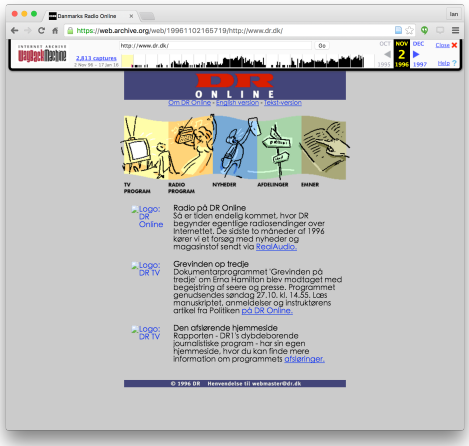
WebARChive (WARC) File

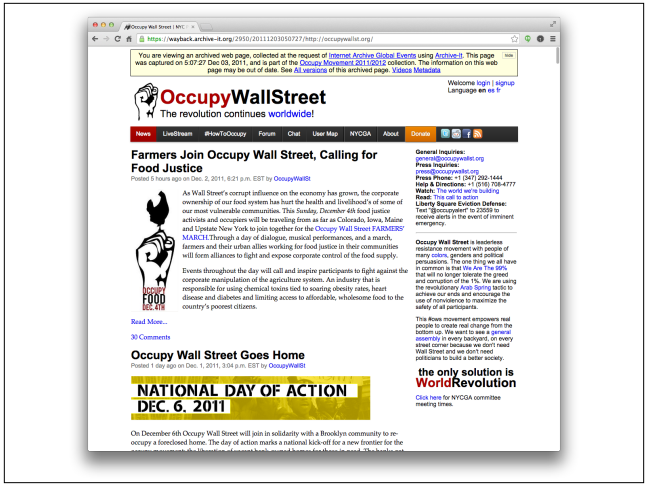
The WARC file, or WebARChive file.

This, I argue, is what historians need to wrap their heads around. Because it represents the biggest challenge to our profession since that old man in the last slide first tried to professionalize our discipline over a hundred years ago.

So what is a WARC file? It's an International Standard, ISO 28500:2009 for those keeping track, and has become the international standard to preserve content and information "from mainstream Internet application layer protocols, such as the Hypertext Transfer Protocol (HTTP), Domain Name System (DNS), and File Transfer Protocol (FTP)" with attached metadata.

What does this really mean? In a nutshell, I can take the entirety of the University of Copenhagen's website today, and pop it all into a WARC file. It preserves the website as it is today, and I can either access it through the WaybackMachine to experience it, or I can use a suite of text-based tools to extract meaningful information quickly.


Danmarks Radio Online
https://web.archive.org/web/19961102165719/http://www.dr.dk/
http://www.dr.dk/
2,813 captures
OCT
NOV
2
1996
DEC
Close
Help
DR
ONLINE
Om DR Online - english version - next version
TV PROGRAM
RADIO PROGRAM
NYHEDER
AFDELINGER
EMNER
Logo: DR Online
Radio på DR Online
Så er tiden endelig kommet, hvor DR begynder egentlige radiosendinger over internettet. De sidste to måneder af 1996 kører vi et forsøg med nyheder og magasinstof sendt via RealAudio.
Logo: DR TV
Grevinden op tredje
Dokumentarprogrammet 'Grevinden på tredje' om Erna Hamilton blev modtaget med begejstring af seere og presse. Programmet genudsendes søndag 27.10. kl. 14.55. Læs manuskriptet, anmeldelser og instruktørens artikel fra Politiken på DR Online.
Logo: DR TV
Den ofslørende hjemmeside
Rapporten - DR1's dybdeborende journalistiske program - har sin egen hjemmeside, hvor du kan finde mere information om programmets afsløringer.
© 1996 DR Henvendelse til webmaster@dr.dk



So what does this mean?

Imagine a future historian, taking on a central question of social and cultural life in the middle of the first decade of the twenty-first century, such as how Canadians understood Idle No More or the Occupy movement through social meia? What would her archive look like?

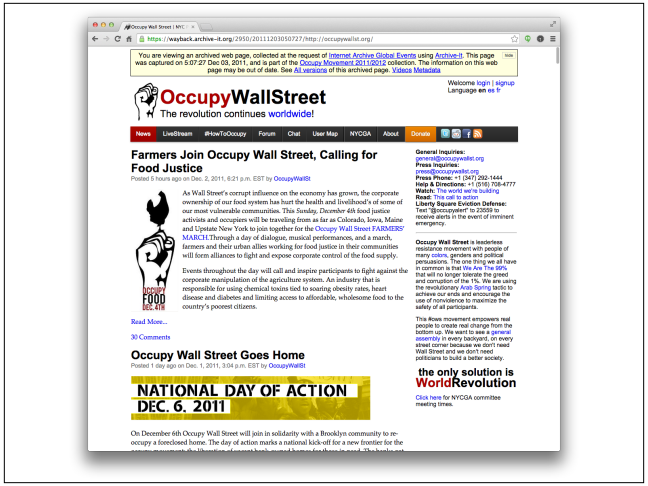
I like to imagine it as boxes stretching out into the distance, tapering off, without any immediate reference points to bring it into perspective.

Many of these digital archives will be without human-generated finding aids, would have perhaps no dividers or headings within “files,” and certainly no archivist with a comprehensive grasp of the collection contents.

Now this all means that historians need to get involved a bit earlier, of course. Of all those Occupy Wall Street sites - created in the heat of a movement, etc., they didn't have data management plans. And so two years later, only 41% of those sites are still active today.

I like to throw this in because it shows that we need to active maintain our digital sources, and historians need to begin thinking about retaining the data that we generate today.

If I wrote a book in 1991 and put it on a shelf in my basement, and came back in 2014, chances are I can read it. I'm damn sure that if I wrote a digital object, it would be unusable. We ned to be active.

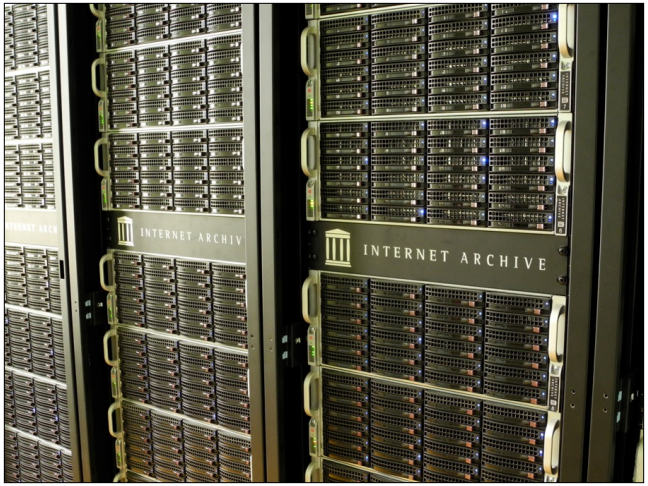


But I digress. Some of it will be retained, and it's going to be retained on a massive scale thanks to the efforts of institutions like the Internet Archive and some legal deposit institutions in Europe. Remember scarcity? **We're in the era of historical abundance**.

Take Twitter, for example. During the #IdleNoMore protests, there were an stoning 55,334 tweets on 11 January 2013. Each tweet can be up to 140 characters. Through a complicated bit of math that I whipped together, I argue that that's over 1,800 pages if we take 300 words per page. That's a MASSIVE book, and you've got one for every day of the big protest.
You can't read that yourselves, you're going to have to learn how to program. And that, my friends today, I hold to you is why historians need to be leading the charge up the digital humanities hill. But we're not.

Early theories of information saw the Library of Congress as the pinnacle, the largest quantity of information conceivable by information theorist Claude Shannon in 1949. It is, at least by some measures - the British Library disagrees - the largest repository of traditional print information in the world. Simply walking along its 838 miles, or 1,349 kilometres of shelving would take weeks – without even stopping to open up a single book.

The Library of Congress is no longer the pinnacle of information storage.



For that, we have the Internet Archive.

Now, comparing miles of shelves to the Internet Archive's 11,000 hard drives is an "apples versus oranges" issue, but it can be done as a rough thought experiment. Trying to put a firm data figure on an analog collection is difficult: a widely distributed figure is that the Library of Congress's print collection amounts to 10 terabytes. That is too low, and a more accurate figure is somewhere in the ballpark of 200 to 250 terabytes (if one digitized each book at 300 DPI, leading to a rough figure of eighty megabytes per book). As a petabyte is a thousand terabytes, if we take the latter figure we arrive at a 1:40 ratio.

1 (Library of Congress):
40 (Internet Archive)

The Internet Archive is massively bigger. The Wayback Machine continues to grow, and grow, and grow, and historians are now confronted with historical sources on an entirely new order of magnitude.

“…. [n]ow expectations have inverted. Everything may be recorded and preserved, at least potentially.”

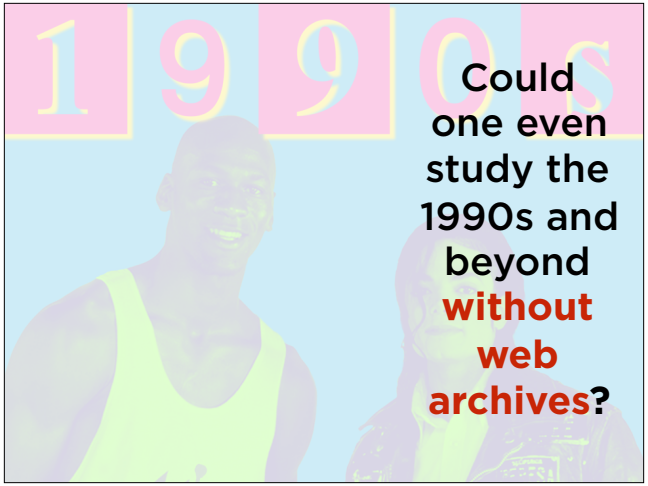
- James Gleick

These sources present both a boon and a challenge to historians. If the norm until the digital era was to have human information vanish, “[n]ow expectations have inverted. Everything may be recorded and preserved, at least potentially.” Useful historical information is being preserved at a rate that is accelerating with every passing day.

I think James is perhaps overstating the point here. Remember the previous process I outlined to you before - event happens, some traces of the past are left behind, and historians take those little traces to write their histories.

On a philosophical level, this is still true.

But, more traces than ever before are being left behind.



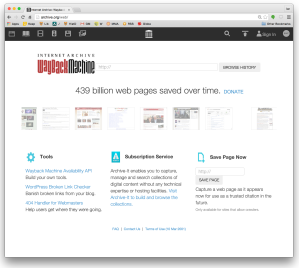
Let's not fool ourselves, this stuff matters. Things that would never have been preserved before are, and there are unparalleled opportunities for social historians to capture the broad context of the time that they are studied.

And, most importantly, I want to underscore that I don't think you can do justice to the 1990s and beyond if you do not consider the World Wide Web. This holds true for all branches of our historical profession.

Political historians cannot do justice to elections without understanding the tweets, the blogs, the websites that surround not only elections, but the everyday process of making policy, understanding public sentiment, and reaching out to the electorate on a new level. Military historians will have the voices of rank-and-file soldiers, playing out on discussion boards and other parts of the Web. Cultural and social historians, the source base is even more evident: the impact of online and offline culture playing out, and the voices of everyday people that would not be lost.

Yes, the Web is not a perfect democracy: there are many lower-income people who still do not have access to the Web, and there are age and racial cleavages as well. We cannot forget those. But we are still expanding our percentage of 'traces of the past' that are preserved, and we cannot forget that either.

Nightmare Scenario



This won't be enough!

Because if we don't, I want to outline a nightmare scenario.

[make sure to note the URL only restriction of the Wayback Machine]

… but what will our search engines look like?

Just like digitized newspapers.

My fear is that we'll query large databases, of literally hundreds of millions of documents, look at the first 100 results or even 1000, and we'll just be reaffirming what we already think.

Nightmare Scenario

- Historians rely uncritically on **date-ordered keyword search results**, putting them at mercy of search algorithms they do not understand (similar to digitized newspapers);
- 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…

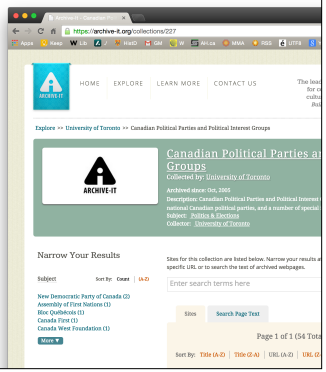
[and emphasize the comparison with historians and newspapers! i.e. we have these black boxes]

We already see some disturbing trends in this area, as with the Culturomics project.

What can we do to access this information?

Building Portals

- Democratizing access so that historians can use them.
- Building **transparent indexes**.
- But they have to be useful and tested…



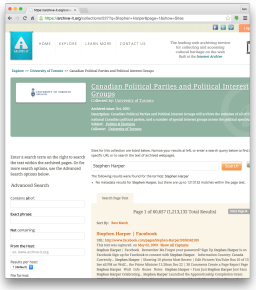
Pivotal Changes in Canadian Politics, 2005-2015

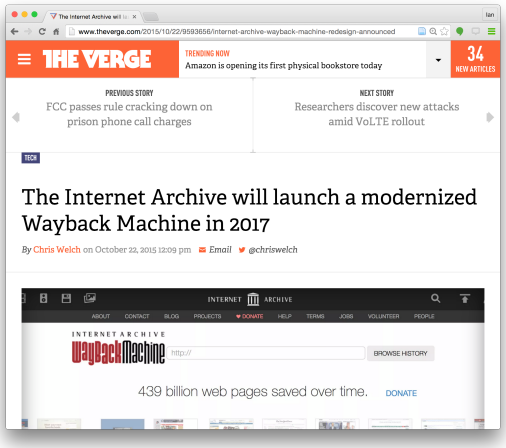
- Militarization of Canadian society?
- Change from 'natural governing party' of Liberals to Conservatives
- Major policy changes on foreign policy, environment, etc.
- How to measure?

collections SHOULD have been used a lot more

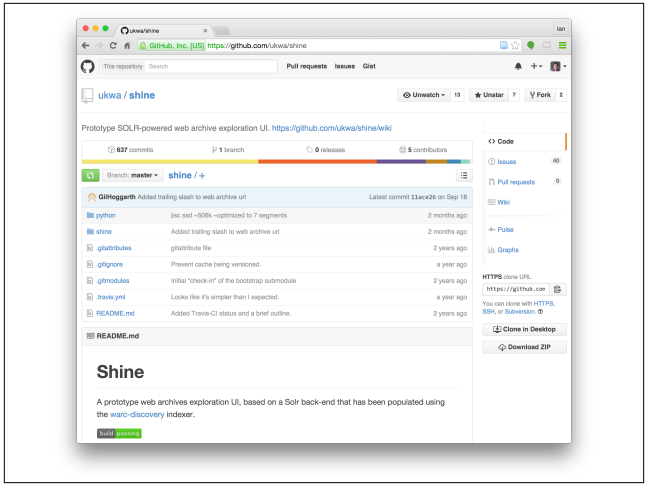
Current Interface

- Very limited - simple search engine, some advanced options; no facets
- Great collections.. but nobody uses them!





Some of this is in the area of providing sophisticated researcher access to web archives, and let people do what they need to do without having to be programmers. The upcoming rebuild of the Internet Archive's Wayback Machine is extremely exciting, and I've already set my countdown timer to 2017. I'm a bit afraid that it will deprive me of the straw man argument that I like to use around its URL-centric shortcomings, but I think the field will benefit.

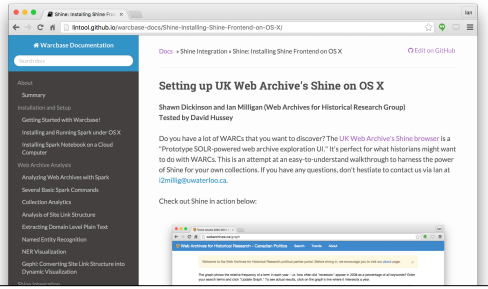


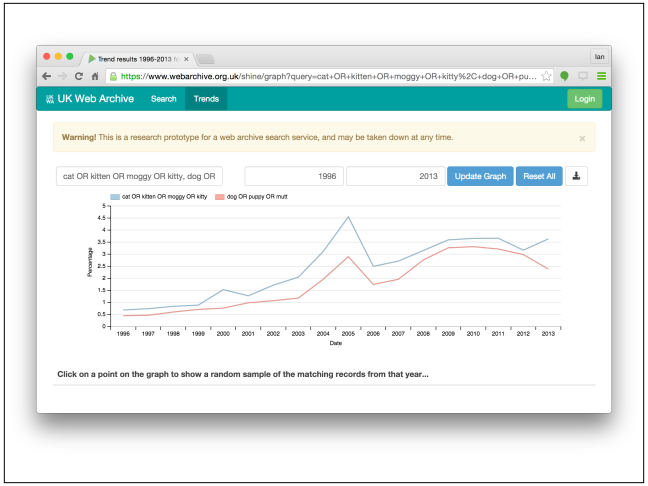
It’s been things like the United Kingdom Web Archive’s SHINE interface, though, that I think herald a lot of promise in the realm of DEMOCRATIZING ACCESS.

<describe shine>

Walkthroughs at:

http://lintool.github.io/warcbase-docs/Shine-Installing-Shine-Frontend-on-OS-X/



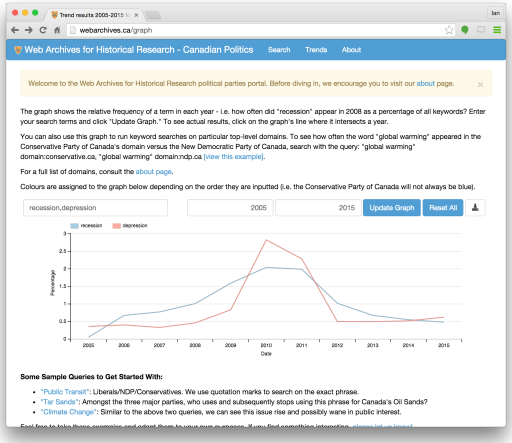


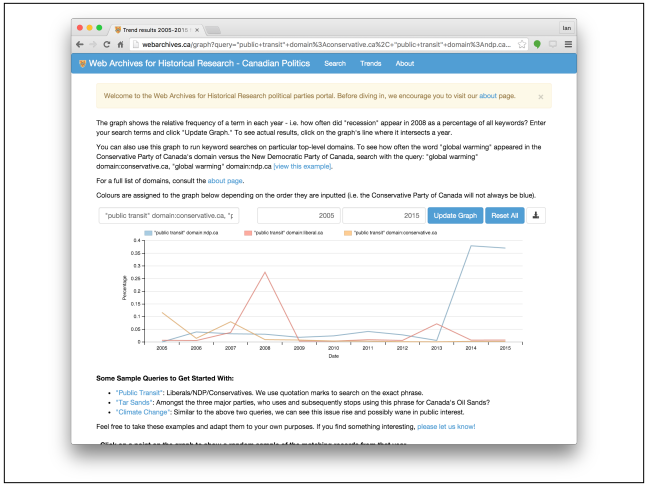
Jason Webber's cat example

With Nick Ruest (York University), Nich Worby (University of Toronto), Jimmy Lin (University of Waterloo)

We were really inspired by this as a way to build user engagement, and so on the eve of the recent Canadian federal election, we launched WebArchives.ca using the Shine build.

We used the University of Toronto's CPP collection, which was some 50 websites: all major political parties, minor ones, and a nebulous group of political interest groups, from David Suzuki's environmental organization to the Assembly of First Nations, to groups fighting for equal marriage.


Trend results 2005-2015
webarchives.ca/graph
Web Archives for Historical Research - Canadian Politics
Search
Trends
About
Welcome to the Web Archives for Historical Research political parties portal. Before diving in, we encourage you to visit our about page.
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
recession
depression
Percentage
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.

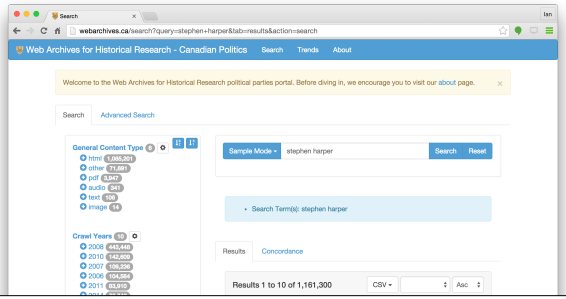

Trend results 2005-2015
webarchives.ca/graph?query="public+transit"+domain%3Aconservative.ca%2C+"public+transit"+domain%3Andp.ca...
Web Archives for Historical Research - Canadian Politics
Search
Trends
About
Welcome to the Web Archives for Historical Research political parties portal. Before diving in, we encourage you to visit our about page.
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).
"public transit" domain:conservative.ca, "p
2005
2015
Update Graph
Reset All
"public transit" domain:ndp.ca
"public transit" domain:liberal.ca
"public transit" domain:conservative.ca
Percentage
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!


webarchives.ca
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.
The Canadian Political Parties and Political Interest Groups Portal
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.
Curious how the Liberal Party of Canada responded to the 2008 financial crisis (a search for "recession" in 2008, liberal.ca)? How the Canadian Centre for Policy Alternatives reacted to Michael Ignatieff? Now you can check it all out.
Options include:
Basic keyword searching [Example: "Rob Ford", only Liberal.ca]
Graphing trends over time [Example: Liberal Opposition Leaders, 2005-2015]
Advanced search, including words in proximity to each other [Example: environmental and tax within 25 words of each other]
Below, here are all of the links for the entire time period, visualized below.
twitter.com
facebook.com
youtube.com
google.com
canadianlabour.ca
theglobeandmail.com
parl.gc.ca
cbc.ca
hotmail.com
canada.com
gmail.com

Five Things We 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

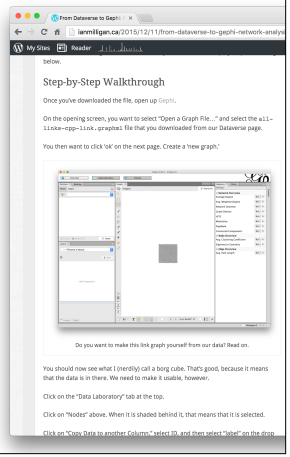
Good for public engagement - but limited for scholarship....



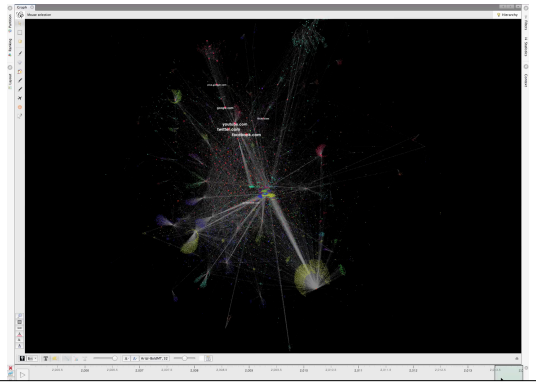
Getting over my bias towards content **and embracing metadata**

Gephi 0.9
(http://gephi.github.io/)

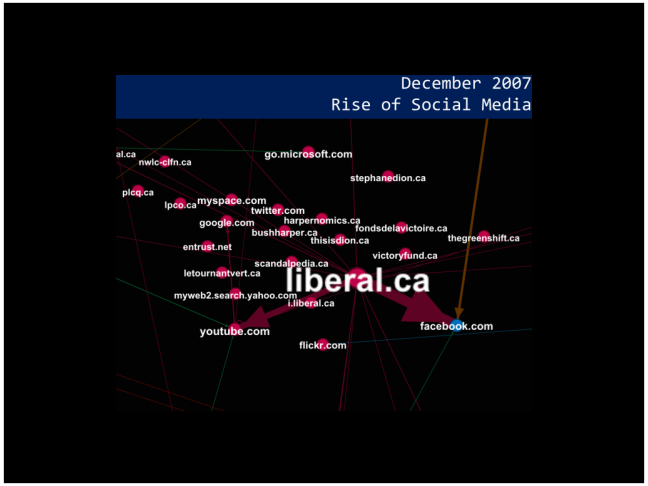
Walkthrough at ianmilligan.ca: “From Dataverse to Gephi” - try it on this data!

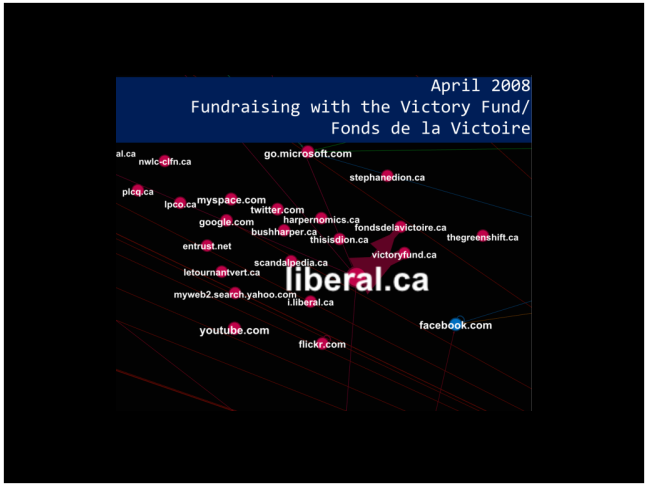


Metadata Extraction

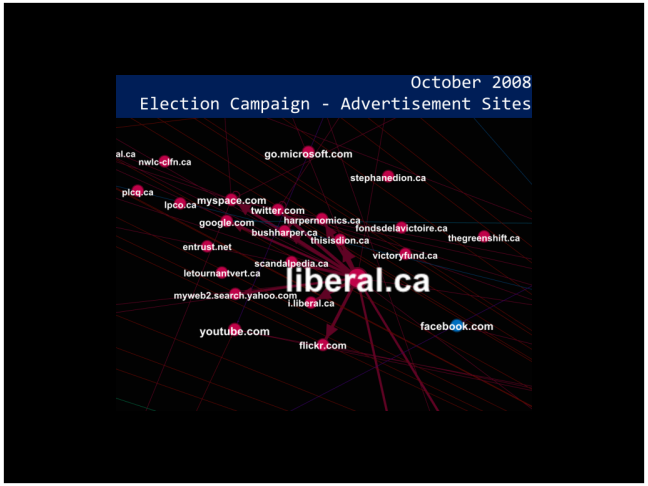


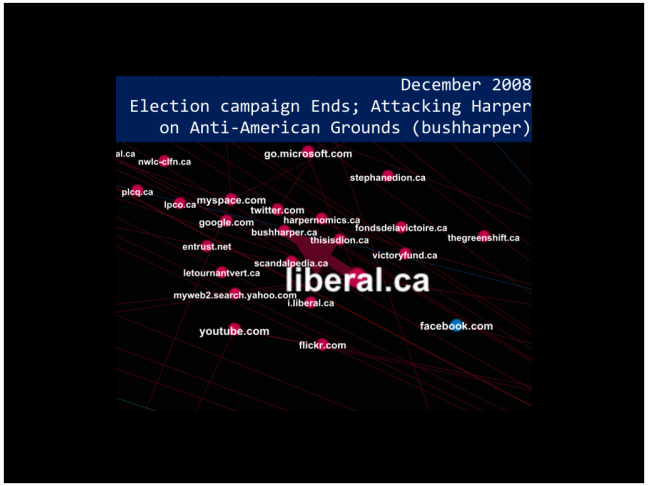

December 2006
Stephane Dion Elected Leader of Party
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
harpernomics.ca
google.com
bushharper.ca
fondsdelavictoire.ca
thegreenshift.ca
thisisdion.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca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

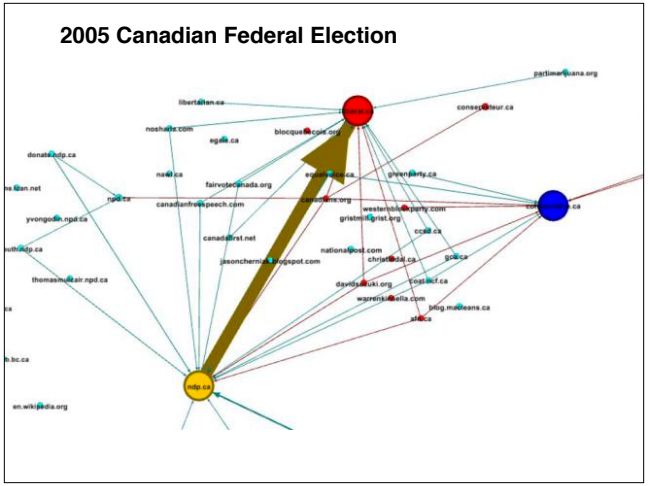

December 2007
Rise of Social Media
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
google.com
harpernomics.ca
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com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

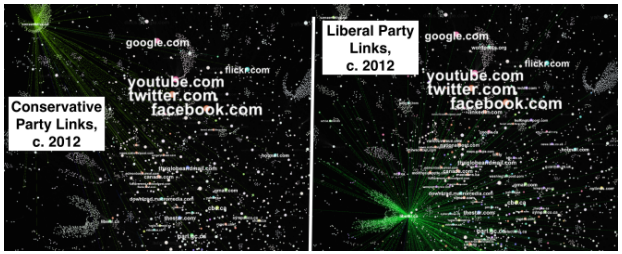

July 2008
The Green Shift Announced!
al.ca
nwlc-clfn.ca
go.microsoft.com
stephanedion.ca
plcq.ca
lpco.ca
myspace.com
twitter.com
harpernomics.ca
google.com
fondsdelavictoire.ca
bushharper.ca
thisisdion.ca
thegreenshift.ca
entrust.net
victoryfund.ca
scandalpedia.ca
letournantvert.ca
liberal.ca
myweb2.search.yahoo.com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

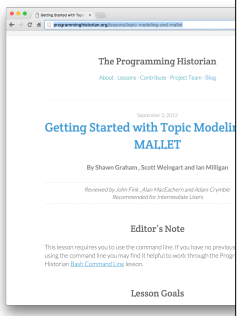

2005 Canadian Federal Election
partimarijuana.org
libertarian.ca
conservateur.ca
noshame.com
egale.ca
donate.ndp.ca
fairvotecanada.org
greenparty.ca
canadianfreespeech.com
yvongodin.npd.ca
canadafirst.net
ccsd.ca
gristmillgrist.org
nationalpost.com
gpc.ca
jasonchermak.blogspot.com
davidsuzuki.org
warrenkinsella.com
blog.macleans.ca
thomasmulcair.npd.ca
en.wikipedia.org

Metadata Extraction



Topic Modelling using MALLET (http://mallet.cs.umass.edu/)

Walkthrough:
http://programminghistorian.org/lessons/topic-modeling-and-mallet



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.

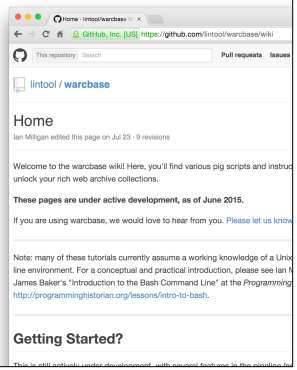
Metadata Extraction

- 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.

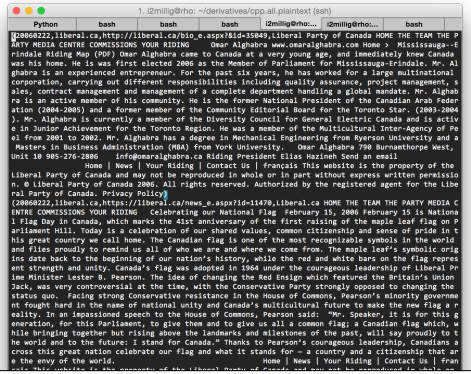
Interdisciplinary

Warcbase

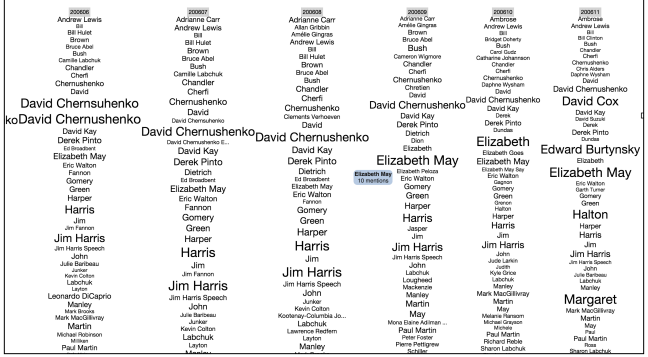
- Jimmy Lin (main developer), Ian Milligan, Jeremy Wiebe, Alice Zhou, all University of Waterloo
- Developing code, walkthroughs, and instructions for historians to be able to take a directory of WARCs and…



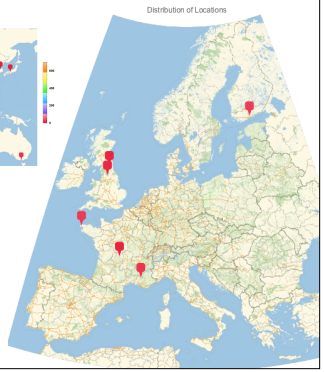
Extract all Plain Text



Extract Entities



Extract Entities

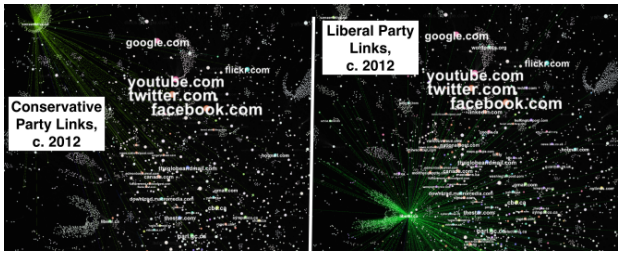



```
>>> int = SemanticInterpretation[0].0/@processedfreq[[All, 1]]
```

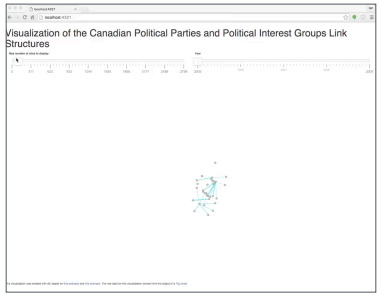
Out[]= Canada , Calgary , Colombia , Montreal , $Failed, Ontario, Canada , Afghanistan ,
Ottawa , Manitoba, Canada , British Columbia, Canada , Toronto , Nova Scotia, Canada ,
Saskatchewan, Canada , Quebec, Canada , Alberta, Canada , Winnipeg , Washington ,
Canada , Nunavut, Canada , White Horse , United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

Extract Entities

Extract Links/Gephi Connector

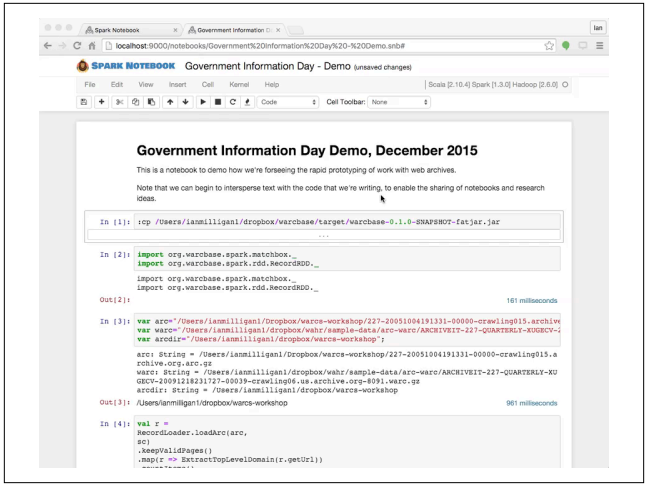


Or D3.js link networks in browser



All walkthroughs at:
docs.warcbase.org

Bringing it all together in
a notebook environment


Spark Notebook
Government Information
Ian
localhost:9000/notebooks/Government%20Information%20Day%20-%20Demo.snb#
SPARK NOTEBOOK
Government Information Day - Demo (unsaved changes)
File Edit View Insert Cell Kernel Help
Scala [2.10.4] Spark [1.3.0] Hadoop [2.6.0]
Code
Cell Toolbar: None
Government Information Day Demo, December 2015
This is a notebook to demo how we're forseeing the rapid prototyping of work with web archives.
Note that we can begin to intersperse text with the code that we're writing, to enable the sharing of notebooks and research ideas.
In [1]: :cp /Users/ianmilligan1/dropbox/warcbase/target/warcbase-0.1.0-SNAPSHOT-fatjar.jar
In [2]: import org.warcbase.spark.matchbox._
import org.warcbase.spark.rdd.RecordRDD._
import org.warcbase.spark.matchbox._
import org.warcbase.spark.rdd.RecordRDD._
Out[2]: 161 milliseconds
In [3]: var arc="/Users/ianmilligan1/Dropbox/warcs-workshop/227-20051004191331-00000-crawling015.archiv
var warc="/Users/ianmilligan1/dropbox/wahr/sample-data/arc-warc/ARCHIVEIT-227-QUARTERLY-XUGECV-
var arcdir="/Users/ianmilligan1/dropbox/warcs-workshop";
arc: String = /Users/ianmilligan1/Dropbox/warcs-workshop/227-20051004191331-00000-crawling015.a
rchive.org.arc.gz
warc: String = /Users/ianmilligan1/dropbox/wahr/sample-data/arc-warc/ARCHIVEIT-227-QUARTERLY-XU
GECV-20091218231727-00039-crawling06.us.archive.org-8091.warc.gz
arcdir: String = /Users/ianmilligan1/dropbox/warcs-workshop
Out[3]: /Users/ianmilligan1/dropbox/warcs-workshop 961 milliseconds
In [4]: val r =
RecordLoader.loadArc(arc,
sc)
.keepValidPages()
.map(r => ExtractTopLevelDomain(r.getUrl))

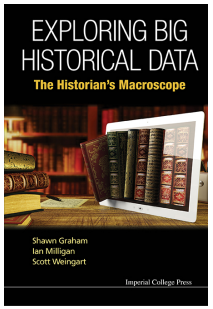
Where to learn?

Programming Historian

- Network Analysis Lessons
- Topic Modeling Lessons
- Command Line Lessons
- etc.

Exploring Big Historical Data

- Check out our draft at **macroscope.org**
  - Conceptual introduction to topic modelling
  - Network analysis
  - Visualizations
  - Field of digital humanities



Events

- **Software Carpentry** – in-person events, looking into building connections with *Programming Historian*
- **Interdisciplinary hackathons** - *Archives Unleashed* (Toronto, March 2016; Washington, June 2016 - TBA)
- **Conferences** - Like this one, or others

… but most of all, a willingness to learn and fail.

Because, as I hope I have shown today..

it’s worth it.


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

Thanks very much!

Questions?

Ian Milligan
Assistant Professor
@ianmilligan1

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

And would love your questions today.