

How Web Archives and Tweets are Reshaping Today's Historical Record

**University of Saskatchewan
11 March 2016**

Ian Milligan
Assistant Professor
@ianmilligan1



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

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

... we need to think
about big data ...

Today's Talk

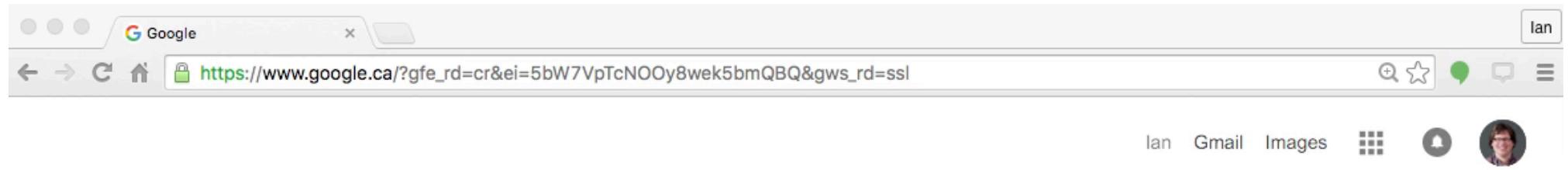
- **1. Prologue:** Big Data is everywhere
- **2. The Web Age:** Will accelerate this process
- **3. What can we do with big data?**

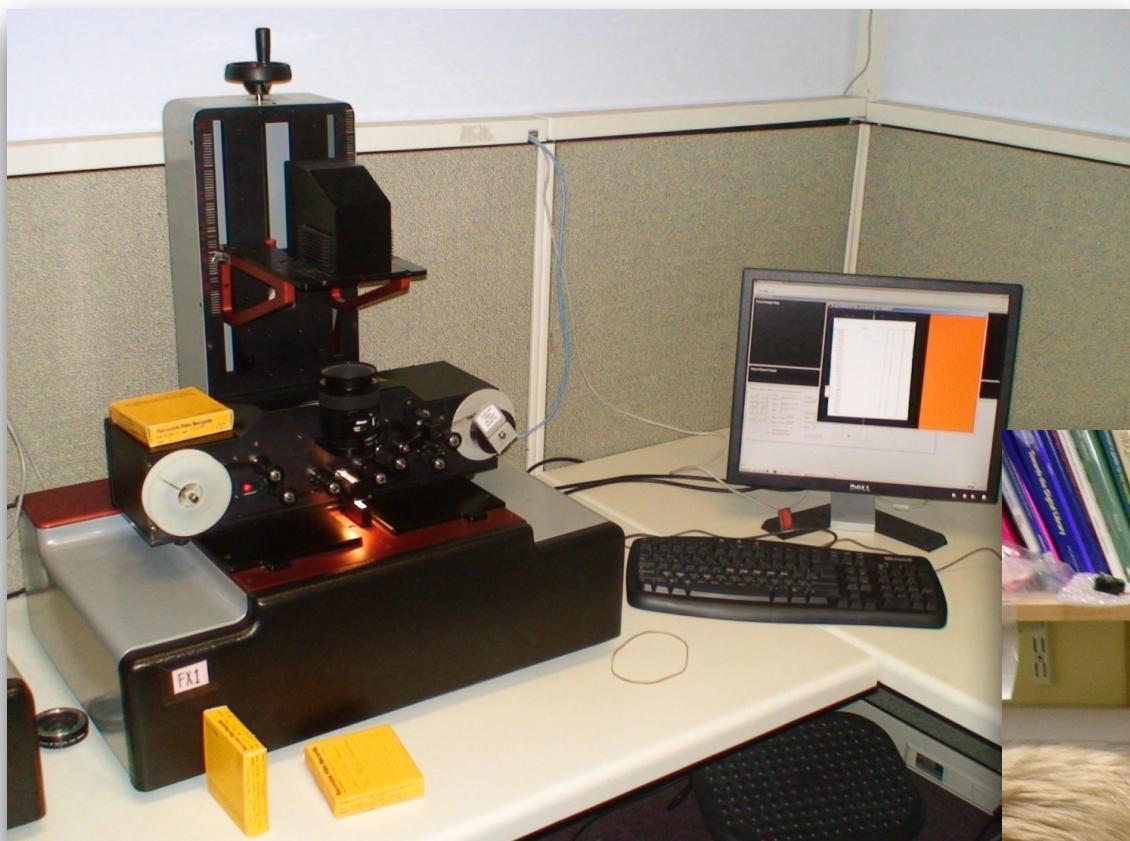
**A Prologue:
Big Data is
Everywhere**

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*)

**Why is it
everywhere?**





Advanced Search - ProQuest

search.proquest.com/hnptorontostar/advanced/accountid=14906

Searching: 1 database ▾

0 Recent searches | 0 Selected items | My Research | Exit

« All databases | News & Newspapers databases

Preferences | English ▾ | Help ?

ProQuest | ProQuest Historical Newspapers: Toronto Star (1894-2011)

Basic Search | Advanced ▾ | Obituaries | Publications

Advanced Search

Look Up Citation | Command Line | Find Similar

Field codes | Search tips

in Anywhere

in Anywhere

in Anywhere

AND ([] OR []) AND ([] OR [])

Add a row | Remove a row

Search | Clear form

Search options

Publication date: All dates

Sort results by: Publication date (most recent first)

Items per page: 50

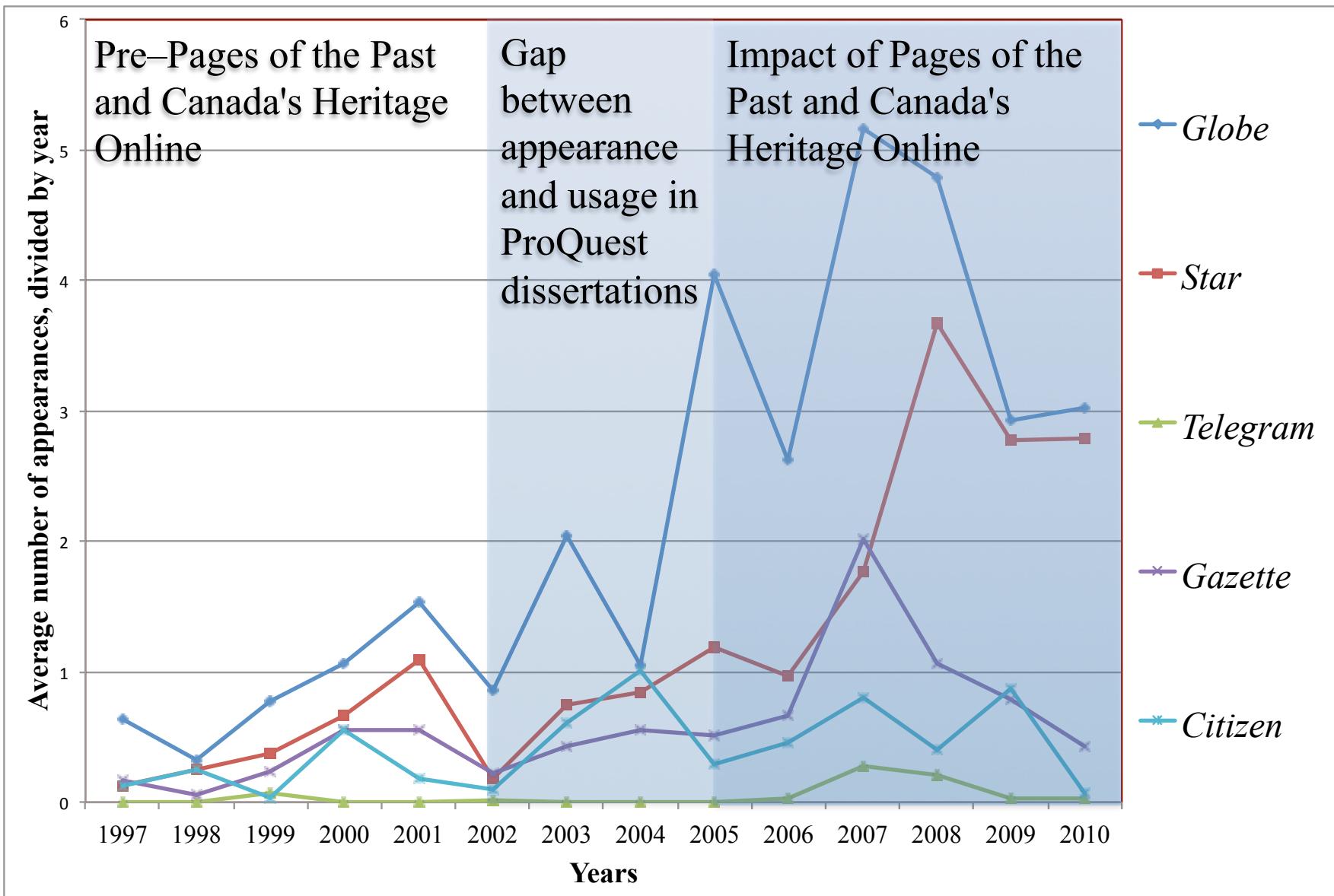
Duplicates: Include duplicate documents *i*

Search | Clear form

Search subject areas

Use search forms customized for each subject.

	The Arts
	Business
	Dissertations & Theses
	Health & Medicine
	History
	Literature & Language
	News & Newspapers



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

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

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

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

**A process that is only
now beginning to
accelerate.**

First - more data than ever before being preserved;

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

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL
SERIES I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

370

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL
SERIES I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

371

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
SERIES I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

372

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
SERIES I
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

373

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series II
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29

374

CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series II
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Subseries F Democratic Party
BOX 30

JOHN J. BURNS LIBRARY
BOSTON COLLEGE

375

Scarcity

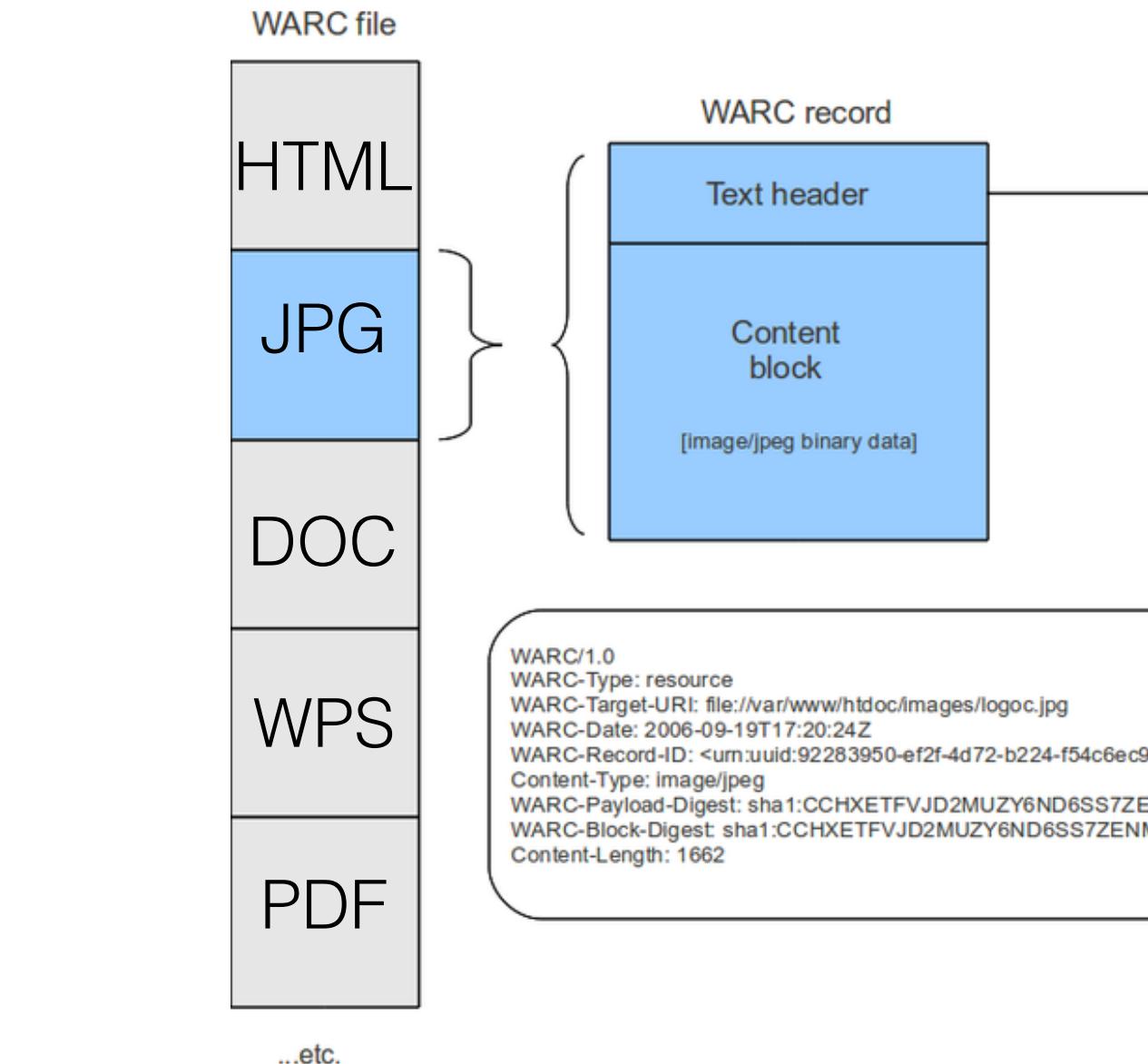




WebARChive (WARC) File

????

- What does this really mean?



Department of History

web.archive.org/web/19970128205440/http://www.usask.ca/history/

INTERNET ARCHIVE
Wayback Machine

204 captures 28 Jan 97 - 9 Sep 15

http://www.usask.ca/history/ Go DEC JAN JUL 28 1996 1997 1998 Close Help ?

DEPARTMENT Of HISTORY

UNIVERSITY OF SASKATCHEWAN

9 Campus Drive
Saskatoon, SK S7N 5A5



copyright © 1995, Department of History, University of Saskatchewan

Last Update (*this page only*): Wednesday, 22-Jan-97 15:02:45 CST
[See [Recent Changes](#) to History Web Files]

Telephone: (306)966-5792 Department Head: *Bill Waiser*
Fax: (306)966-5852 Graduate Director: *Michael Hayden*
E-mail: hist.dept@usask.ca Undergraduate Director: *Michael Swan*

The History Department of the [University of Saskatchewan](#) in [Saskatoon](#) is one of the largest departments in the [College of Arts and Science](#). Nevertheless, the three staff and twenty-one faculty members work to create an atmosphere in which students can find the information and advice they need about courses, careers, and the discipline of history.

SOME **HOT SPOTS** TO VISIT ...

Canadian Journal of History	1912-Present Theses in History at UofS	H-Canada [English] [French]	UofS Library & Univ. Archives Essay Award	NEW Quiz Show	Other Web Sites
Canadian Historical Assoc. 1997	Now Then newsletter	What's UP?	Send a Postcard NEW	Amazon Book Search NEW	...

You are viewing an archived web page, collected at the request of [Internet Archive Global Events](#) using [Archive-It](#). This page was captured on 5:07:27 Dec 03, 2011, and is part of the [Occupy Movement 2011/2012](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. [Videos Metadata](#)

Welcome [login](#) | [signup](#)
Language [en](#) [es](#) [fr](#)

OccupyWallStreet

The revolution continues [worldwide!](#)

[News](#) [LiveStream](#) [#HowToOccupy](#) [Forum](#) [Chat](#) [User Map](#) [NYCGA](#) [About](#) [Donate](#)

Farmers Join Occupy Wall Street, Calling for Food Justice

Posted 5 hours ago on Dec. 2, 2011, 6:21 p.m. EST by [OccupyWallSt](#)



As Wall Street's corrupt influence on the economy has grown, the corporate ownership of our food system has hurt the health and livelihood's of some of our most vulnerable communities. This *Sunday, December 4th* food justice activists and occupiers will be traveling from as far as Colorado, Iowa, Maine and Upstate New York to join together for the [Occupy Wall Street FARMERS' MARCH](#). Through a day of dialogue, musical performances, and a march, farmers and their urban allies working for food justice in their communities will form alliances to fight and expose corporate control of the food supply.

Events throughout the day will call and inspire participants to fight against the corporate manipulation of the agriculture system. An industry that is responsible for using chemical toxins tied to soaring obesity rates, heart disease and diabetes and limiting access to affordable, wholesome food to the country's poorest citizens.

[Read More...](#)

[30 Comments](#)

Occupy Wall Street Goes Home

Posted 1 day ago on Dec. 1, 2011, 3:04 p.m. EST by [OccupyWallSt](#)



On December 6th Occupy Wall Street will join in solidarity with a Brooklyn community to re-occupy a foreclosed home. The day of action marks a national kick-off for a new frontier for the [occupy movement: the liberation of vacant bank owned homes for those in need](#). The banks are

General Inquiries: general@occupywallst.org
Press Inquiries: press@occupywallst.org
Press Phone: +1 (347) 292-1444
Help & Directions: +1 (516) 708-4777
Watch: [The world we're building](#)
Read: [This call to action](#)
Liberty Square Eviction Defense: Text "@occupyalert" to 23559 to receive alerts in the event of imminent emergency.

Occupy Wall Street is leaderless resistance movement with people of many [colors](#), genders and political persuasions. The one thing we all have in common is that [We Are The 99%](#) that will no longer tolerate the greed and corruption of the 1%. We are using the revolutionary [Arab Spring](#) tactic to achieve our ends and encourage the use of nonviolence to maximize the safety of all participants.

This #ows movement empowers real people to create real change from the bottom up. We want to see a [general assembly](#) in every backyard, on every street corner because we don't need Wall Street and we don't need politicians to build a better society.

the only solution is WorldRevolution

[Click here](#) for NYCGA committee meeting times.





This webpage is not available

[Details](#)

You are viewing an archived web page, collected at the request of [Internet Archive Global Events](#) using [Archive-It](#). This page was captured on 5:07:27 Dec 03, 2011, and is part of the [Occupy Movement 2011/2012](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page. [Videos Metadata](#)

Welcome [login](#) | [signup](#)
Language [en](#) [es](#) [fr](#)

OccupyWallStreet

The revolution continues [worldwide!](#)

[News](#) [LiveStream](#) [#HowToOccupy](#) [Forum](#) [Chat](#) [User Map](#) [NYCGA](#) [About](#) [Donate](#)

Farmers Join Occupy Wall Street, Calling for Food Justice

Posted 5 hours ago on Dec. 2, 2011, 6:21 p.m. EST by [OccupyWallSt](#)



As Wall Street's corrupt influence on the economy has grown, the corporate ownership of our food system has hurt the health and livelihood's of some of our most vulnerable communities. This *Sunday, December 4th* food justice activists and occupiers will be traveling from as far as Colorado, Iowa, Maine and Upstate New York to join together for the [Occupy Wall Street FARMERS' MARCH](#). Through a day of dialogue, musical performances, and a march, farmers and their urban allies working for food justice in their communities will form alliances to fight and expose corporate control of the food supply.

Events throughout the day will call and inspire participants to fight against the corporate manipulation of the agriculture system. An industry that is responsible for using chemical toxins tied to soaring obesity rates, heart disease and diabetes and limiting access to affordable, wholesome food to the country's poorest citizens.

[Read More...](#)

[30 Comments](#)

Occupy Wall Street Goes Home

Posted 1 day ago on Dec. 1, 2011, 3:04 p.m. EST by [OccupyWallSt](#)



On December 6th Occupy Wall Street will join in solidarity with a Brooklyn community to re-occupy a foreclosed home. The day of action marks a national kick-off for a new frontier for the [occupy movement: the liberation of vacant bank owned homes for those in need](#). The banks are

General Inquiries: general@occupywallst.org
Press Inquiries: press@occupywallst.org
Press Phone: +1 (347) 292-1444
Help & Directions: +1 (516) 708-4777
Watch: [The world we're building](#)
Read: [This call to action](#)
Liberty Square Eviction Defense: Text "@occupyalert" to 23559 to receive alerts in the event of imminent emergency.

Occupy Wall Street is leaderless resistance movement with people of many [colors](#), genders and political persuasions. The one thing we all have in common is that [We Are The 99%](#) that will no longer tolerate the greed and corruption of the 1%. We are using the revolutionary [Arab Spring](#) tactic to achieve our ends and encourage the use of nonviolence to maximize the safety of all participants.

This #ows movement empowers real people to create real change from the bottom up. We want to see a [general assembly](#) in every backyard, on every street corner because we don't need Wall Street and we don't need politicians to build a better society.

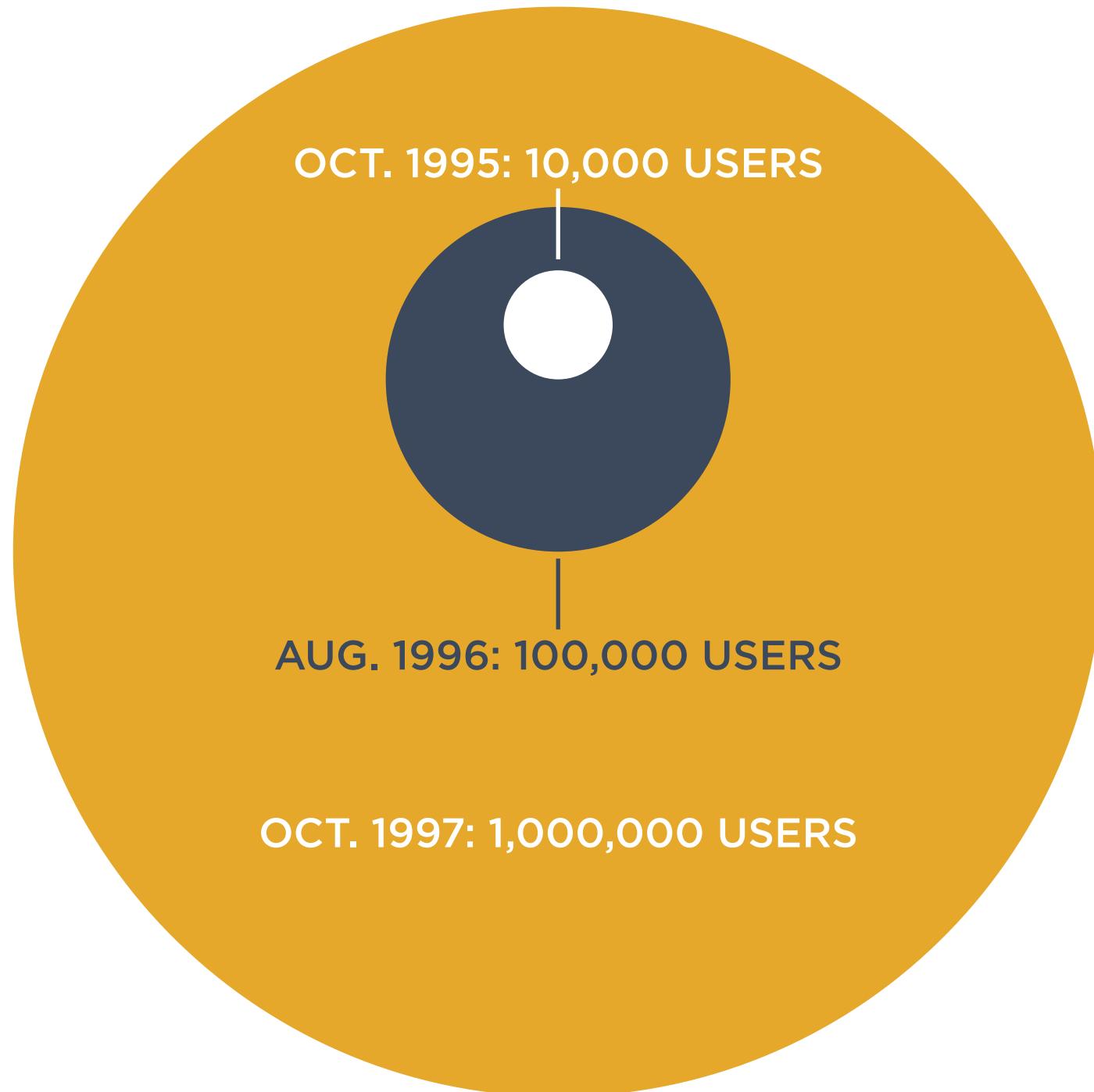
the only solution is WorldRevolution

[Click here](#) for NYCGA committee meeting times.

Scarcity Abundance



GEOCITIES USERS:



This is a scale that **boggles**
the mind - compare it to the
Old Bailey (197,745 trials
between 1674 and 1913)

RIC'S GRILL
STEAK SEAFOOD
& CHOP HOUSE



October 17, 2015

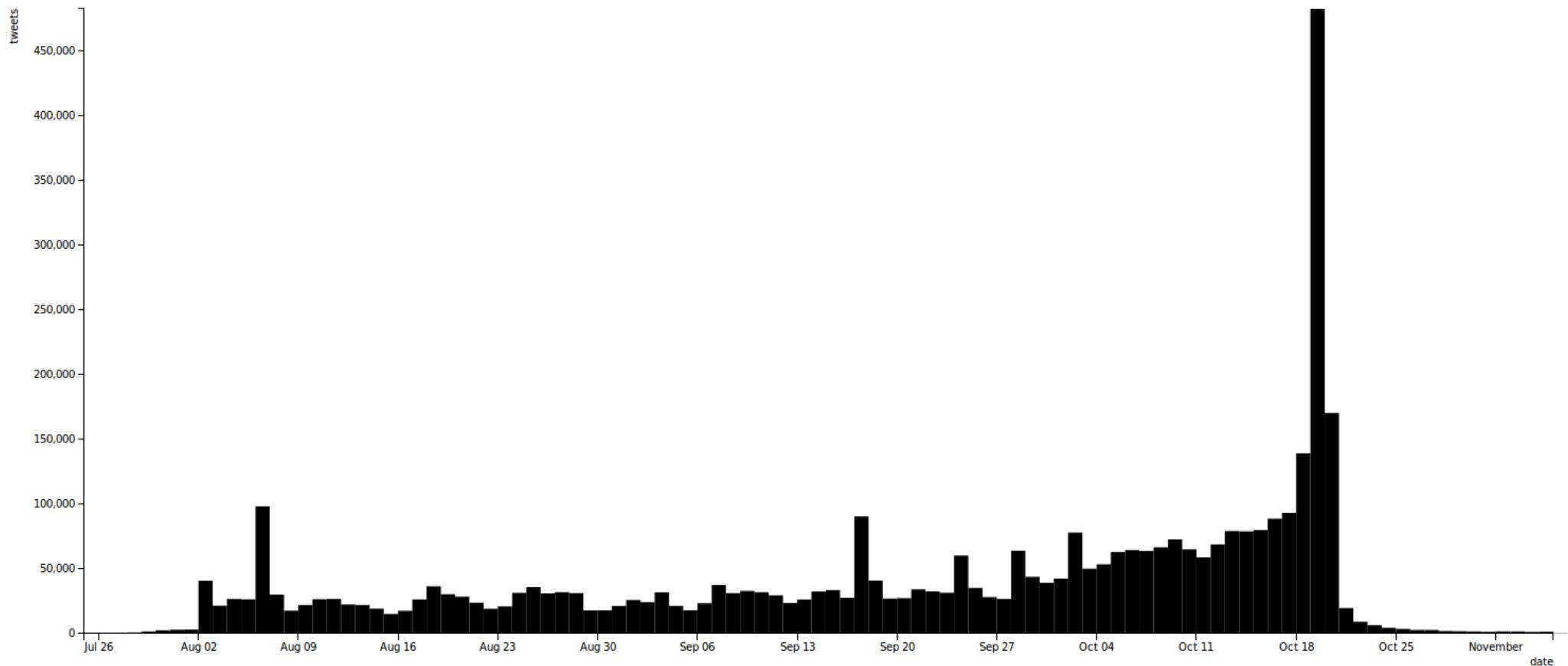
October 18, 2015

October 19, 2015

October 20, 2015

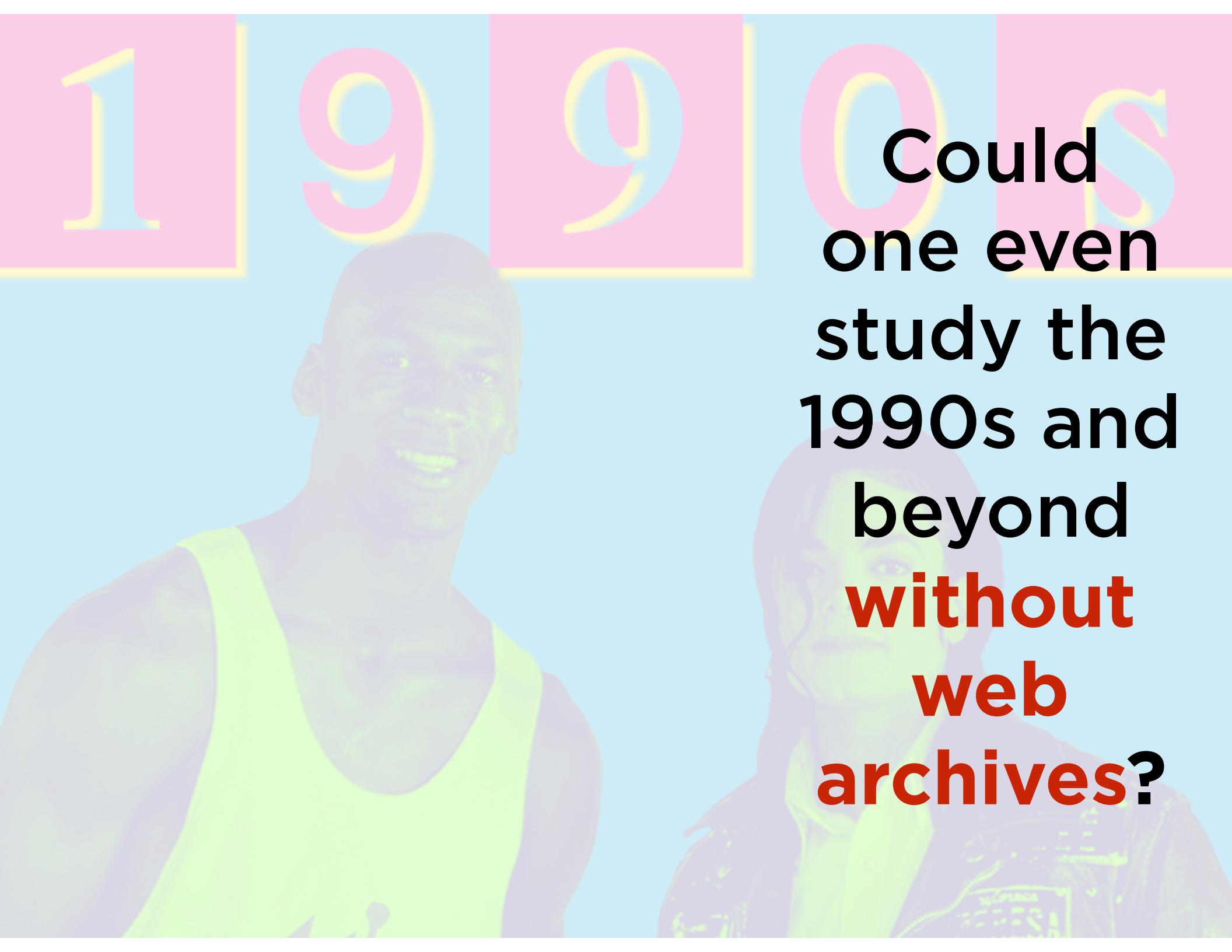
Frequency on the #ELXN42 Hashtag

2015-07-25 17:56:45 EDT to 2015-11-05 06:46:45 EST



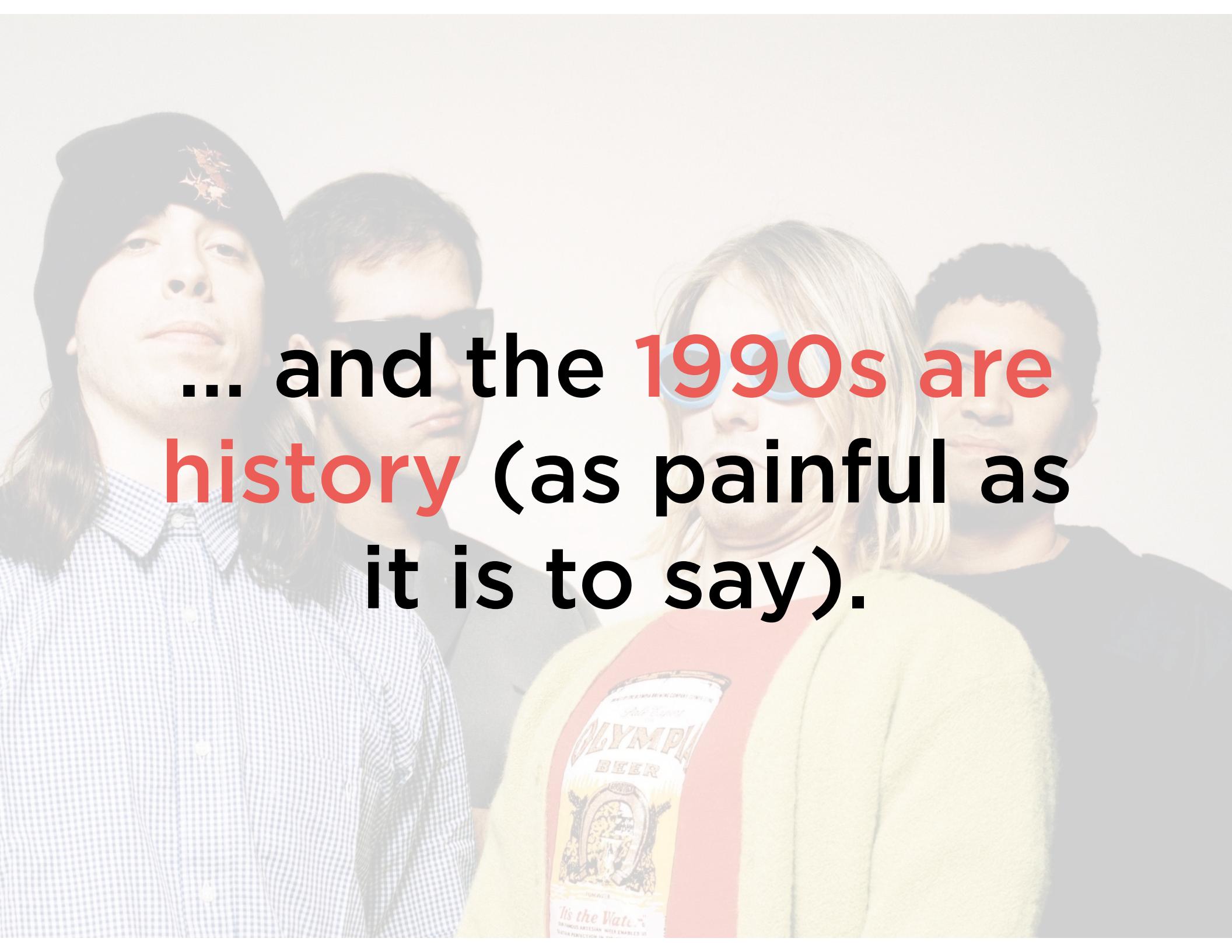
“.... [n]ow expectations have inverted. Everything may be recorded and preserved, at least potentially.”

- James Gleick



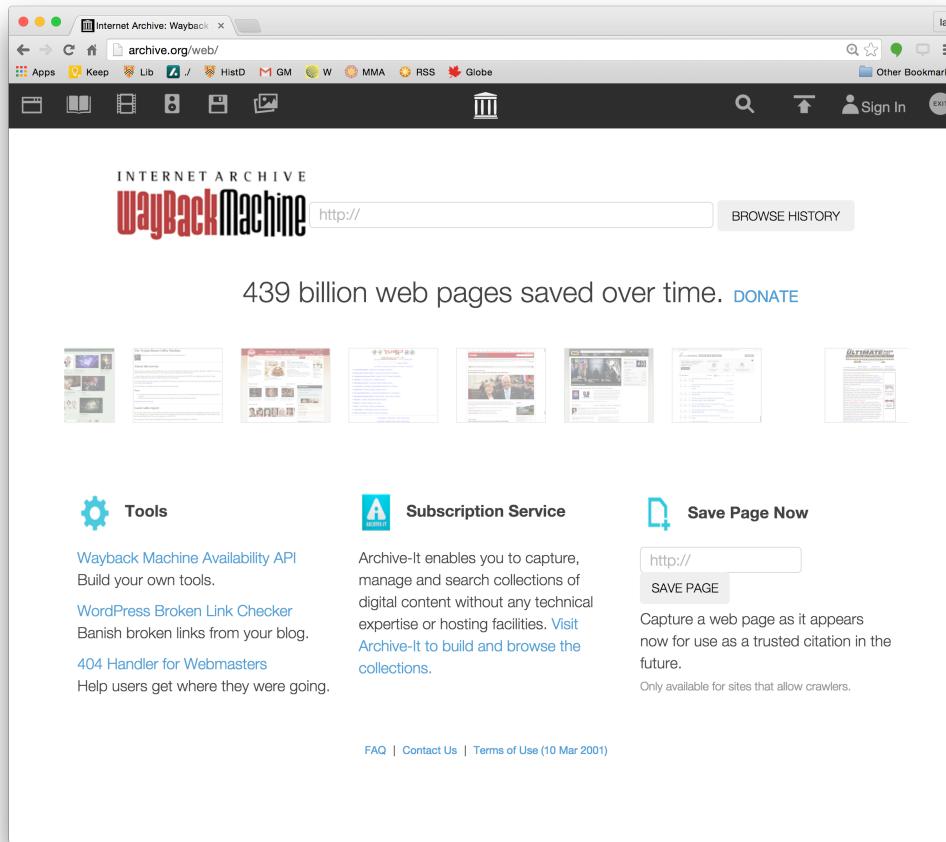
Could
one even
study the
1990s and
beyond
without
web
archives?

1990s



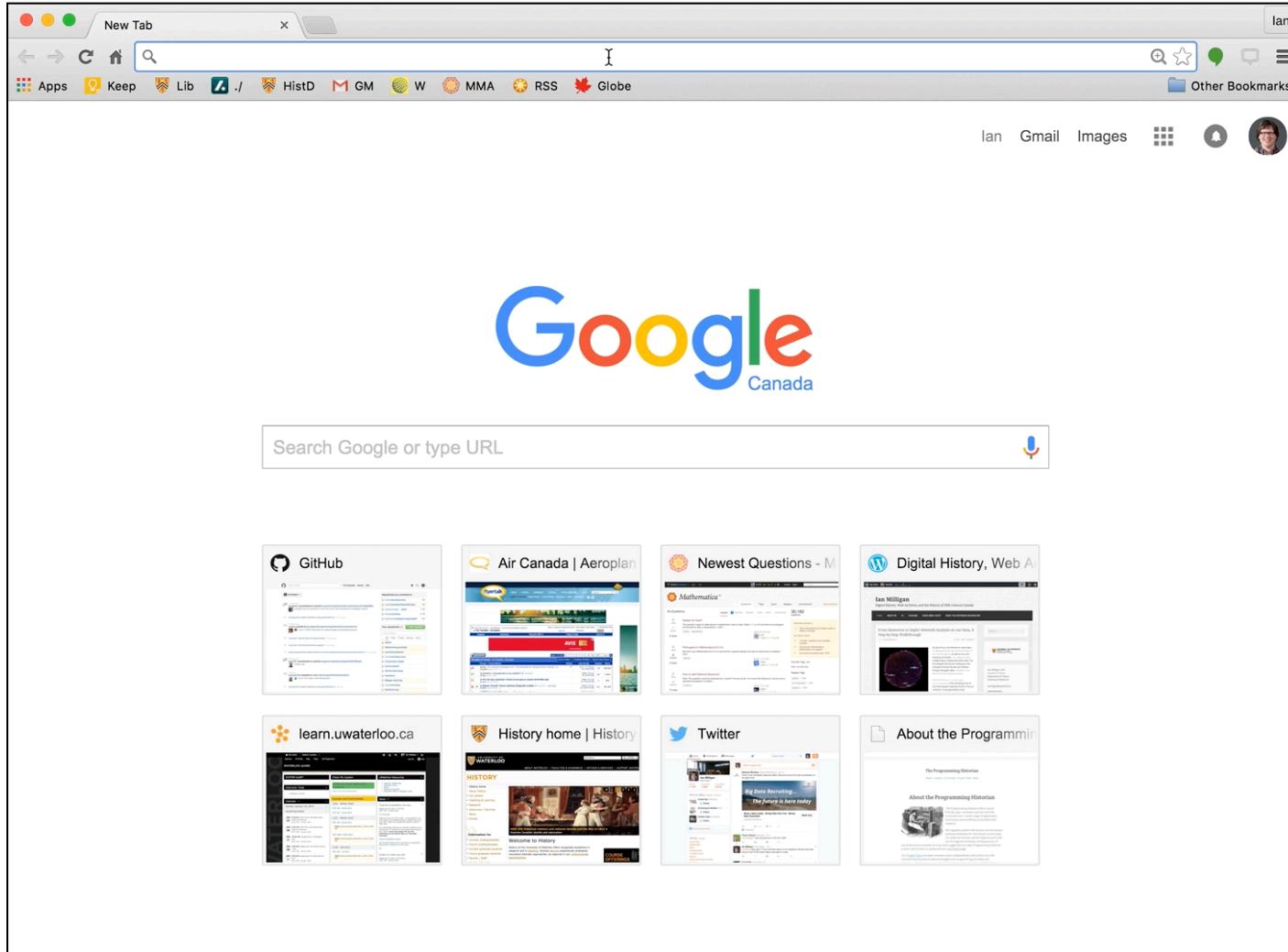
... and the 1990s are
history (as painful as
it is to say).

Nightmare Scenario



This won't be enough!

Nightmare Scenario



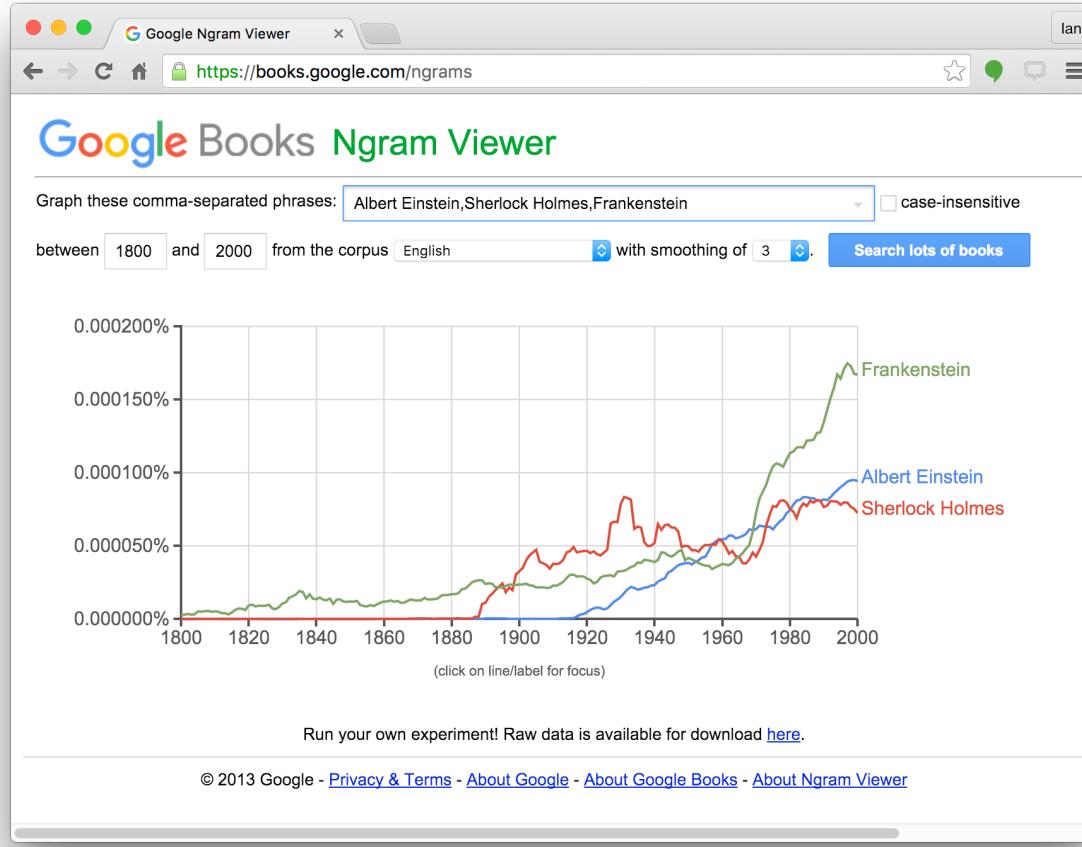
This won't be enough!



**... but what will our
search engines look
like?**

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);



My deepest fear:
Historians are completely left out of
post-1996 research, letting everybody else
do the work (a la Culturomics project/
Science magazine article);
Our profession gets left behind...

The historians who came to the meeting were intelligent, kind, and encouraging. But they didn't seem to have a good sense of how to wield quantitative data to answer questions, didn't have relevant computational skills, and didn't seem to have the time to dedicate to a big multiauthor collaboration. It's not their fault: these things don't appear to be taught or encouraged in history departments right now.

- Erez Leiberman Aiden and Jean-Baptiste Michel

**What can we do to
access this information
and avoid my nightmare?**

Building Portals

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

The screenshot shows a web browser window with the title "Archive-It - Canadian Political Parties and Political Interest Groups". The URL in the address bar is <https://archive-it.org/collections/227>. The page features the Archive-It logo and navigation links for HOME, EXPLORE, LEARN MORE, and CONTACT US. Below this, it displays the title "Canadian Political Parties Groups" collected by "University of Toronto". It notes the collection was archived since Oct, 2005, and describes it as containing national Canadian political parties and a number of specific political interest groups. The subject is listed as "Politics & Elections". A "Narrow Your Results" section allows users to filter by subject, with options like New Democratic Party of Canada (2), Assembly of First Nations (1), Bloc Québécois (1), Canada First (1), and Canada West Foundation (1). A search bar at the bottom right allows users to enter search terms here. The footer includes links for "Sites" and "Search Page Text", and a page number "Page 1 of 1 (54)".

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.
- Sweeping change to how we understand our history
- How to measure?



Canadian Political Parties & Political Interest Group Collection

- 50 Websites
 - All major political parties
 - Minor political parties
 - Political interest groups
- Collected quarterly between 2005 & present.



Current Interface

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

The screenshot shows a web browser window displaying the Archive-It collection page for "Canadian Political Parties and Political Interest Groups". The URL in the address bar is <https://archive-it.org/collections/227?q=Stephen+Harper&page=1&show=Sites>. The page features a header with the Archive-It logo, navigation links for HOME, EXPLORE, LEARN MORE, and CONTACT US, and a tagline about collecting and accessing cultural heritage on the web. Below the header, it says "Explore >> University of Toronto >> Canadian Political Parties and Political Interest Groups". The main content area displays the "Canadian Political Parties and Political Interest Groups" collection, which was collected by the University of Toronto and archived since Oct, 2005. It includes a brief description, subject information (Politics & Elections), and a collector note. A search bar at the bottom allows users to search for specific terms within the collection results.

ukwa/shine lan

GitHub, Inc. [US] <https://github.com/ukwa/shine>

This repository Search Pull requests Issues Gist

ukwa / shine Unwatch 13 Unstar 7 Fork 2

Prototype SOLR-powered web archive exploration UI. <https://github.com/ukwa/shine/wiki>

637 commits 1 branch 0 releases 5 contributors

Branch: master → shine / +

GilHoggarth Added trailing slash to web archive url Latest commit 11ace26 on Sep 18

File	Description	Time
python	jisc ssd ~506k ~optimized to 7 segments	2 months ago
shine	Added trailing slash to web archive url	2 months ago
.gitattributes	gitattribute file	2 years ago
.gitignore	Prevent cache being versioned.	a year ago
.gitmodules	Initial "check-in" of the bootstrap submodule	2 years ago
.travis.yml	Looks like it's simpler than I expected.	a year ago
README.md	Added Travis-CI status and a brief outline.	2 years ago

README.md

Shine

A prototype web archives exploration UI, based on a Solr back-end that has been populated using the [warc-discovery](#) indexer.

build passing

Code Issues Pull requests Wiki Pulse Graphs

HTTPS clone URL <https://github.com>

You can clone with [HTTPS](#), [SSH](#), or [Subversion](#).

Clone in Desktop Download ZIP

**Great research question that our
contemporary historians were
studying (Canada changing)**

+

**Great collection (all the political
parties + many interest groups)**

+

Ope Source Software





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



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

The Canadian Political Parties and Political Interest Groups Portal

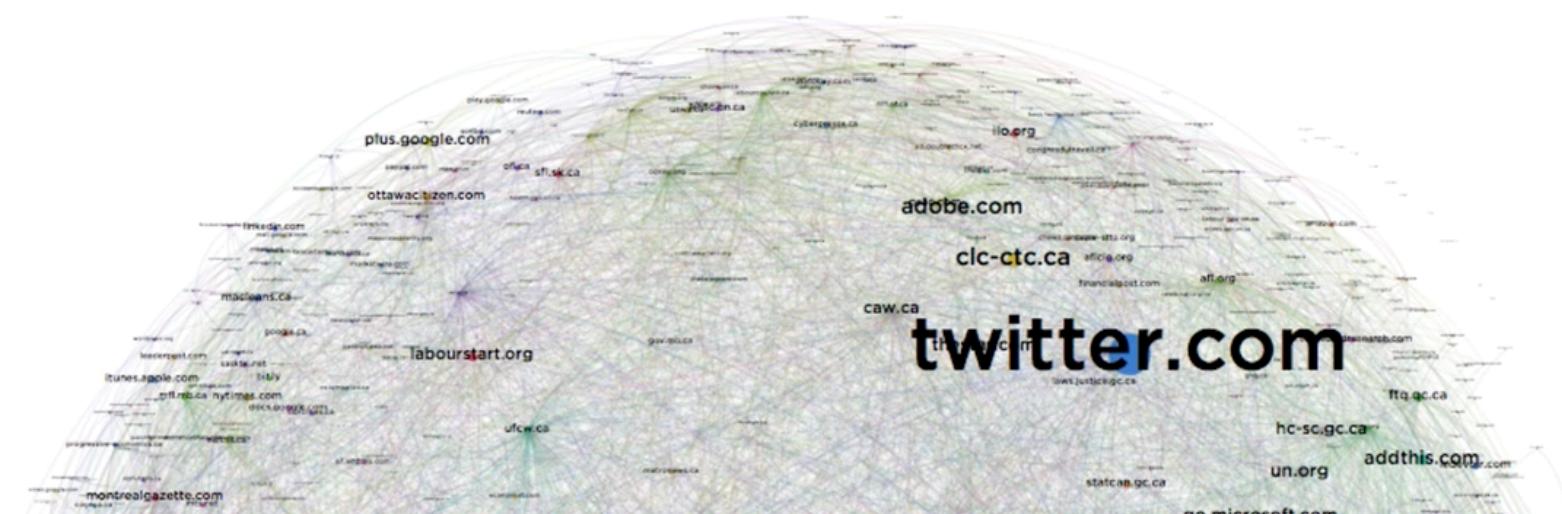
On this website, you can search web archived content from 50 political parties and political interest groups, from October 2005 to March 2015.

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.



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....

The screenshot shows a web browser window with the following details:

- Address Bar:** webarchives.ca/search?query=stephen+harper&tab=results&action=search
- Page Title:** Web Archives for Historical Research - Canadian Politics
- Header:** Search, Trends, About
- Welcome Message:** Welcome to the Web Archives for Historical Research political parties portal. Before diving in, we encourage you to visit our [about](#) page.
- Search Options:** Search, Advanced Search
- General Content Type:** html (1,085,201), other (71,691), pdf (3,947), audio (341), text (106), image (14)
- Sample Mode:** stephen harper (Search, Reset)
- Search Term(s):** stephen harper
- Crawl Years:** 2008 (443,448), 2010 (142,609), 2007 (109,236), 2006 (104,564), 2011 (83,910), 2014 (70,746)
- Results:** Results (selected), Concordance
- Page Footer:** Results 1 to 10 of 1,161,300, CSV ▾, Asc ▾

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

Gephi 0.8.2

(<http://gephi.github.io/>)

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

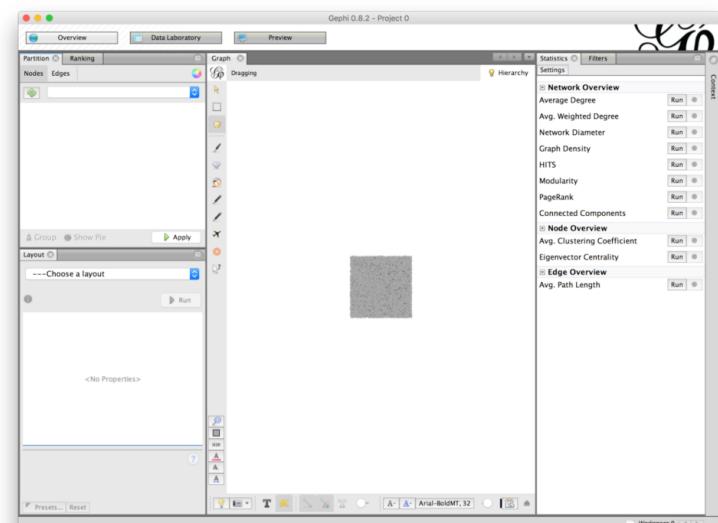
below.

Step-by-Step Walkthrough

Once you've downloaded the file, open up Gephi.

On the opening screen, you want to select “Open a Graph File...” and select the `all-links-cpp-link.graphml` file that you downloaded from our Dataverse page.

You then want to click ‘ok’ on the next page. Create a ‘new graph.’



Do you want to make this link graph yourself from our data? Read on.

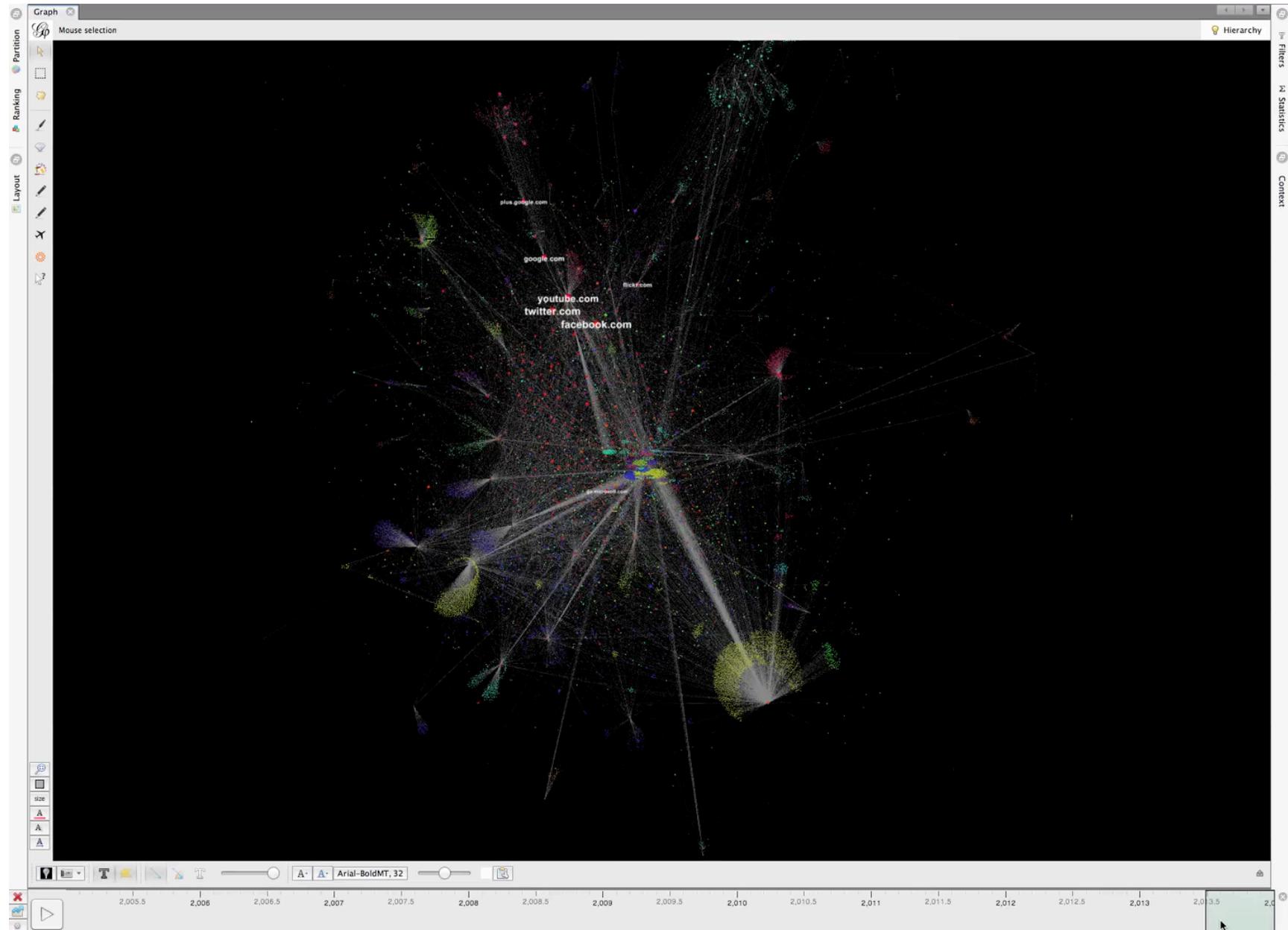
You should now see what I (nerdily) call a borg cube. That's good, because it means that the data is in there. We need to make it usable, however.

Click on the “Data Laboratory” tab at the top.

Click on “Nodes” above. When it is shaded behind it, that means that it is selected.

Click on “Copy Data to another Column,” select ID, and then select “label” on the drop

Metadata Extraction



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!



October 2008

Election Campaign - Advertisement Sites

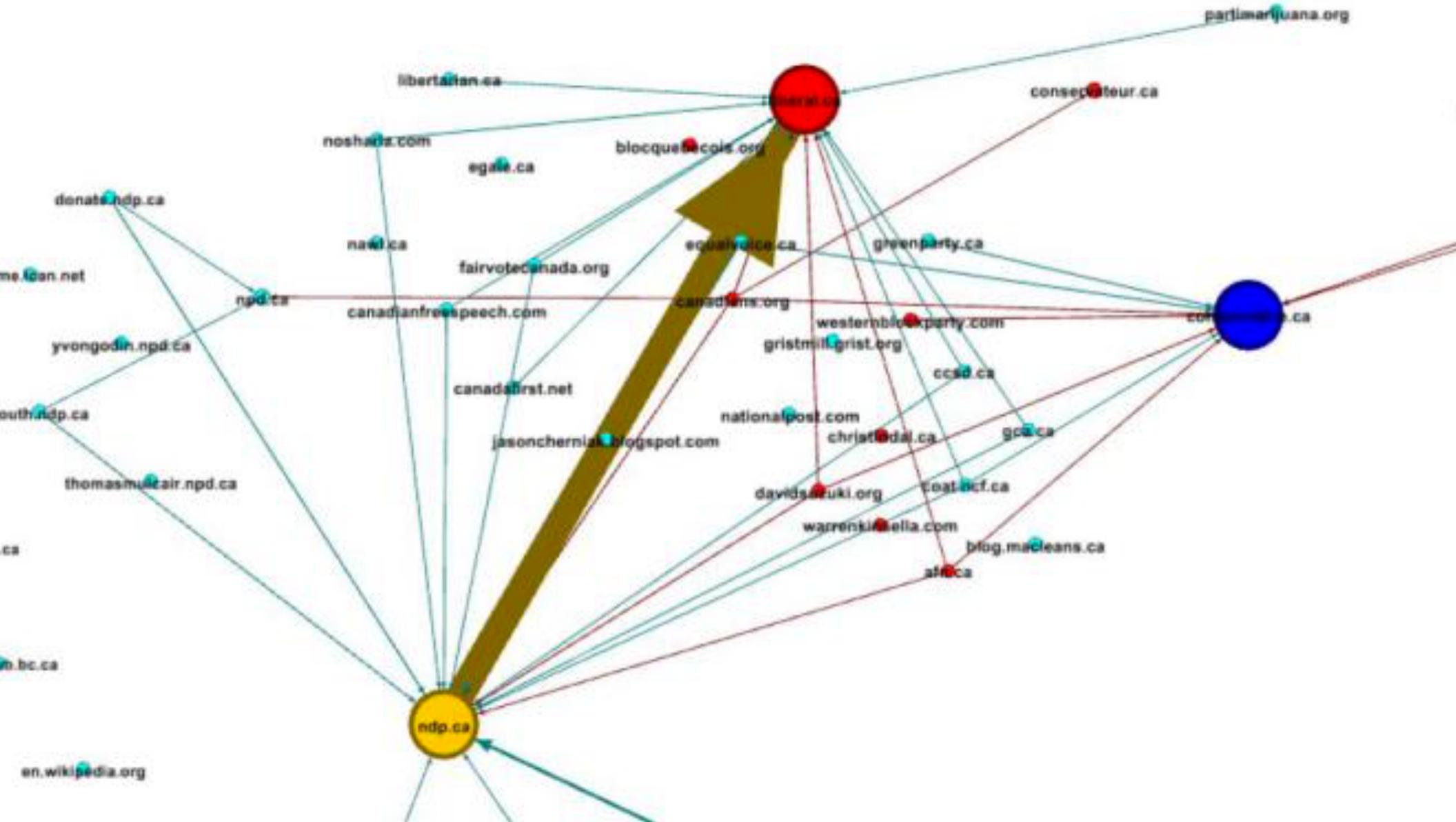


December 2008

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



2005 Canadian Federal Election



**This requires
interdisciplinary
collaboration**

Warcbase (soon to be twarcbase, too!)

- 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...

The screenshot shows a GitHub repository page for 'lintool/warcbase'. The browser title bar reads 'Home · lintool/warcbase'. The GitHub header includes the repository name, a 'This repository' button, a search bar, and links for 'Pull requests' and 'Issues'. The main content area is titled 'Home' and displays the following text:
Welcome to the warcbase wiki! Here, you'll find various pig scripts and instructions to unlock your rich web archive collections.
These pages are under active development, as of June 2015.
If you are using warcbase, we would love to hear from you. [Please let us know](#).

Note: many of these tutorials currently assume a working knowledge of a Unix line environment. For a conceptual and practical introduction, please see Ian MacLennan and James Baker's "Introduction to the Bash Command Line" at the *Programming Historian*: <http://programminghistorian.org/lessons/intro-to-bash>.

Getting Started?

This is still actively under development with several features in the pipeline (no

Warcbase

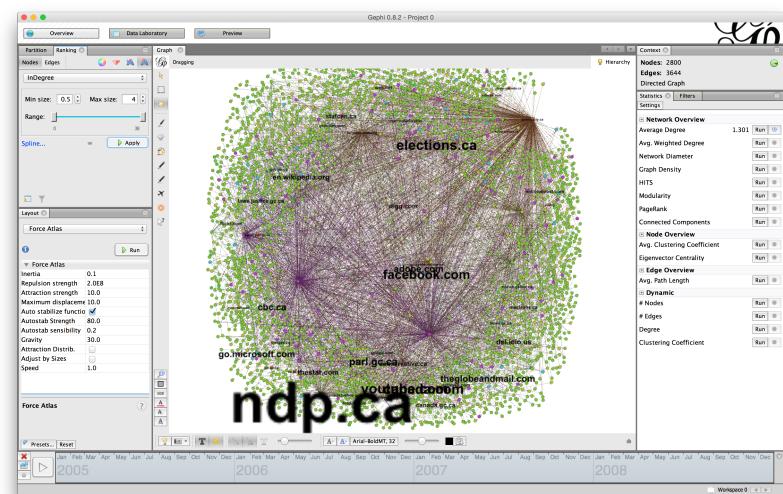
An open-source platform for managing web archives

<http://warcbase.org>

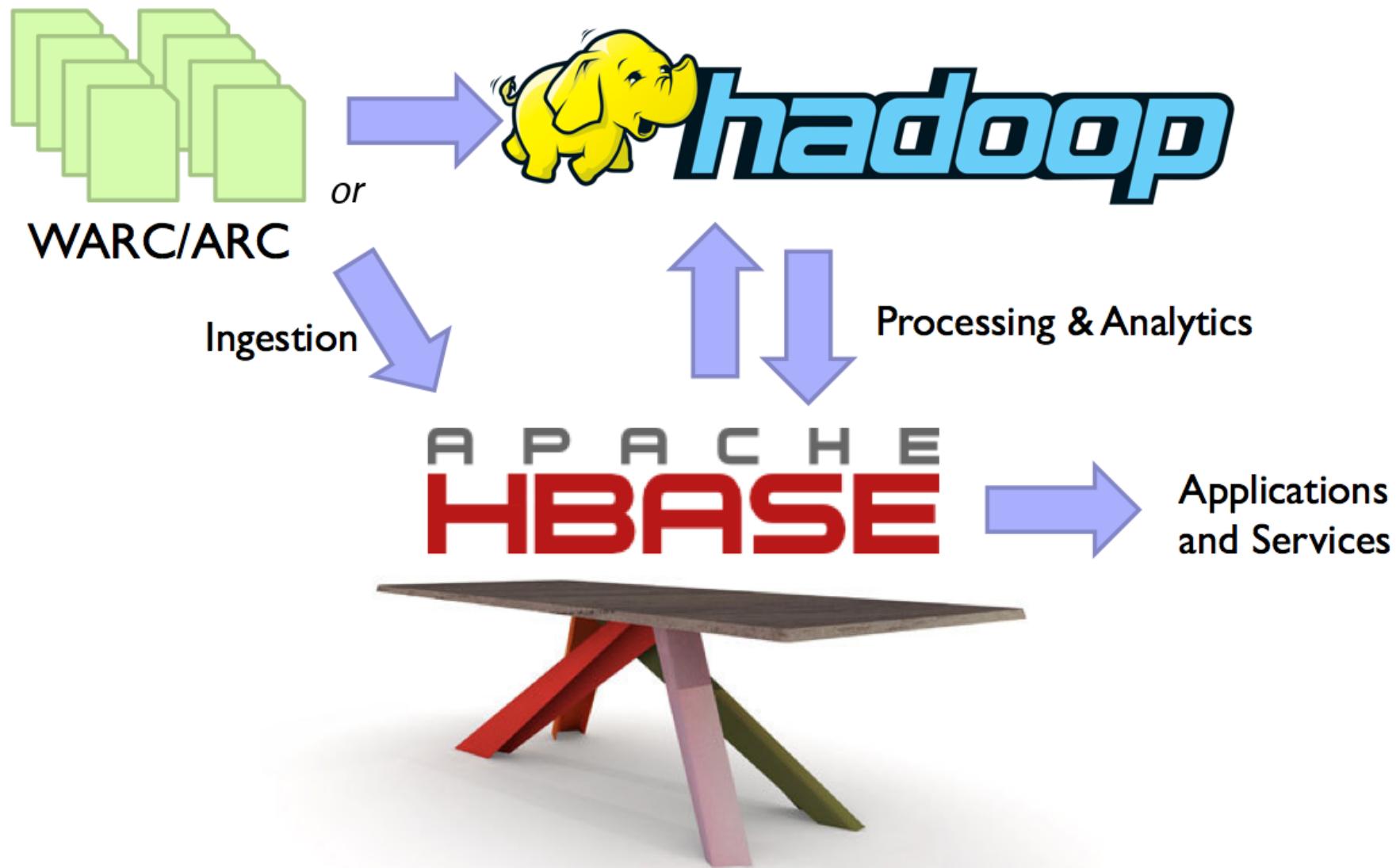
Two main facets

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

Funded by Mellon and (now) SSHRC.



Warcbase



Warcbase

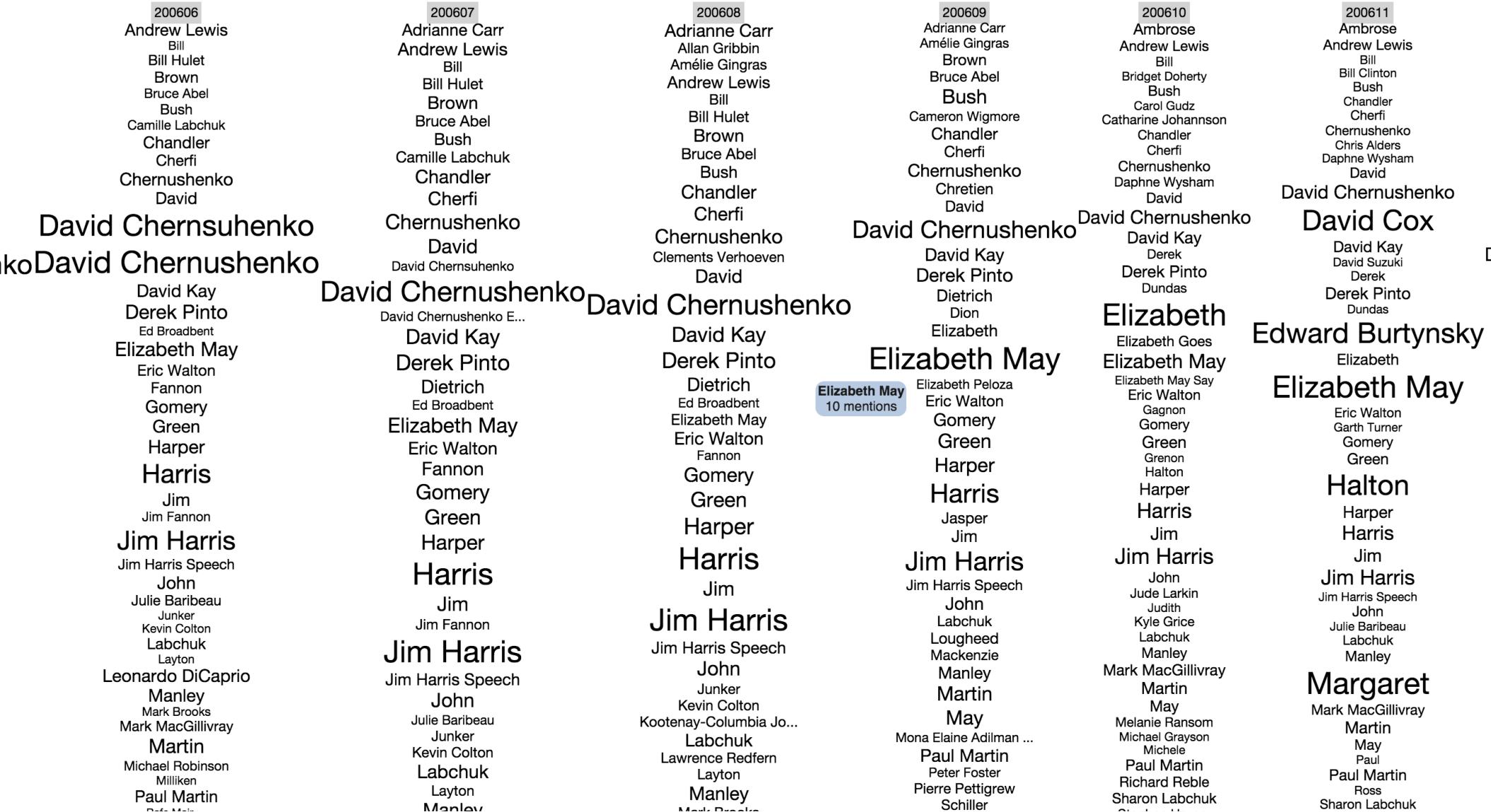
- Framework for distributed storage and distributed processing of very big data
- Scalable
 - From Raspberry Pi, to laptop, to powerful desktop, to single-node beefy server, to cluster
- Potentially very powerful
 - *Trantor*: 1.2PB of disk, 25 compute nodes (each w/ 128GB memory, 2×6-core Intel Xeon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



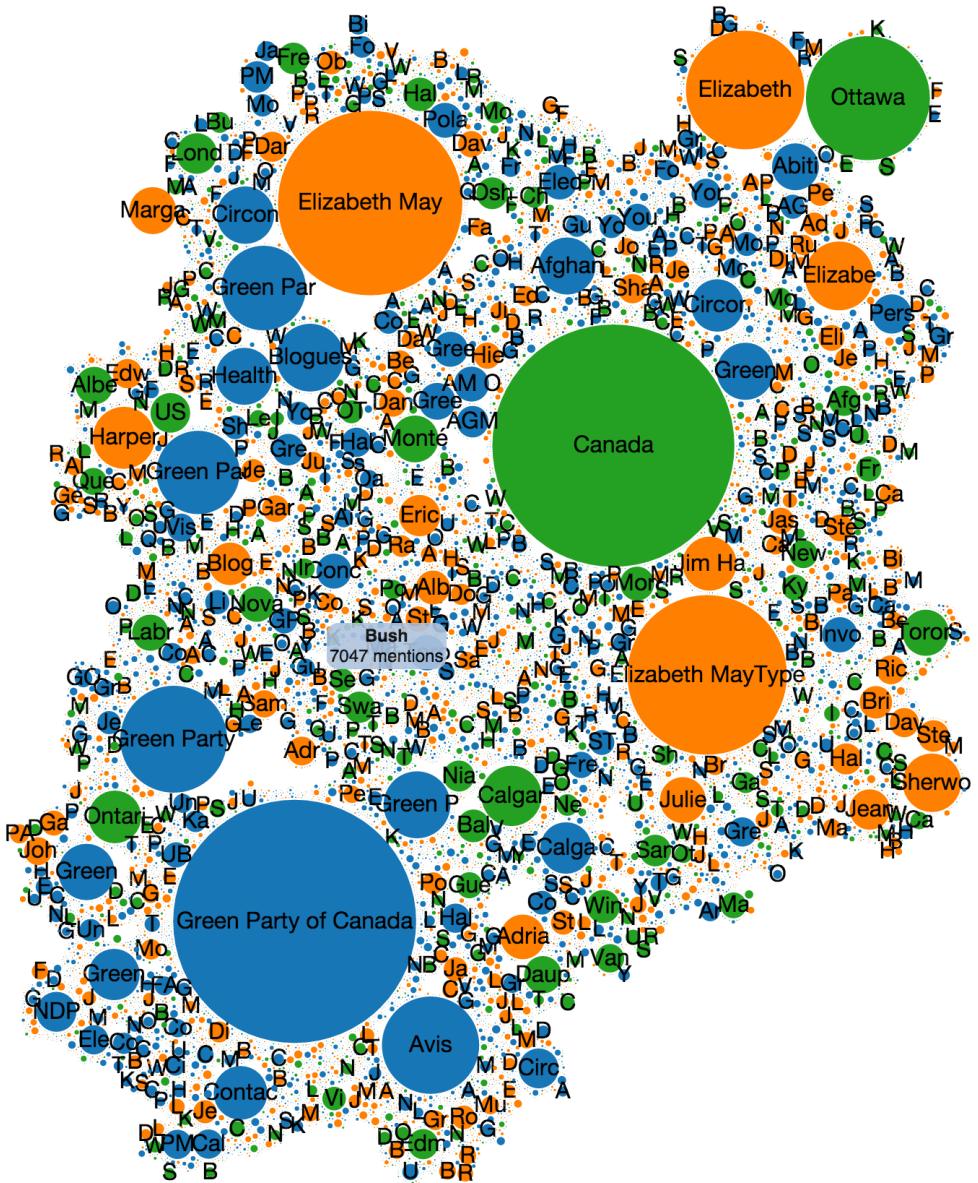
Extract all Plain Text

```
1. i2millig@rho: ~/derivatives/cpp.all.plaintext (ssh)
Python      bash      bash      bash      i2millig@rho:... i2millig@rho:... bash
(20060222,liberal.ca,http://liberal.ca/bio_e.aspx?id=35049,Liberal Party of Canada HOME THE TEAM THE P
ARTY MEDIA CENTRE COMMISSIONS YOUR RIDING      Omar Alghabra www.omaralghabra.com Home > Mississauga--E
rindale Riding Map (PDF) Omar Alghabra came to Canada at a very young age, and immediately knew Canada
was his home. He was first elected 2006 as the Member of Parliament for Mississauga-Erindale. Mr. Al
ghabra is an experienced entrepreneur. For the past six years, he has worked for a large multinational
corporation, carrying out different responsibilities including quality assurance, project management, s
ales, contract management and management of a complete department handling a global mandate. Mr. Alghab
ra is an active member of his community. He is the former National President of the Canadian Arab Feder
ation (2004-2005) and a former member of the Community Editorial Board for the Toronto Star. (2003-2004
). Mr. Alghabra is currently a member of the Diversity Council for General Electric Canada and is activ
e in Junior Achievement for the Toronto Region. He was a member of the Multicultural Inter-Agency of Pe
ople from 2001 to 2002. Mr. Alghabra has a degree in Mechanical Engineering from Ryerson University and a
Masters in Business Administration (MBA) from York University.    Omar Alghabra 790 Burnamthorpe West,
Unit 10 905-276-2806   info@omaralghabra.ca Riding President Elias Hazineh Send an email
Home | News | Your Riding | Contact Us | français This website is the property of the
Liberal Party of Canada and may not be reproduced in whole or in part without express written permission.
© Liberal Party of Canada 2006. All rights reserved. Authorized by the registered agent for the Libe
ral Party of Canada. Privacy Policy)
(20060222,liberal.ca,https://liberal.ca/news_e.aspx?id=11470,Liberal.ca HOME THE TEAM THE PARTY MEDIA C
ENTRE COMMISSIONS YOUR RIDING  Celebrating our National Flag  February 15, 2006 February 15 is Nationa
l Flag Day in Canada, which marks the 41st anniversary of the first raising of the maple leaf flag on P
arliament Hill. Today is a celebration of our shared values, common citizenship and sense of pride in t
his great country we call home. The Canadian flag is one of the most recognizable symbols in the world
and flies proudly to remind us all of who we are and where we come from. The maple leaf's symbolic orig
ins date back to the beginning of our nation's history, while the red and white bars on the flag repres
ent strength and unity. Canada's flag was adopted in 1964 under the courageous leadership of Liberal Pr
ime Minister Lester B. Pearson. The idea of changing the Red Ensign which featured the Britain's Union
Jack, was very controversial at the time, with the Conservative Party strongly opposed to changing the
status quo. Facing strong Conservative resistance in the House of Commons, Pearson's minority governme
nt fought hard in the name of national unity and Canada's multicultural future to make the new flag a r
eality. In an impassioned speech to the House of Commons, Pearson said: "Mr. Speaker, it is for this g
eneration, for this Parliament, to give them and to give us all a common flag; a Canadian flag which, w
hile bringing together but rising above the landmarks and milestones of the past, will say proudly to t
he world and to the future: I stand for Canada." Thanks to Pearson's courageous leadership, Canadians a
cross this great nation celebrate our flag and what it stands for – a country and a citizenship that ar
e the envy of the world.
Home | News | Your Riding | Contact Us | fran
cais This website is the property of the Liberal Party of Canada and may not be reproduced in whole or
```

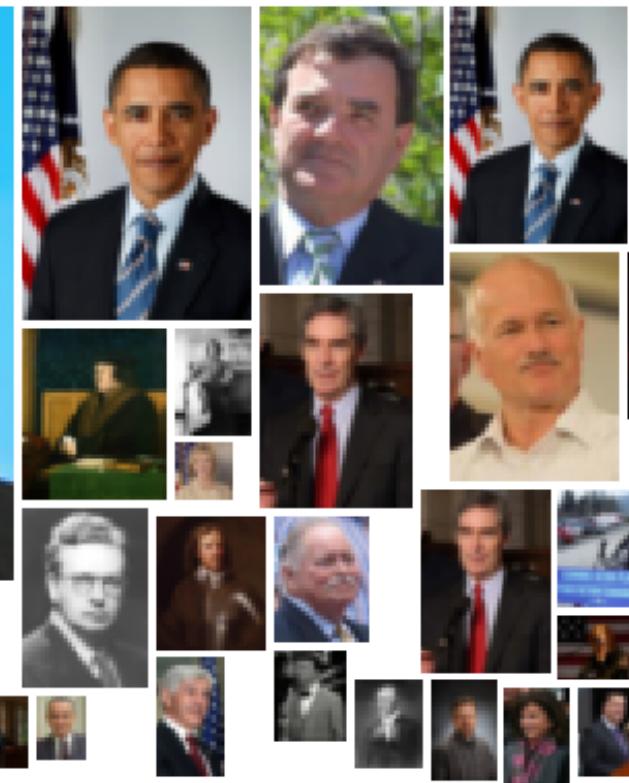
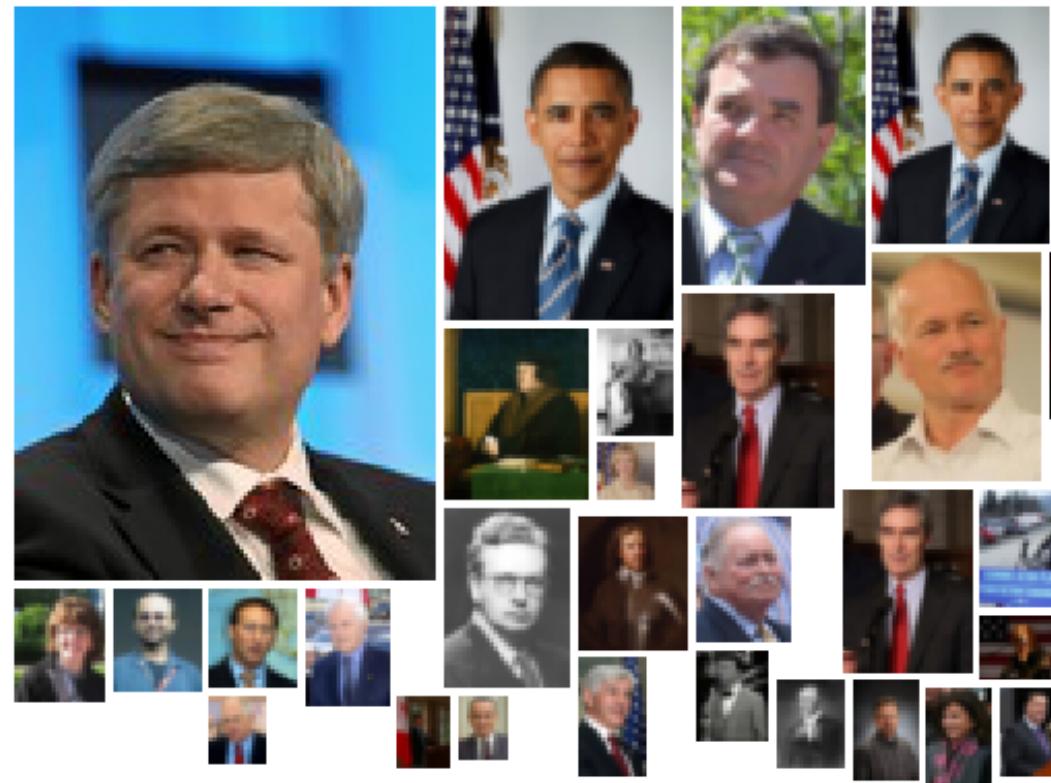
Extract Entities



Extract Entities



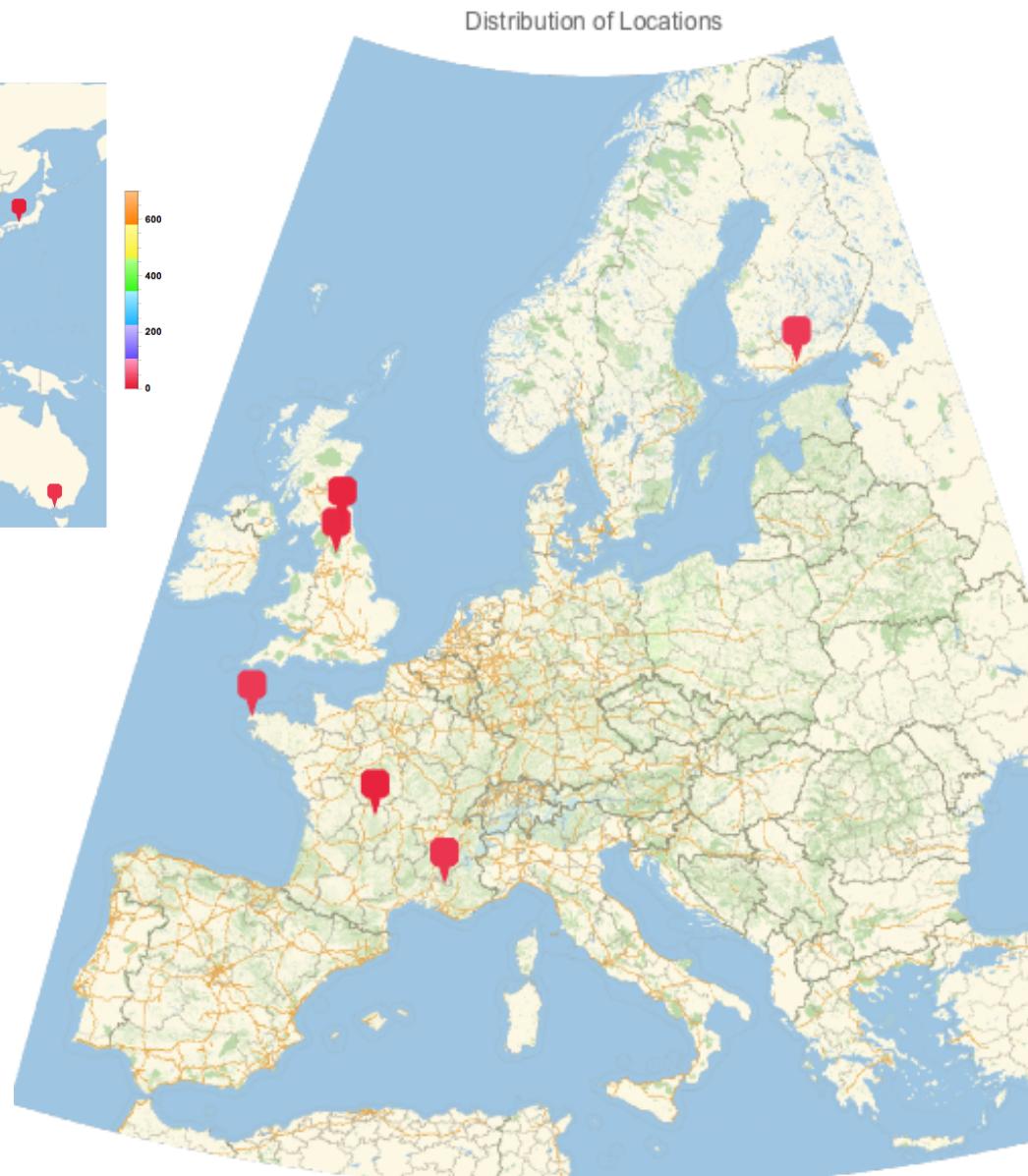
Extract Entities



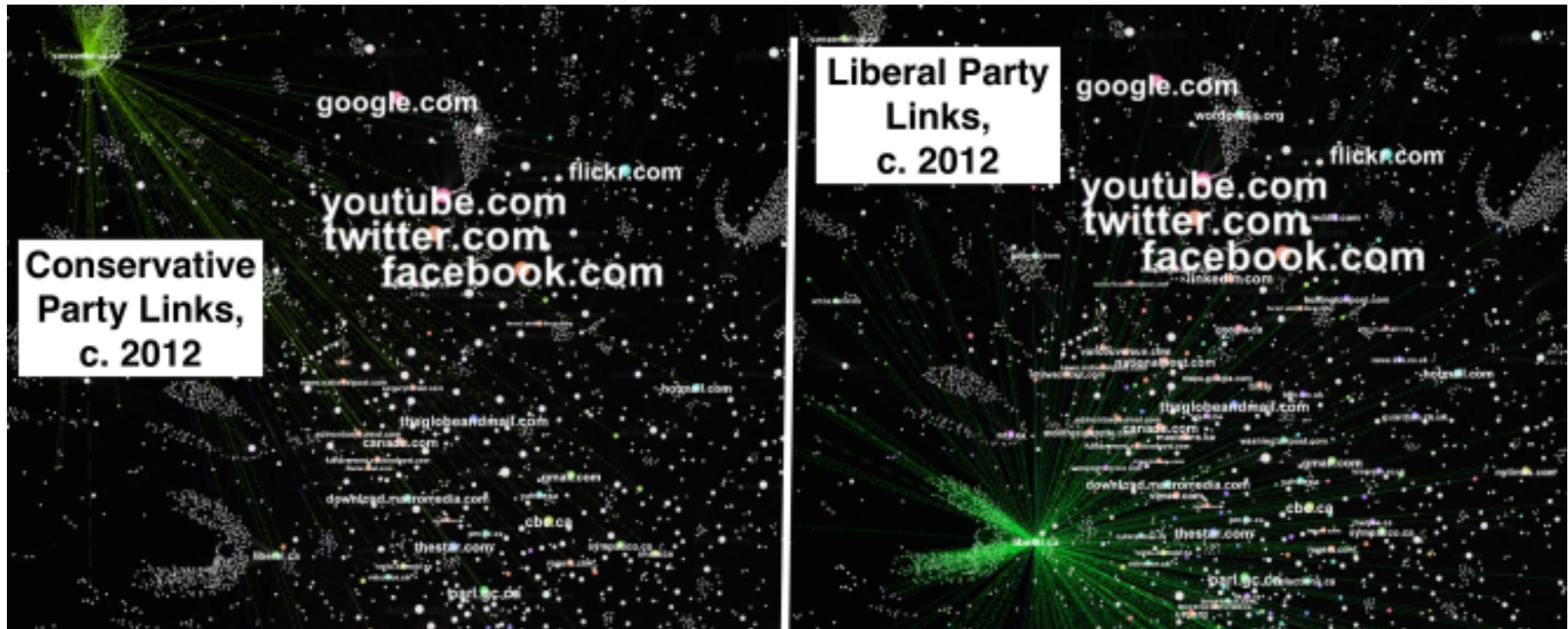
Extract Entities



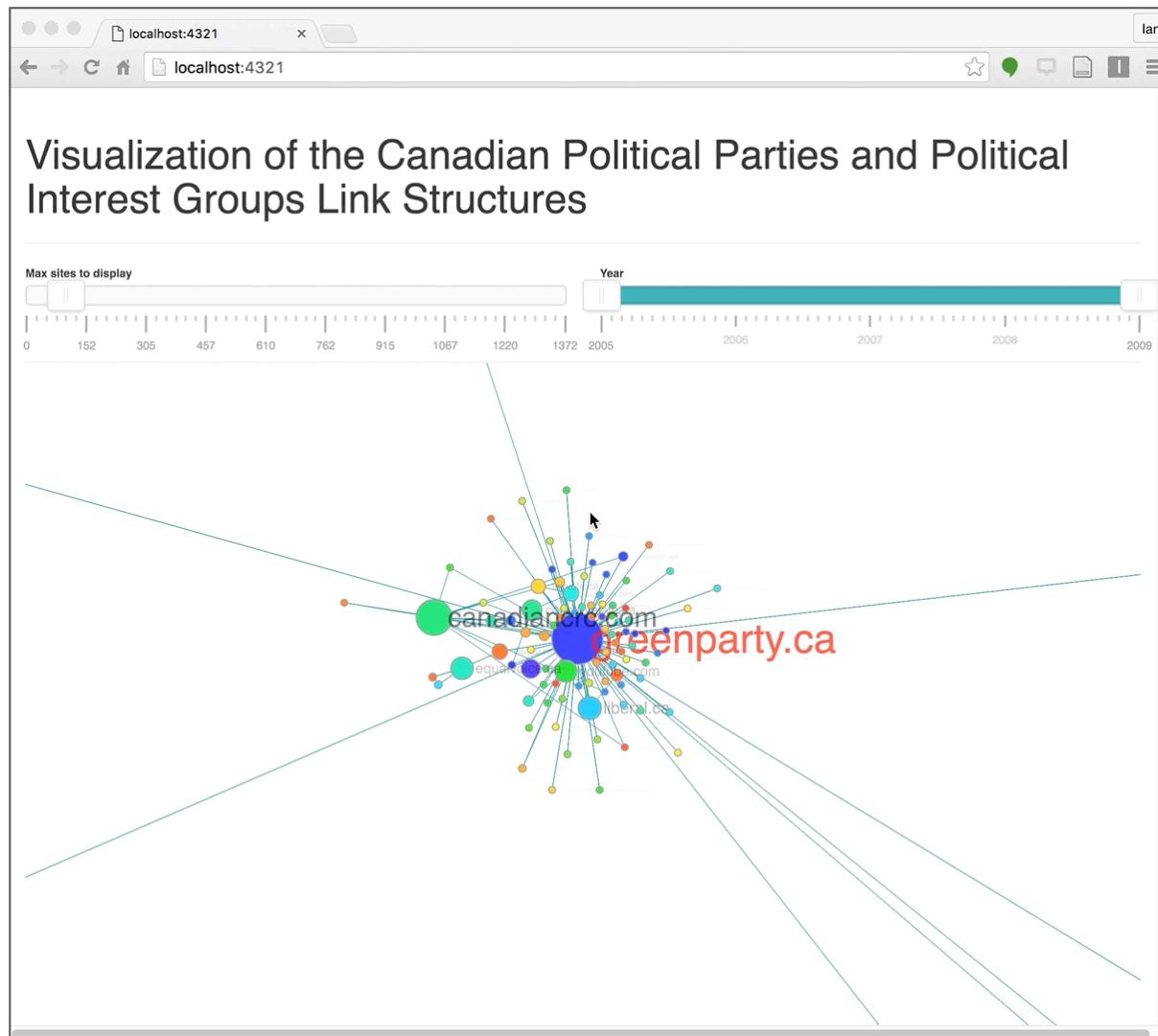
```
In[95]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]  
Out[95]= {Canada, Calgary, Colombia, Montreal, $Failed, Ontario, Canada, Afghanistan,  
Ottawa, Manitoba, Canada, British Columbia, Canada, Toronto, Nova Scotia, Canada,  
Saskatchewan, Canada, Quebec, Canada, Alberta, Canada, 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,
```



Extract Links/Gephi Connector



Or D3.js link networks in browser



Bringing it all together
in a notebook
environment



Spark Notebook

localhost:9000

Apps Keep Lib ./ HistD GM W MMA RSS Globe Global Online Enrollment Historical Statistics Other Bookmarks



Files Running Clusters

To import a notebook, drag the file onto the listing below or [click here](#).

New ▾



adam

anomalyDetection

cassandra

core

graphx

misc

mllib

sql

streaming

tachyon

viz

Spark Notebook Demo

Duplicate

Shutdown

TTOW

Duplicate

Delete

Tachyon Test

Duplicate

Delete

Untitled1

Duplicate

Delete

Untitled2

Duplicate

Delete

Web Archives 2015, Demo

Duplicate

Delete

All walkthroughs at:
docs.warcbase.org



compute • calcul
CANADA

**Coming soon to
Canadian web archive
collections near you!**

Follow along w/ GeoCities?

Exploring the GeoCities Web Archive

https://ianmilligan.ca/2016/02/14/exploring-the-geocities-web-archive-with-warbase-spark-getting-started/

My Sites Reader and research into web archival use more generally.

Assistant Professor
Department of History
University of Waterloo
i2milligan@uwaterloo.ca

CC BY SA

ianmilligan.ca by Ian Milligan is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Step One: Ingesting the Data

Once the hard drives arrived, it was fun to watch the data populate our server as Nick supervised the time-consuming job of moving over 4TB of data from two hard drives onto our server at York University.

Ian Milligan (@ianmilligan) 11 Feb
Wowow, @ruebot and I now have a "geocities" directory on our server. Filling with WARCs and WATs. Thanks @jefferson_bail! #WebArchiving

nick ruest (@ruebot) Follow
. @ianmilligan1 @jefferson_bail copy copy copy pic.twitter.com/gsj3r1HxWl
9:49 AM - 11 Feb 2016 · Toronto, Ontario, Canada



GitHub

Recent Blog Posts

Exploring the GeoCities Web Archive with Warcbase & Spark: Getting Started 14 February 2016



```
1. i2mllig@rho: /mnt/vol1/data_sets/geocities/warc$ (ssh)
bash                                bash                                i2mllig@rho: /mnt/vol1/data...
GEOCITIES-20091029114236-00191-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029115416-00171-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029123634-00172-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029138439-00173-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029134536-00174-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029140344-00192-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029141553-00193-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029141726-00175-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029144445-00176-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029152151-00177-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029160824-00178-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029164941-00179-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029165037-00194-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029170431-00195-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029171605-00180-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029174154-00181-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029180818-00182-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029182725-00183-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029185858-00184-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029193728-00185-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029194541-00196-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029195911-00197-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029202041-00186-crawling88.us.archive.org.warc.gz
GEOCITIES-20091029221340-00198-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029222459-00199-ia400110.us.archive.org.warc.gz
GEOCITIES-20091030021147-00197-ia400103.us.archive.org.warc.gz
GEOCITIES-20091030021444-00198-ia400103.us.archive.org.warc.gz
GEOCITIES-20091030022413-00171-ia400104.us.archive.org.warc.gz
i2mllig@rho: /mnt/vol1/data_sets/geocities/warc$ du -h
4.1T .
i2mllig@rho: /mnt/vol1/data_sets/geocities/warc$
```

```
ianmilligan@ians-MacBook-Pro:~$ rho
i2millig@rho.library.yorku.ca's password:
Welcome to Ubuntu 14.04.2 LTS (GNU/Linux 3.13.0-32-generic x86_64)

 * Documentation: https://help.ubuntu.com/

System information as of Mon Mar  7 13:43:20 EST 2016

System load:  0.99          Users logged in:      1
Usage of /:   34.7% of 744.67GB  IP address for em1:    130.63.180.18
Memory usage: 16%
Swap usage:   6%           IP address for em2:    10.0.0.18
Processes:    359          IP address for docker0: 172.17.0.1

Graph this data and manage this system at:
  https://landscape.canonical.com/

242 packages can be updated.
130 updates are security updates.

Last login: Mon Mar  7 13:43:21 2016 from 38.123.136.254
i2millig@rho:~$ ./spark-1.5.1/bin/spark-shell --jars ~/warcbase/target/warcbase-0.1.0-SNAPSHOT-fatjar.jar
WARN NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Welcome to

    / _/_/ \
    \ \_\_ - \ \_` /_ /'_ /` /
    /_/_/ . / \_,_/_/ /_ / \_\ \
    /_/_/          version 1.5.1

Using Scala version 2.10.4 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_45)
Type in expressions to have them evaluated.
Type :help for more information.
WARN MetricsSystem - Using default name DAGScheduler for source because spark.app.id is not set.
Spark context available as sc.
SQL context available as sqlContext.

scala> :paste
// Entering paste mode (ctrl-D to finish)

import org.warcbase.spark.matchbox._
import org.warcbase.spark.rdd.RecordRDD._

val r =
RecordLoader.loadWarc("/mnt/vol1/data_sets/geocities/warcs/GEOCITIES-20090808133634-04399-crawling08.us.archive.org.warc.gz", sc)
.keepValidPages()
.map(r => ExtractTopLevelDomain(r.getUrl))
.countItems()
.take(10)

// Exiting paste mode, now interpreting.

INFO WacWarcInputFormat - Loading file:/mnt/vol1/data_sets/geocities/warcs/GEOCITIES-20090808133634-04399-crawling08.us.archive.org.warc.g
z
import org.warcbase.spark.matchbox._
import org.warcbase.spark.rdd.RecordRDD._
r: Array[(String, Int)] = Array((geocities.com,3748), (www.geocities.com,240), (www.myfilehut.com,12), (asiarooms.com,7), (us.geocities.com,6), (www.theginge.com,3), (www.angelfire.com,3), (images.quizilla.com,3), (pub28.bravenet.com,3), (ss.webring.yahoo.com,2))

scala> 
```

Extract all URLs

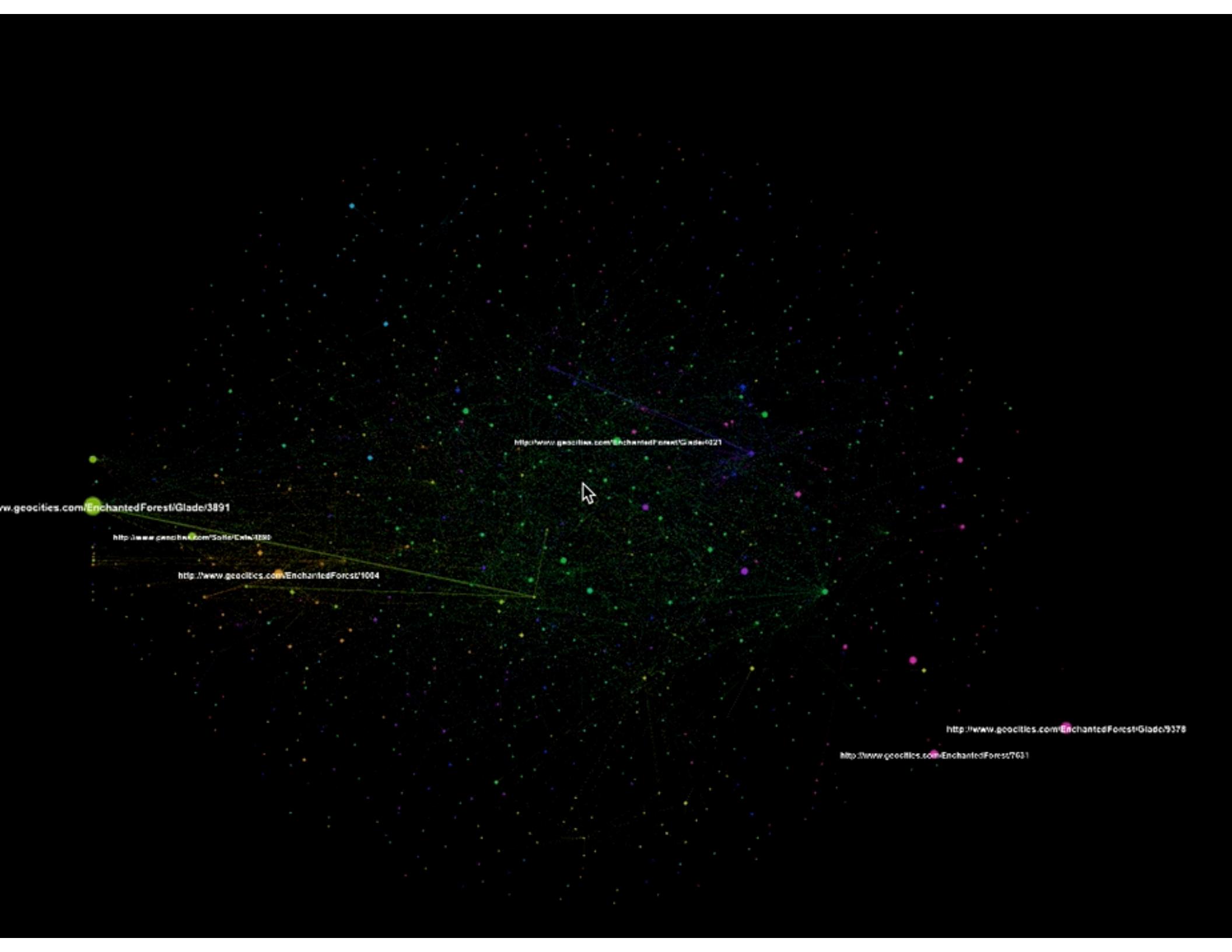
```
1 import org.warcbase.spark.matchbox._  
2 import org.warcbase.spark.rdd.RecordRDD._  
3  
4 val r = RecordLoader.loadWarc("/mnt/vol1/data_sets/geocities/  
      warcs", sc)  
5 .keepValidPages()  
6 .map(r => r.getUrl)  
7 .saveAsTextFile("/mnt/vol1/derivative_data/geocities/url-list")
```

9.9 GB text file, 186,761,346 URLs

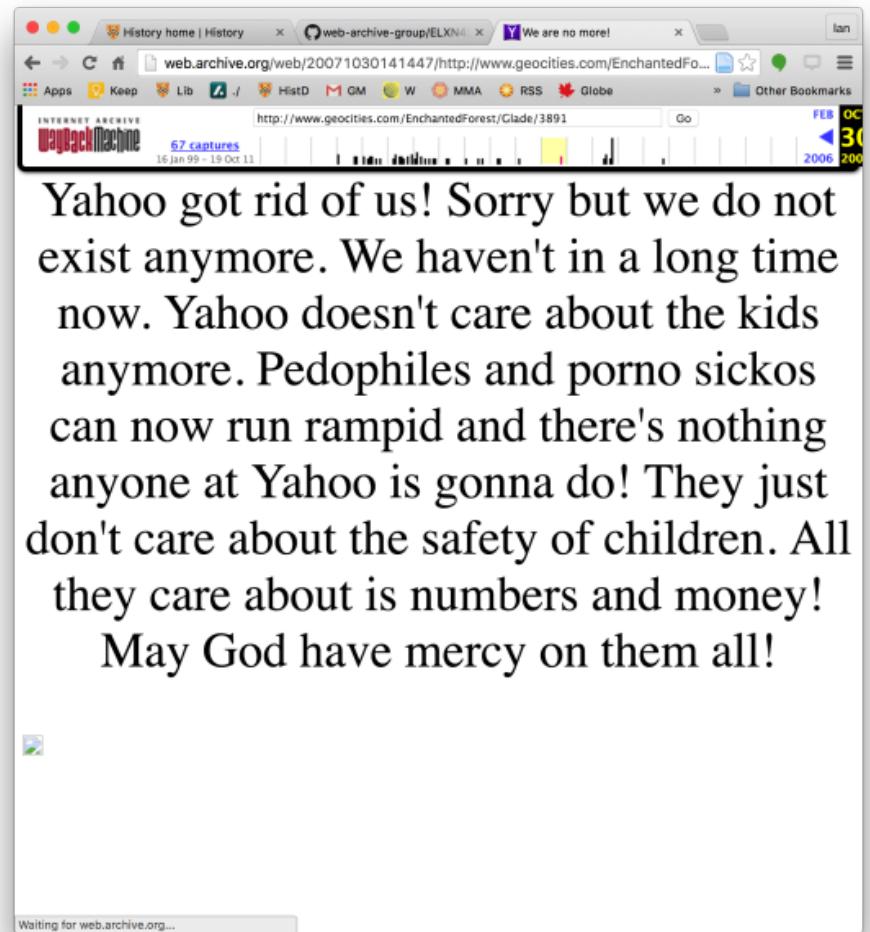
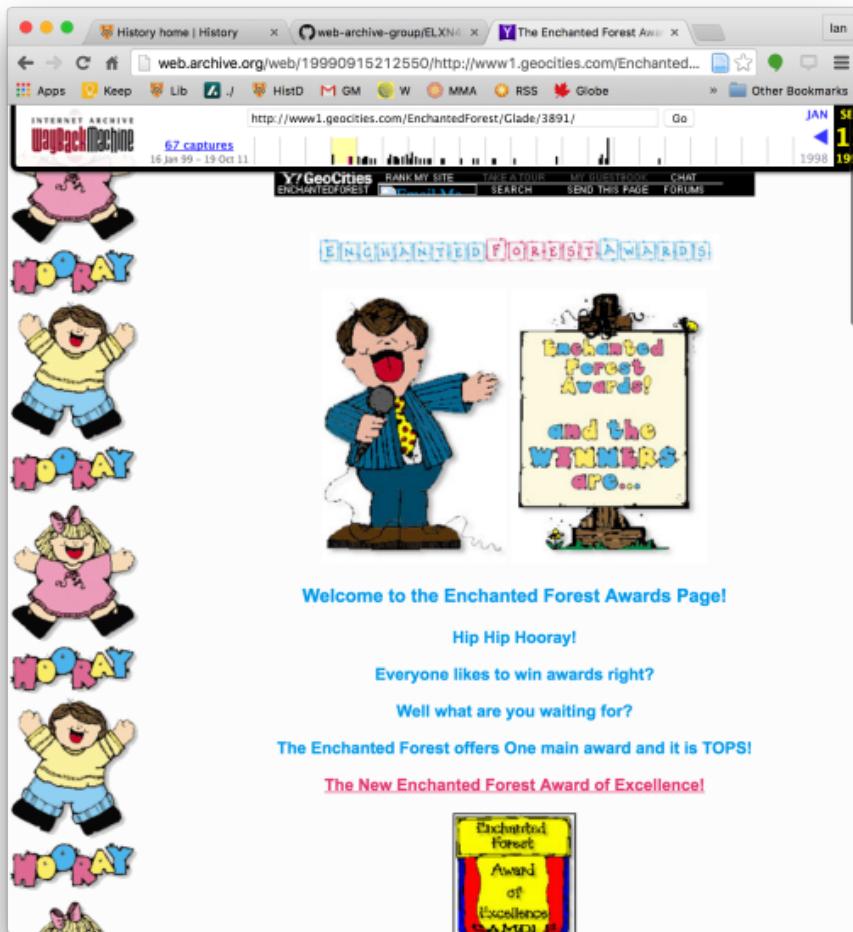
Extract a Link Graph

```
1 import org.warcbase.spark.matchbox.{ExtractTopLevelDomain,  
    ExtractLinks, RecordLoader}  
2 import org.warcbase.spark.rdd.RecordRDD._  
3  
4 RecordLoader.loadArc("/mnt/vol1/data_sets/geocities/warcs/*", sc)  
5 .keepValidPages()  
6 .map(r => (r.getCrawldate, ExtractLinks(r.getUrl, r.  
    getContentString)))  
7 .flatMap(r => r._2.map(f => (r._1, ExtractTopLevelDomain(f._1).  
    replaceAll("^\\s*www\\.", ""), ExtractTopLevelDomain(f._2).  
    replaceAll("^\\s*www\\.", ""))))  
8 .filter(r => r._2 != "" && r._3 != "")  
9 .countItems()  
10 .filter(r => r._2 > 5)  
11 .saveAsTextFile("/mnt/vol1/data_sets/geocities/geocities.  
    sitelinks")
```





Label	▼ PageRank	In-De...	Out-D...
http://www.geocities.com/EnchantedForest/Glade/3891	0.008	62	0
http://www.geocities.com/EnchantedForest/Glade/9378	0.005	2	2
http://www.geocities.com/EnchantedForest/1004	0.005	23	14
http://www.geocities.com/EnchantedForest/7631	0.004	2	1
http://www.geocities.com/SoHo/Cafe/4690	0.004	32	0
http://www.geocities.com/EnchantedForest/Glade/4021	0.004	26	0
http://www.geocities.com/TheTropics/Paradise/4079	0.003	31	0
http://www.geocities.com/RainForest/Vines/4892	0.003	3	3
http://www.geocities.com/EnchantedForest/3278	0.003	15	0
http://www.geocities.com/EnchantedForest/3696	0.003	15	0
http://www.geocities.com/EnchantedForest/Dell/5914	0.003	13	0
http://www.geocities.com/EnchantedForest/1469	0.003	5	20
http://www.geocities.com/EnchantedForest/Tower/9644	0.003	3	13



Where to learn?



The screenshot shows a web browser window with the title bar "About the Programming His x" and the URL "programminghistorian.org". The page content includes the site's name, navigation links, and a section about the project.

The Programming Historian

About · Lessons · Contribute · Project Team · Blog

About the Programming Historian



The Programming Historian offers novice-friendly, peer-reviewed tutorials that help humanists learn a wide range of digital tools, techniques, and workflows to facilitate their research.

We regularly publish new lessons, and we always welcome proposals for new lessons on any topic. Our editorial mentors will be happy to work with you throughout the lesson writing process. If you'd like to be a reviewer or if you have suggestions to make *Programming Historian* a more useful resource, please see our [Contribute](#) page.

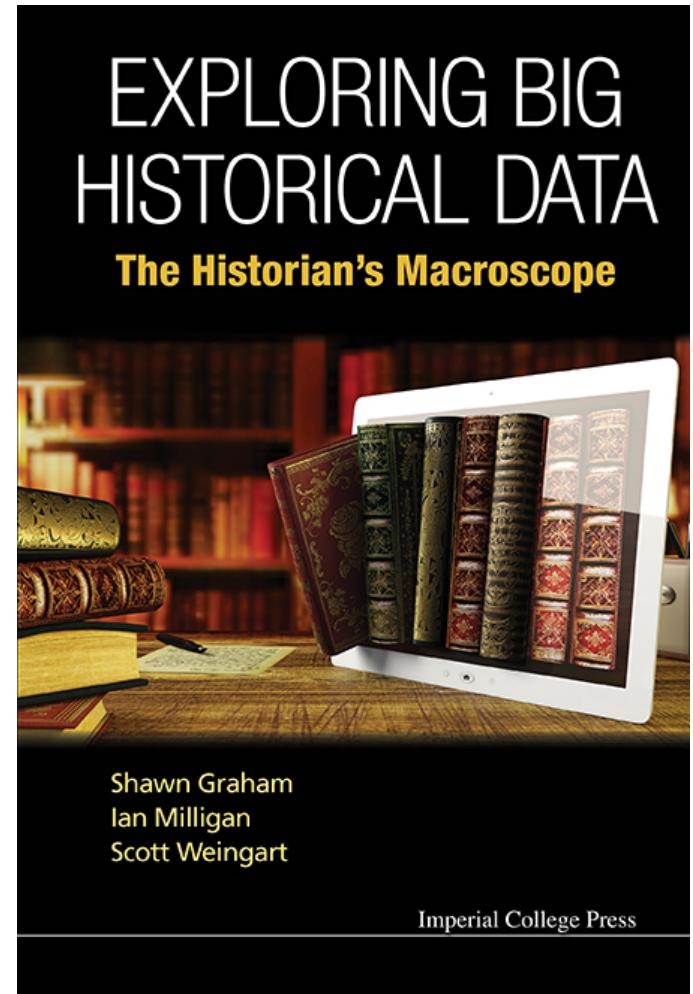
Our [Project Team](#) and peer reviewers work collaboratively with authors to craft tutorials that illustrate fundamental digital and programming principles and

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



In-Person Events

- Hackathons!
- **Toronto, ON.** March 2016 (SSHRC/NSF supported).
- **Hannover, Germany.** May 2016 (WebSci '16).
- **Washington DC**, June 2016 (SSHRC/NSF travel funding for graduate students/PDFs).



**... 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.**



Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada



compute • calcul
CANADA



UNIVERSITY OF
WATERLOO

Thanks very much!

Questions?

Ian Milligan
Assistant Professor
@ianmilligan1



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History