Capturing the Web Today for Tomorrow

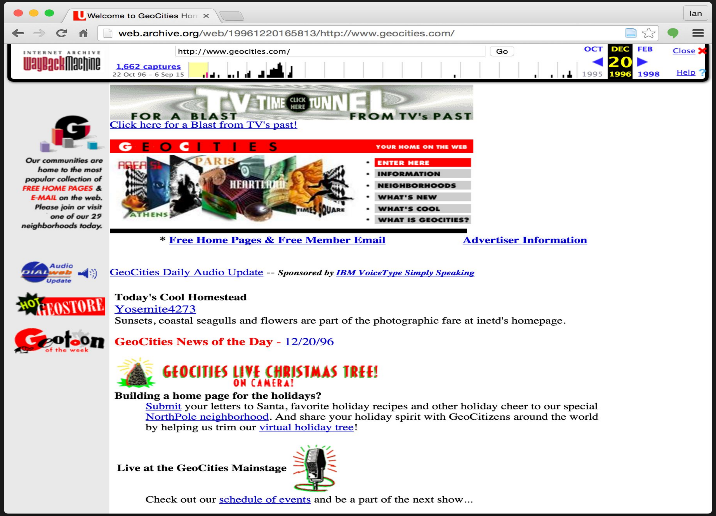
Innovations in capturing and analyzing social media and websites for the new scholarly record

UNIVERSITY OF WATERLOO

Nick Ruest (@ruebot)
Ian Milligan (@ianmilligan1)

YORK UNIVERSITÉ UNIVERSITY

We have a **problem** facing our collective cultural heritage.



Click here for a Blast from TV's past!

Our communities are home to the most popular collection of FREE HOME PAGES & E-MAIL on the web. Please join or visit one of our 29 neighborhoods today.

* Free Home Pages & Free Member Email

Advertiser Information

GeoCities Daily Audio Update -- *Sponsored by IBM VoiceType Simply Speaking*

Today's Cool Homestead

Yosemite4273

Sunsets, coastal seagulls and flowers are part of the photographic fare at inetd's homepage.

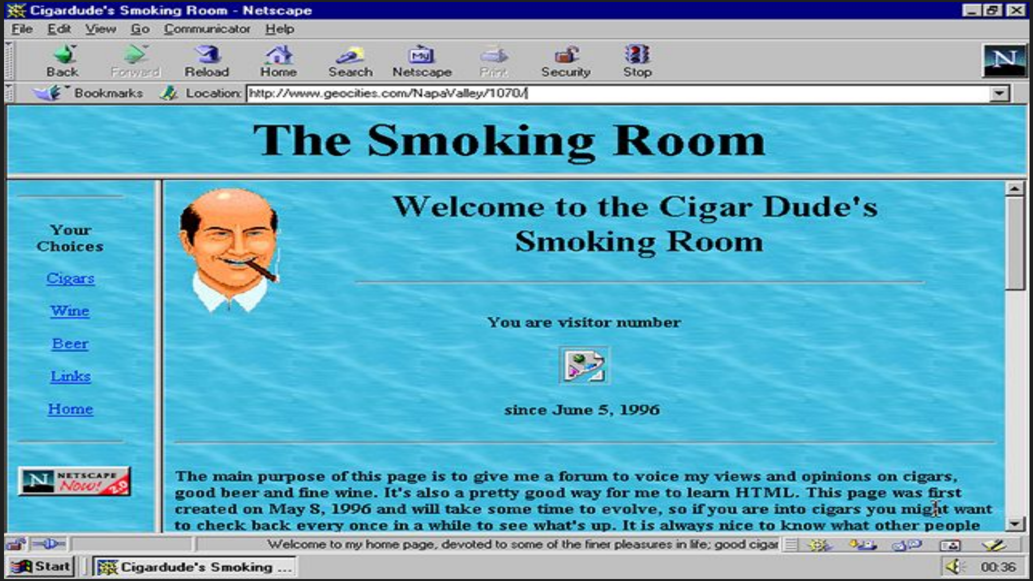
GeoCities News of the Day - 12/20/96

Building a home page for the holidays?

Submit your letters to Santa, favorite holiday recipes and other holiday cheer to our special NorthPole neighborhood. And share your holiday spirit with GeoCitizens around the world by helping us trim our virtual holiday tree!

Live at the GeoCities Mainstage

Check out our schedule of events and be a part of the next show...



The Smoking Room

Welcome to the Cigar Dude's Smoking Room

Your Choices

Cigars

Wine

Beer

Links

Home

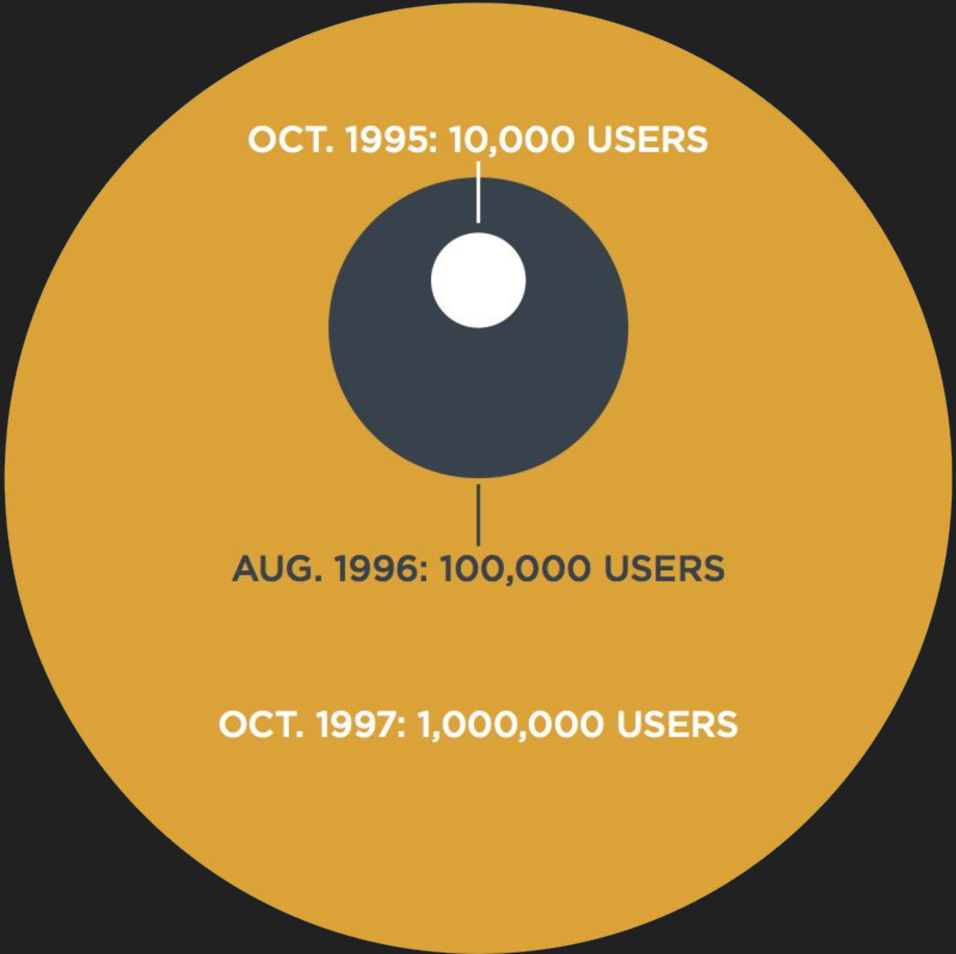
You are visitor number

since June 5, 1996

The main purpose of this page is to give me a forum to voice my views and opinions on cigars, good beer and fine wine. It's also a pretty good way for me to learn HTML. This page was first created on May 8, 1996 and will take some time to evolve, so if you are into cigars you might want to check back every once in a while to see what's up. It is always nice to know what other people

Welcome to my home page, devoted to some of the finer pleasures in life; good cigar

GEOCITIES USERS:



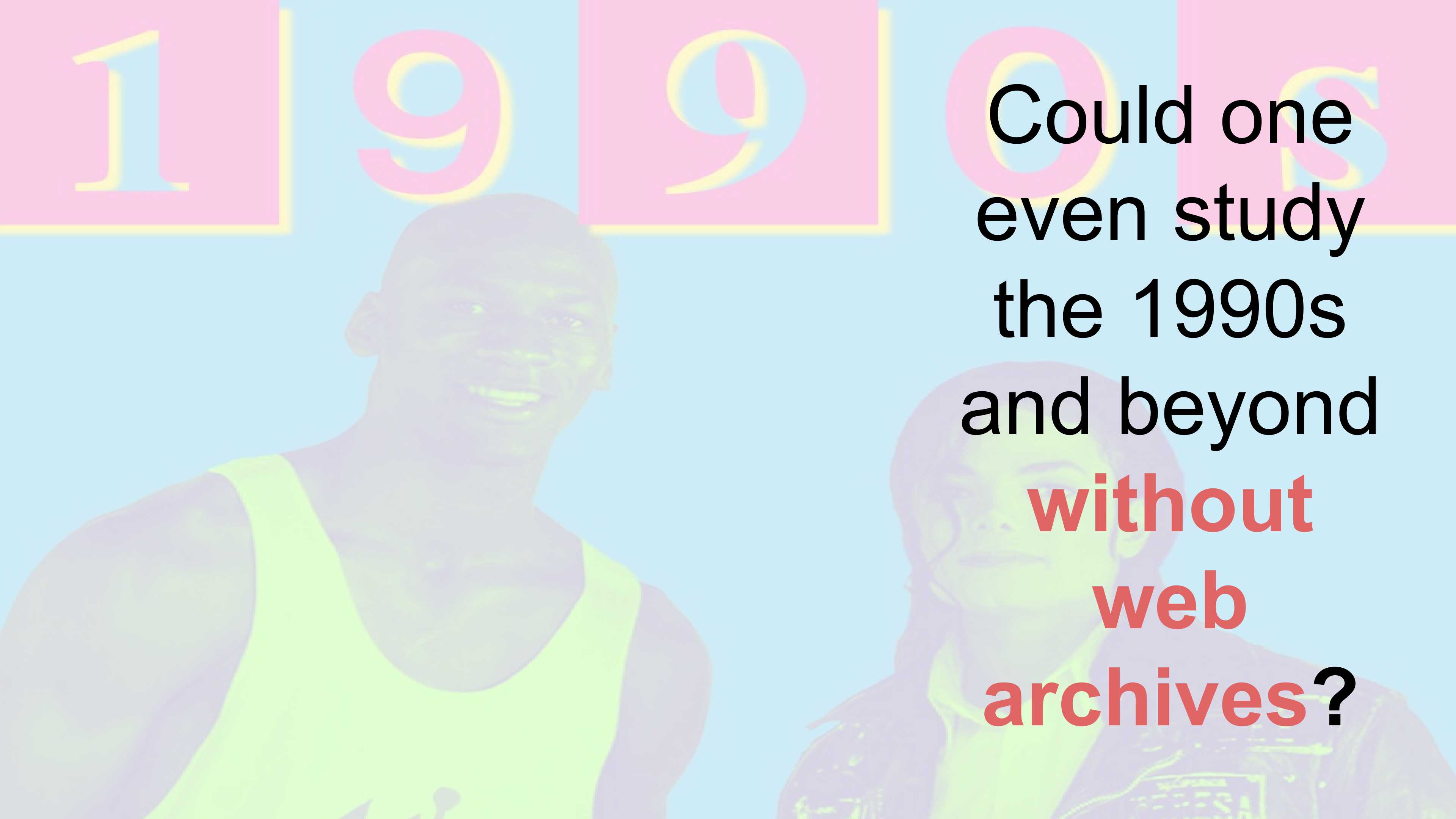
This is a scale that **boggles the mind** – compare it to the Old Bailey 197,745 trials between 1674 and 1913)

Scarcity

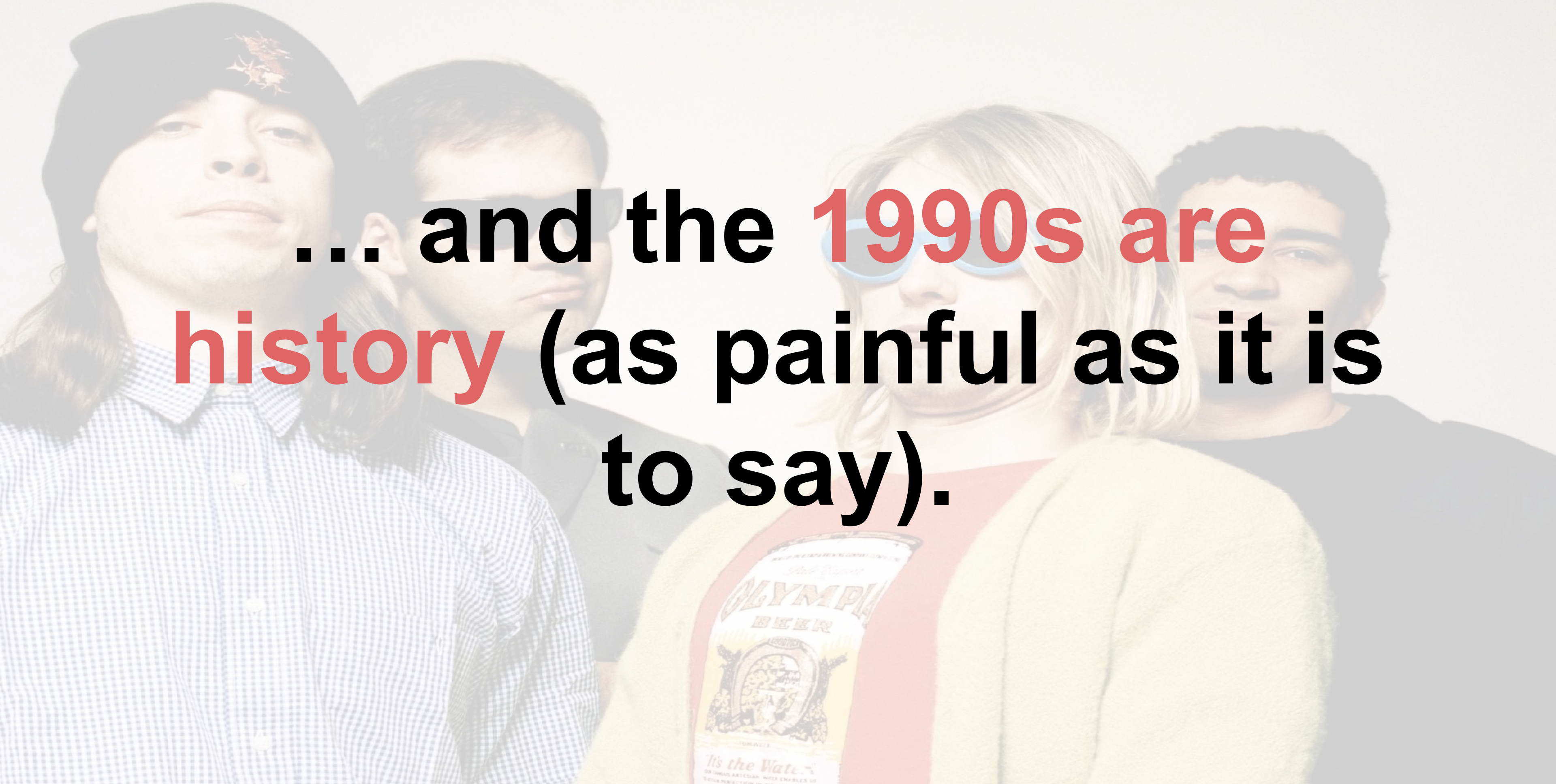
370
371
372
CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series V
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Subseries F
Democratic Party
BOX 28
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
373
CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series V
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Subseries F
Democratic Party
BOX 29
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
374
CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series V
PARTY LEADERSHIP / ADMINISTRATIVE FILE
Subseries F
Democratic Party
BOX 30
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
375

~~Scarcity~~

Abundance



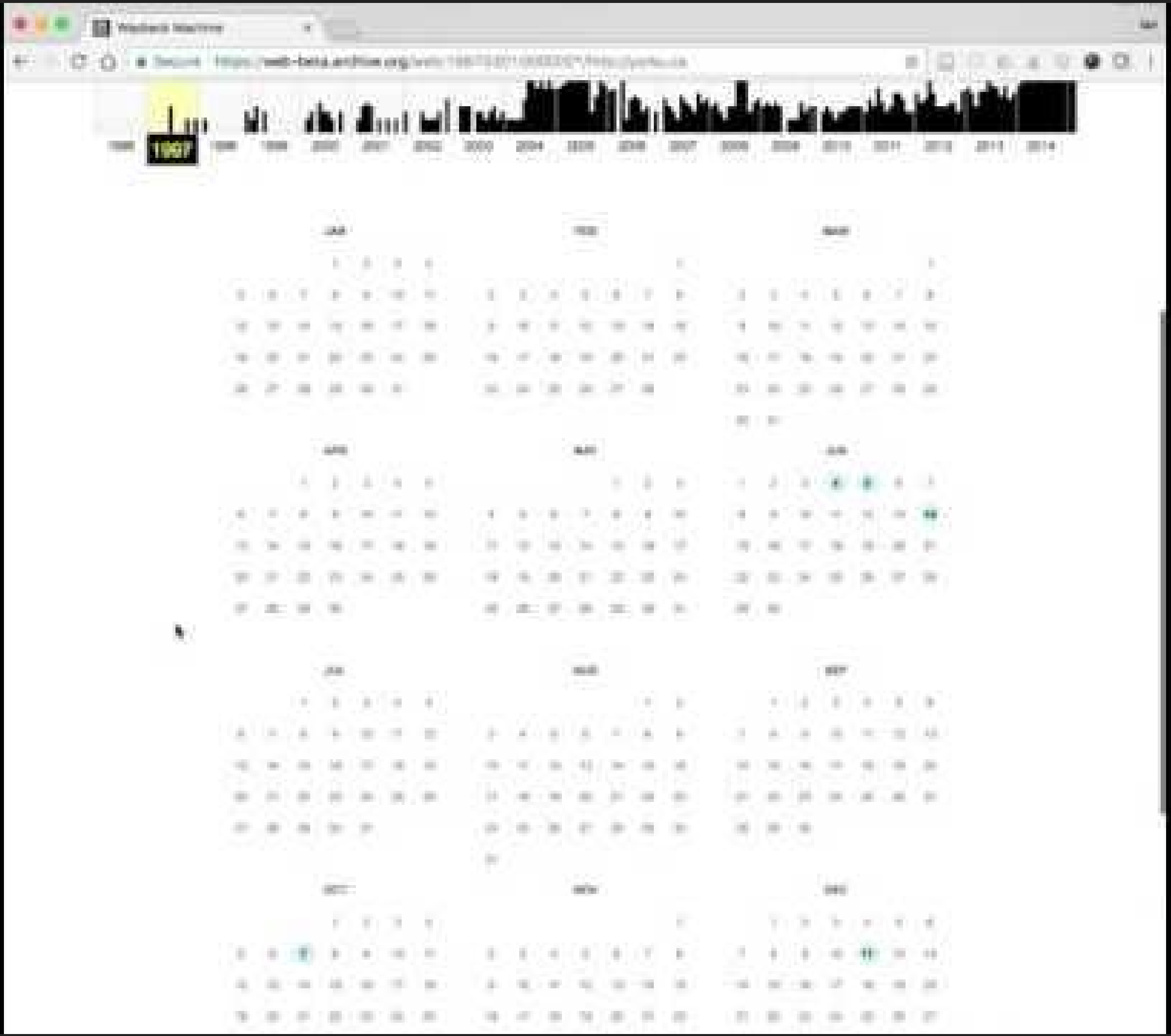
Could one even study the 1990s and beyond **without web archives**?



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

And we have fears...

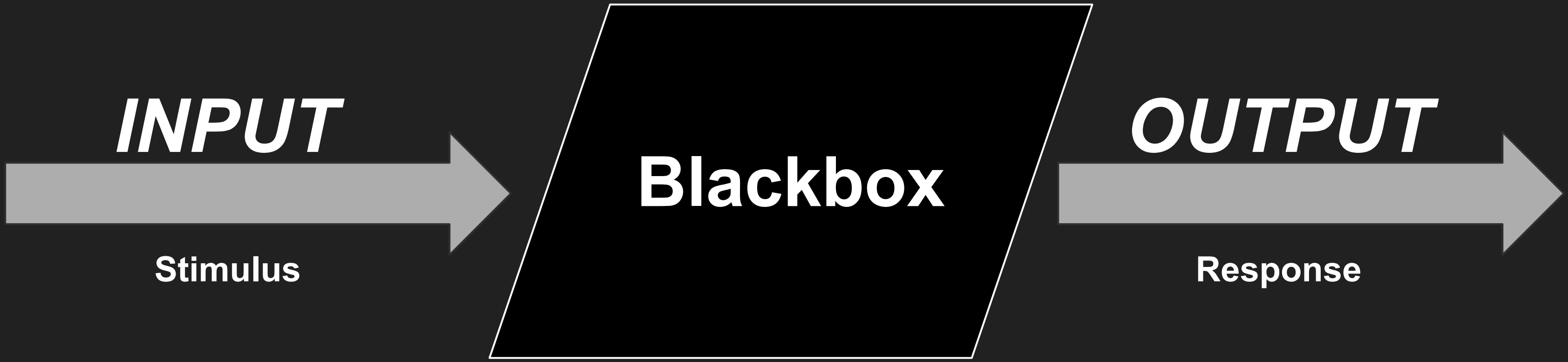
The decisions we make today will lay the foundations for how we work with born-digital cultural heritage.

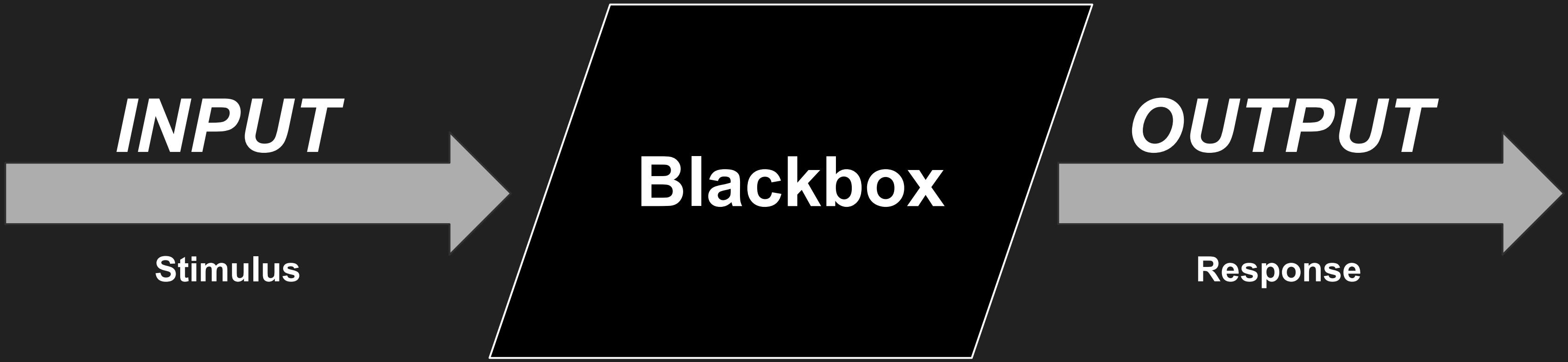


Wayback Machine
Secure
https://web-beta.archive.org/web/1997*/http://yorku.ca
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC

Won’t be enough - we’ll need search engines and discovery tools.

But what will our search engines look like?



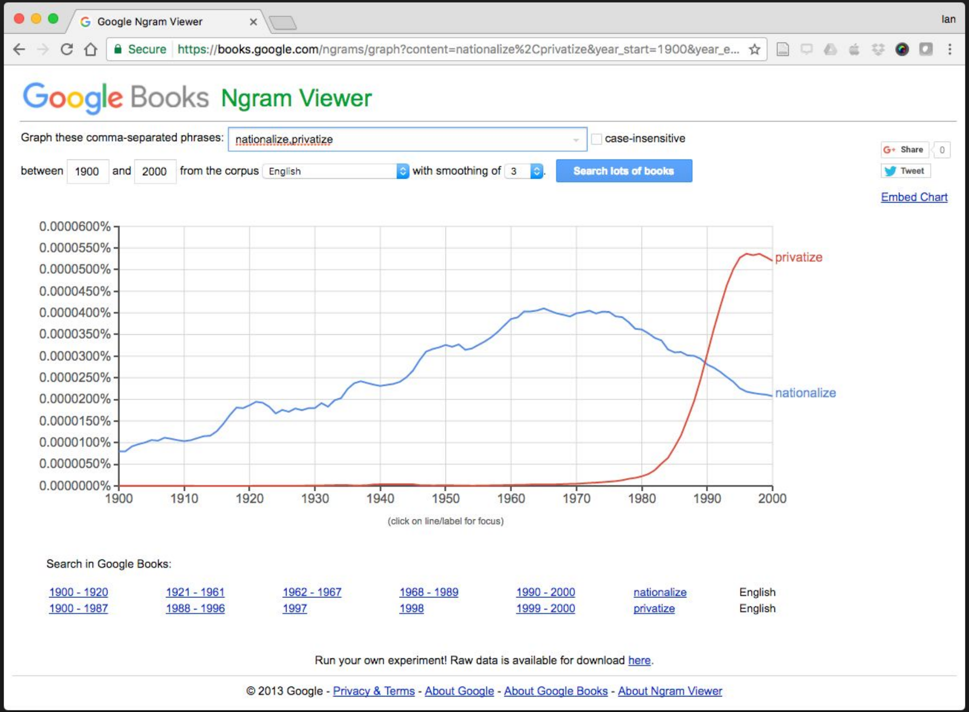


We can’t let the Blackbox write our histories.

Our nightmare:

Historians rely uncritically on date-ordered or algorithmically-ranked keyword search results, putting them at mercy of search algorithms they do not understand.

Some disturbing trends in this area...



Google Ngram Viewer

Secure https://books.google.com/ngrams/graph?content=nationalize%2Cprivatize&year_start=1900&year_e...

Google Books Ngram Viewer

Graph these comma-separated phrases: nationalize,privatize □ case-insensitive

between 1900 and 2000 from the corpus English with smoothing of 3.

Search in Google Books:

1900 - 1920	1921 - 1961	1962 - 1967	1968 - 1989	1990 - 2000	nationalize	English
1900 - 1987	1988 - 1996	1997	1998	1999 - 2000	privatize	English

Run your own experiment! Raw data is available for download here.

© 2013 Google - Privacy & Terms - About Google - About Google Books - About Ngram Viewer

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

We need *interdisciplinary collaboration* to tackle this problem!

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canadä

Ontario

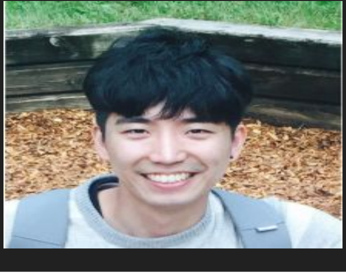
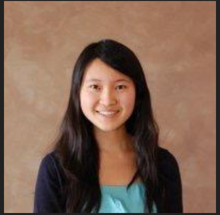
compute canada | calcul canada

Team(s)

Web Archives for Historical Research

Historians

Computer Scientists

Librarians
Archivists

Governance

Projects & Platforms

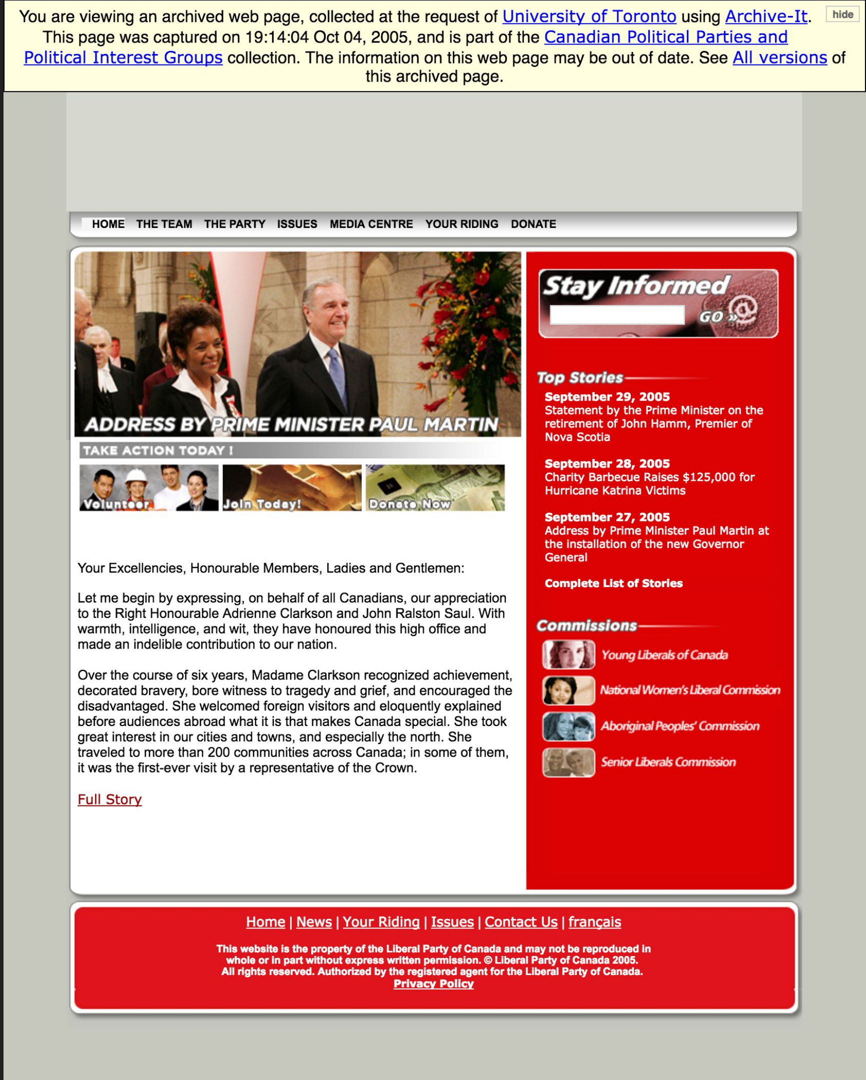
webarchives.ca

Shine

https://github.com/ukwa/shine/

Canadian Political Parties & Political Interest Group Collection (ARCHIVE-IT/Toronto)

- 50 Websites
  - All major political parties
  - Many minor political parties
  - Political interest groups
- Collected quarterly between 2005 and present



hide

You are viewing an archived web page, collected at the request of University of Toronto using Archive-It. This page was captured on 19:14:04 Oct 04, 2005, and is part of the Canadian Political Parties and Political Interest Groups collection. The information on this web page may be out of date. See All versions of this archived page.

HOME THE TEAM THE PARTY ISSUES MEDIA CENTRE YOUR RIDING DONATE
ADDRESS BY PRIME MINISTER PAUL MARTIN
TAKE ACTION TODAY !
Volunteer
Join Today!
Donate Now
Stay Informed
GO »

Your Excellencies, Honourable Members, Ladies and Gentlemen:

Let me begin by expressing, on behalf of all Canadians, our appreciation to the Right Honourable Adrienne Clarkson and John Ralston Saul. With warmth, intelligence, and wit, they have honoured this high office and made an indelible contribution to our nation.

Over the course of six years, Madame Clarkson recognized achievement, decorated bravery, bore witness to tragedy and grief, and encouraged the disadvantaged. She welcomed foreign visitors and eloquently explained before audiences abroad what it is that makes Canada special. She took great interest in our cities and towns, and especially the north. She traveled to more than 200 communities across Canada; in some of them, it was the first-ever visit by a representative of the Crown.

Full Story

Top Stories

September 29, 2005
Statement by the Prime Minister on the retirement of John Hamm, Premier of Nova Scotia

September 28, 2005
Charity Barbecue Raises $125,000 for Hurricane Katrina Victims

September 27, 2005
Address by Prime Minister Paul Martin at the installation of the new Governor General

Complete List of Stories

Commissions

Young Liberals of Canada

National Women's Liberal Commission

Aboriginal Peoples' Commission

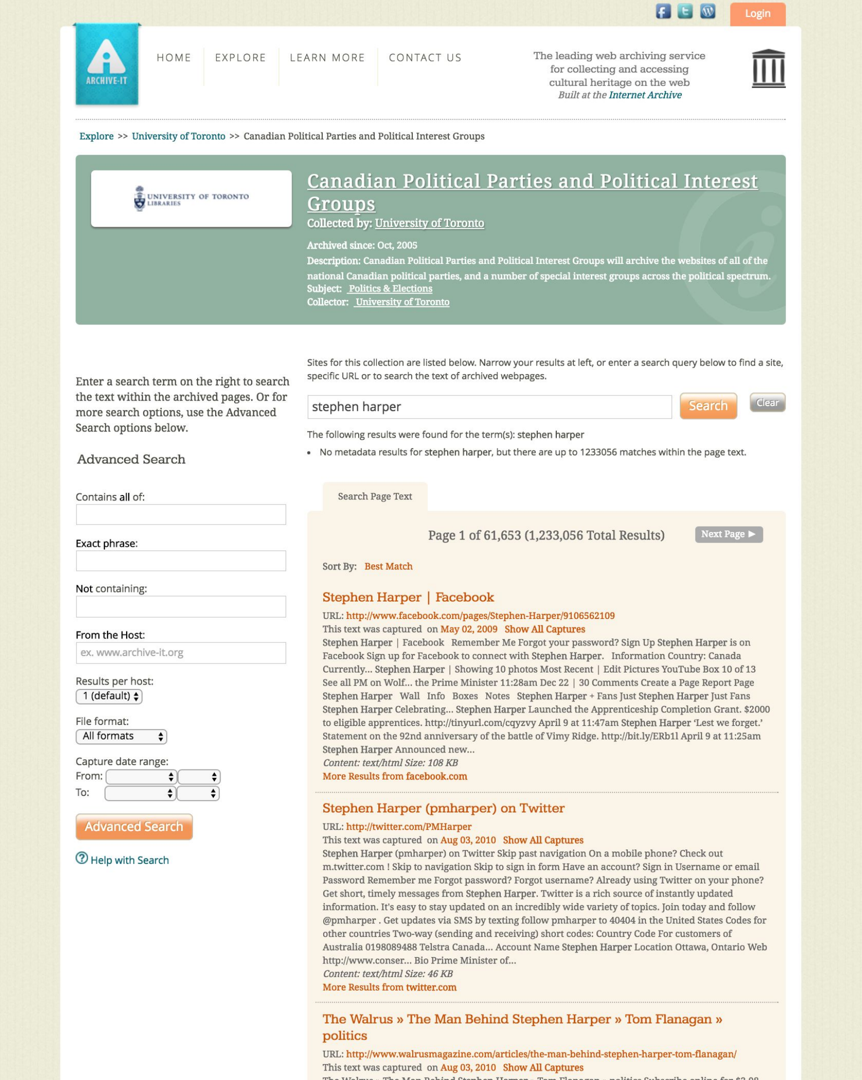
Senior Liberals Commission

Home | News | Your Riding | Issues | 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 2005. All rights reserved. Authorized by the registered agent for the Liberal Party of Canada.
Privacy Policy

The Current Interface..

- **Very limited** - simple search engine, some advanced options; no facets
- **Great collection**.. But nobody uses them.



The leading web archiving service for collecting and accessing cultural heritage on the web
Built at the Internet Archive

Explore >> University of Toronto >> Canadian Political Parties and Political Interest Groups

Canadian Political Parties and Political Interest Groups

Collected by: University of Toronto

Archived since: Oct, 2005

Description: Canadian Political Parties and Political Interest Groups will archive the websites of all of the national Canadian political parties, and a number of special interest groups across the political spectrum.

Subject: Politics & Elections

Collector: University of Toronto

Enter a search term on the right to search the text within the archived pages. Or for more search options, use the Advanced Search options below.

Advanced Search

Contains **all** of:

Exact phrase:

Not containing:

From the Host: ex. www.archive-it.org

Results per host: 1 (default)

File format: All formats

Capture date range:
From:
To:

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

stephen harper

The following results were found for the term(s): stephen harper

- No metadata results for stephen harper, but there are up to 1233056 matches within the page text.

Page 1 of 61,653 (1,233,056 Total Results)

Sort By: Best Match

Stephen Harper | Facebook

URL: http://www.facebook.com/pages/Stephen-Harper/9106562109
This text was captured on May 02, 2009 Show All Captures
Stephen Harper | Facebook Remember Me Forgot your password? Sign Up Stephen Harper is on Facebook Sign up for Facebook to connect with Stephen Harper. Information Country: Canada Currently... Stephen Harper | Showing 10 photos Most Recent | Edit Pictures YouTube Box 10 of 13 See all PM on Wolf... the Prime Minister 11:28am Dec 22 | 30 Comments Create a Page Report Page Stephen Harper Wall Info Boxes Notes Stephen Harper + Fans Just Stephen Harper Just Fans Stephen Harper Celebrating... Stephen Harper Launched the Apprenticeship Completion Grant. $2000 to eligible apprentices. http://tinyurl.com/cqyzvy April 9 at 11:47am Stephen Harper 'Lest we forget.' Statement on the 92nd anniversary of the battle of Vimy Ridge. http://bit.ly/ERb1l April 9 at 11:25am Stephen Harper Announced new...
Content: text/html Size: 108 KB
More Results from facebook.com

Stephen Harper (pmharper) on Twitter

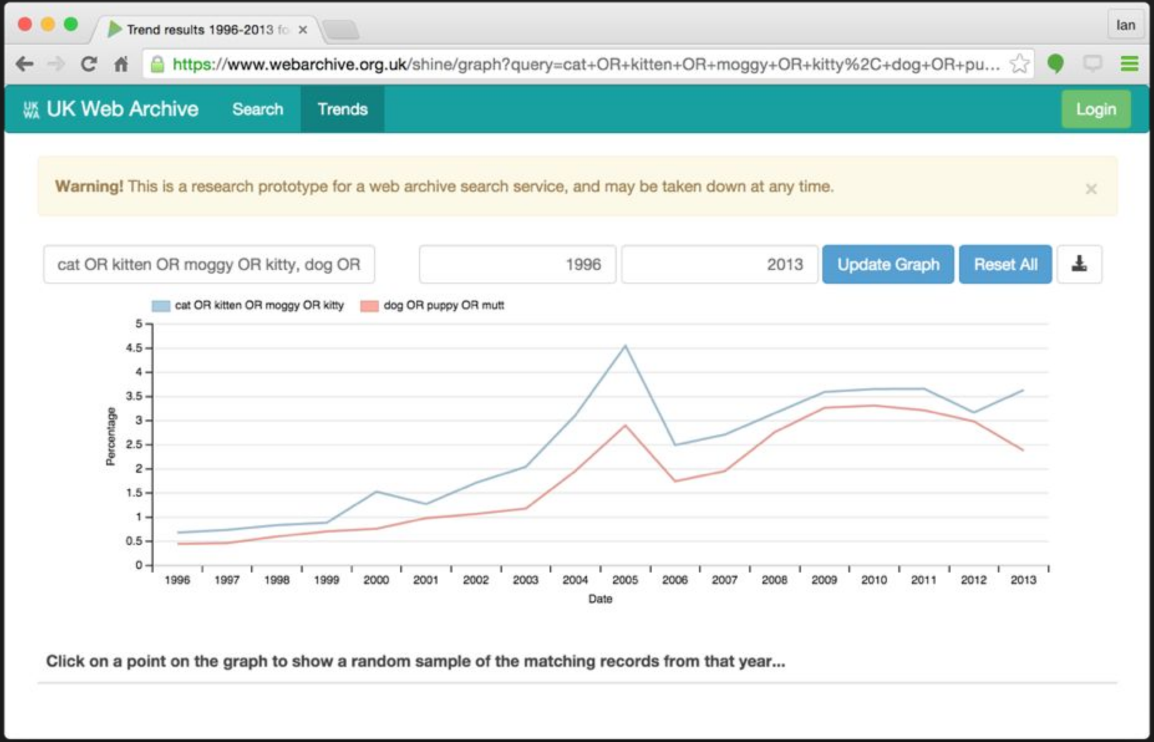
URL: http://twitter.com/PMHarper
This text was captured on Aug 03, 2010 Show All Captures
Stephen Harper (pmharper) on Twitter Skip past navigation On a mobile phone? Check out m.twitter.com ! Skip to navigation Skip to sign in form Have an account? Sign in Username or email Password Remember me Forgot password? Forgot username? Already using Twitter on your phone? Get short, timely messages from Stephen Harper. Twitter is a rich source of instantly updated information. It's easy to stay updated on an incredibly wide variety of topics. Join today and follow @pmharper . Get updates via SMS by texting follow pmharper to 40404 in the United States Codes for other countries Two-way (sending and receiving) short codes: Country Code For customers of Australia 0198089488 Telstra Canada... Account Name Stephen Harper Location Ottawa, Ontario Web http://www.conser... Bio Prime Minister of...
Content: text/html Size: 46 KB
More Results from twitter.com

The Walrus » The Man Behind Stephen Harper » Tom Flanagan » politics

URL: http://www.walrusmagazine.com/articles/the-man-behind-stephen-harper-tom-flanagan/
This text was captured on Aug 03, 2010 Show All Captures

14 Million Solr docs!

Shine



Warning! This is a research prototype for a web archive search service, and may be taken down at any time.

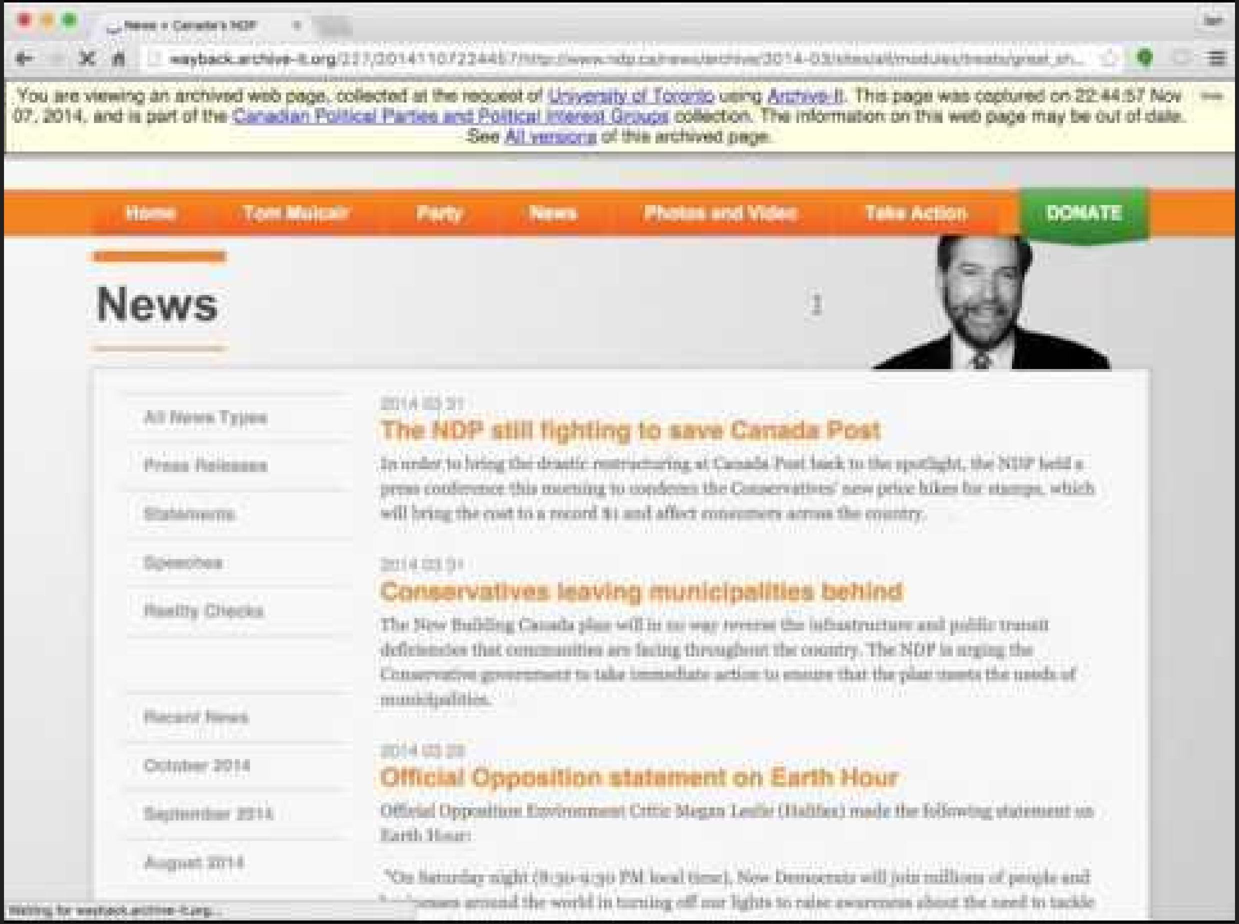
cat OR kitten OR moggy OR kitty, dog OR

1996

2013

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

webarchives.ca



News « Canada's NDP
wayback.archive-it.org/227/20141107224457/http://www.ndp.ca/news/archive/2014-03/sites/all/modules/treats/great_sh...

You are viewing an archived web page, collected at the request of University of Toronto using Archive-It. This page was captured on 22:44:57 Nov 07, 2014, and is part of the Canadian Political Parties and Political Interest Groups collection. The information on this web page may be out of date. See All versions of this archived page.

News

2014 03 31

The NDP still fighting to save Canada Post

In order to bring the drastic restructuring at Canada Post back to the spotlight, the NDP held a press conference this morning to condemn the Conservatives' new price hikes for stamps, which will bring the cost to a record $1 and affect consumers across the country.

2014 03 31

Conservatives leaving municipalities behind

The New Building Canada plan will in no way reverse the infrastructure and public transit deficiencies that communities are facing throughout the country. The NDP is urging the Conservative government to take immediate action to ensure that the plan meets the needs of municipalities.

2014 03 28

Official Opposition statement on Earth Hour

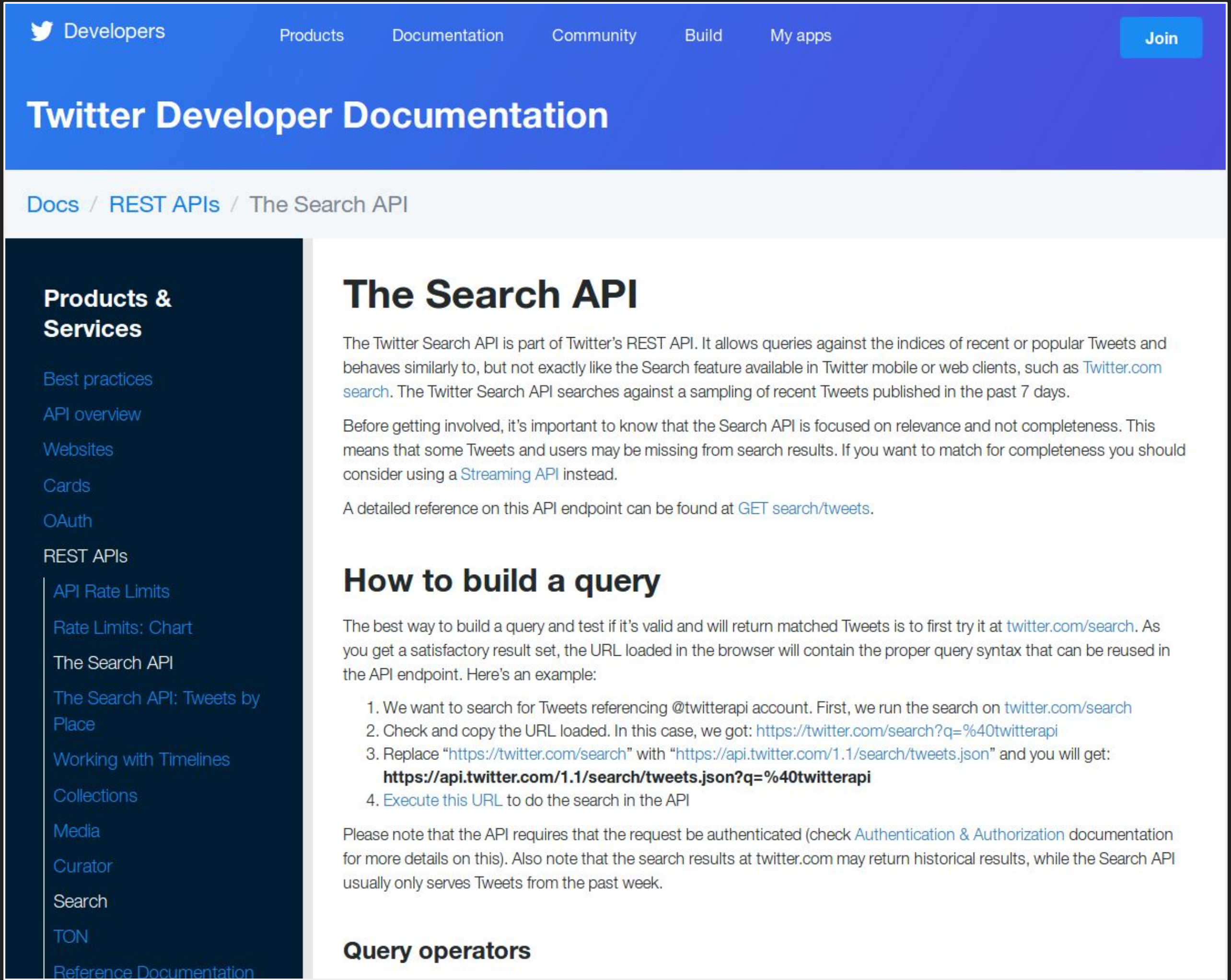
Official Opposition Environment Critic Megan Leslie (Halifax) made the following statement on Earth Hour:

"On Saturday night (8:30-9:30 PM local time), New Democrats will join millions of people and businesses around the world in turning off our lights to raise awareness about the need to tackle

Twitter

twarc

docnow.io



The Search API

The Twitter Search API is part of Twitter's REST API. It allows queries against the indices of recent or popular Tweets and behaves similarly to, but not exactly like the Search feature available in Twitter mobile or web clients, such as Twitter.com search. The Twitter Search API searches against a sampling of recent Tweets published in the past 7 days.

Before getting involved, it's important to know that the Search API is focused on relevance and not completeness. This means that some Tweets and users may be missing from search results. If you want to match for completeness you should consider using a Streaming API instead.

A detailed reference on this API endpoint can be found at GET search/tweets.

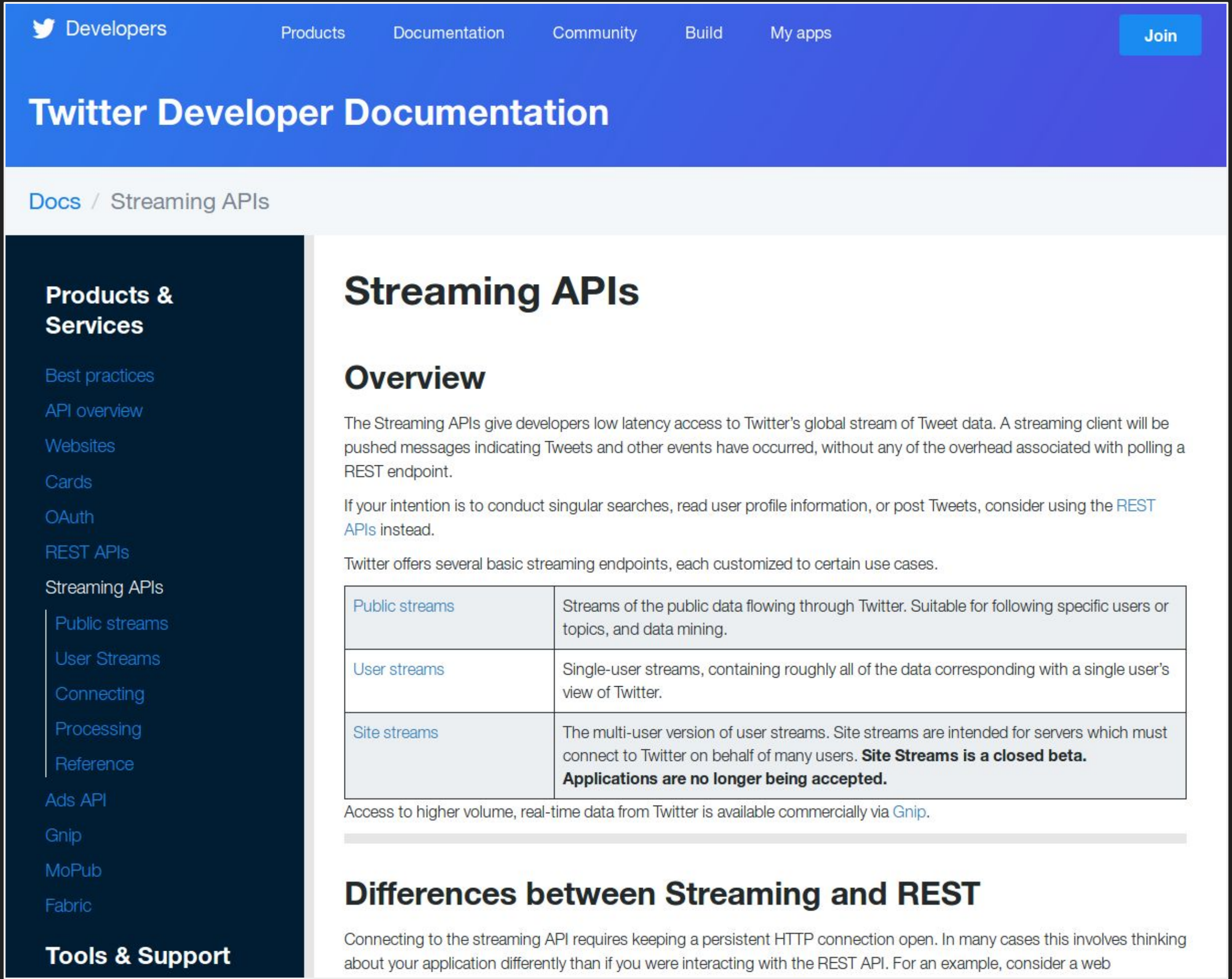
How to build a query

The best way to build a query and test if it's valid and will return matched Tweets is to first try it at twitter.com/search. As you get a satisfactory result set, the URL loaded in the browser will contain the proper query syntax that can be reused in the API endpoint. Here's an example:

1. We want to search for Tweets referencing @twitterapi account. First, we run the search on twitter.com/search
2. Check and copy the URL loaded. In this case, we got: https://twitter.com/search?q=%40twitterapi
3. Replace "https://twitter.com/search" with "https://api.twitter.com/1.1/search/tweets.json" and you will get: **https://api.twitter.com/1.1/search/tweets.json?q=%40twitterapi**
4. Execute this URL to do the search in the API

Please note that the API requires that the request be authenticated (check Authentication & Authorization documentation for more details on this). Also note that the search results at twitter.com may return historical results, while the Search API usually only serves Tweets from the past week.

Query operators



Twitter Developer Documentation

Streaming APIs

Overview

The Streaming APIs give developers low latency access to Twitter's global stream of Tweet data. A streaming client will be pushed messages indicating Tweets and other events have occurred, without any of the overhead associated with polling a REST endpoint.

If your intention is to conduct singular searches, read user profile information, or post Tweets, consider using the REST APIs instead.

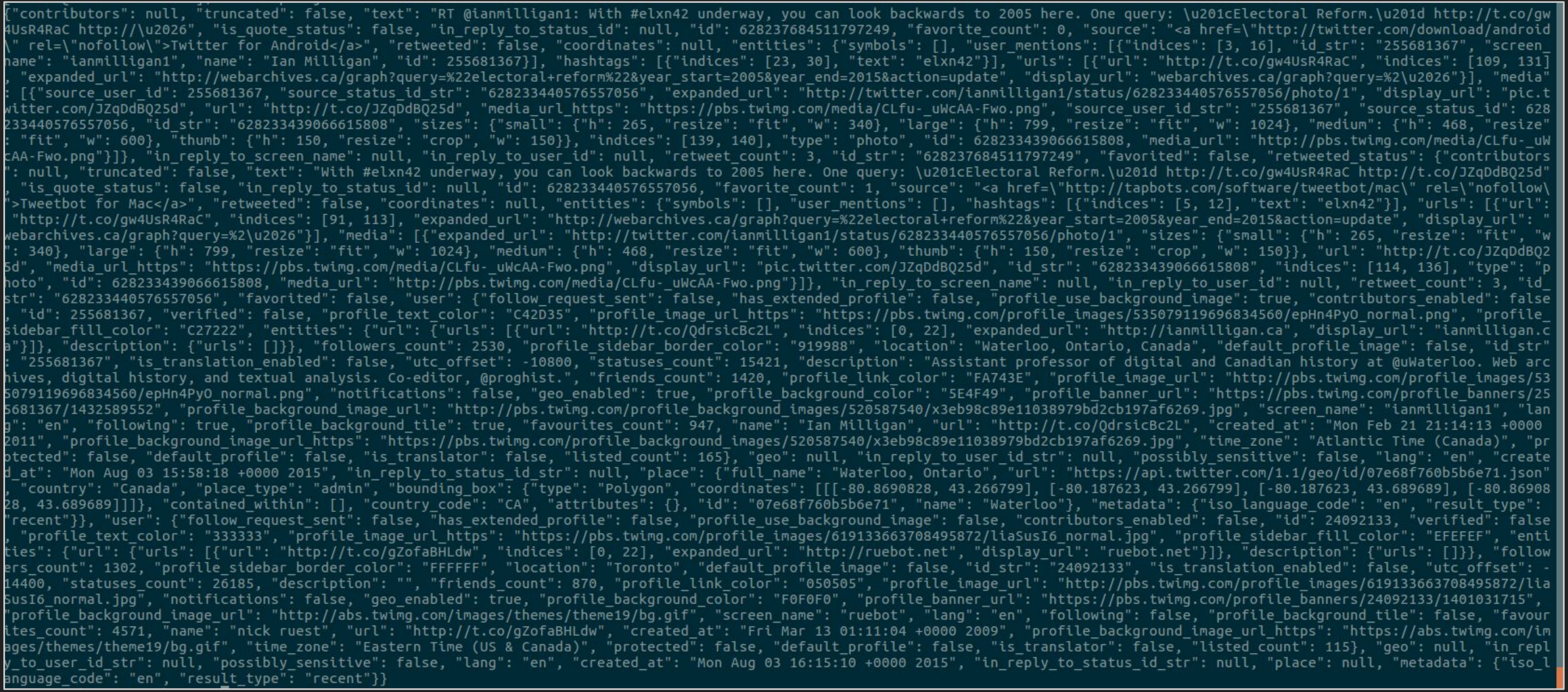
Twitter offers several basic streaming endpoints, each customized to certain use cases.

Public streams	Streams of the public data flowing through Twitter. Suitable for following specific users or topics, and data mining.
User streams	Single-user streams, containing roughly all of the data corresponding with a single user's view of Twitter.
Site streams	The multi-user version of user streams. Site streams are intended for servers which must connect to Twitter on behalf of many users. **Site Streams is a closed beta. Applications are no longer being accepted.**

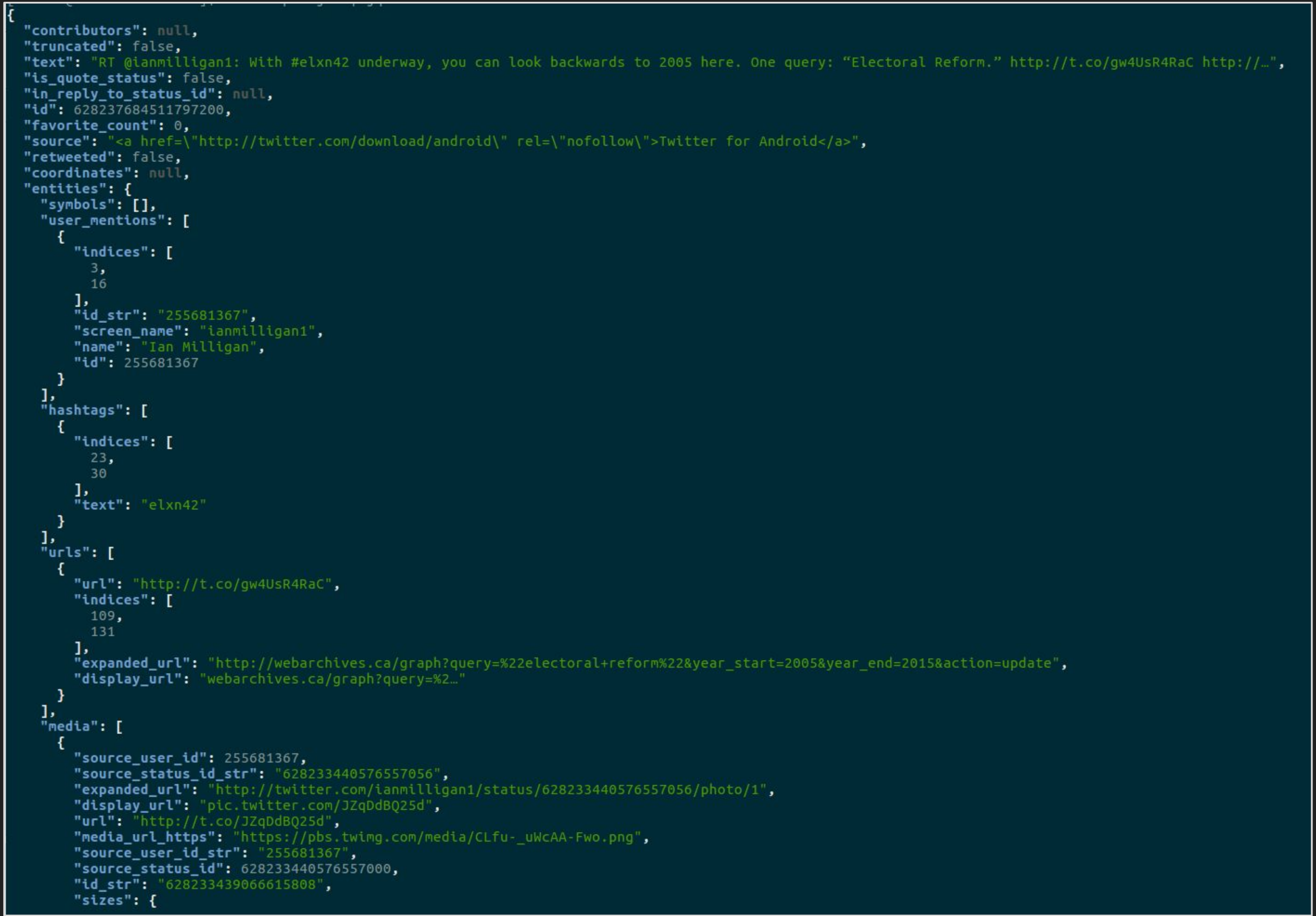
Access to higher volume, real-time data from Twitter is available commercially via Gnip.

Differences between Streaming and REST

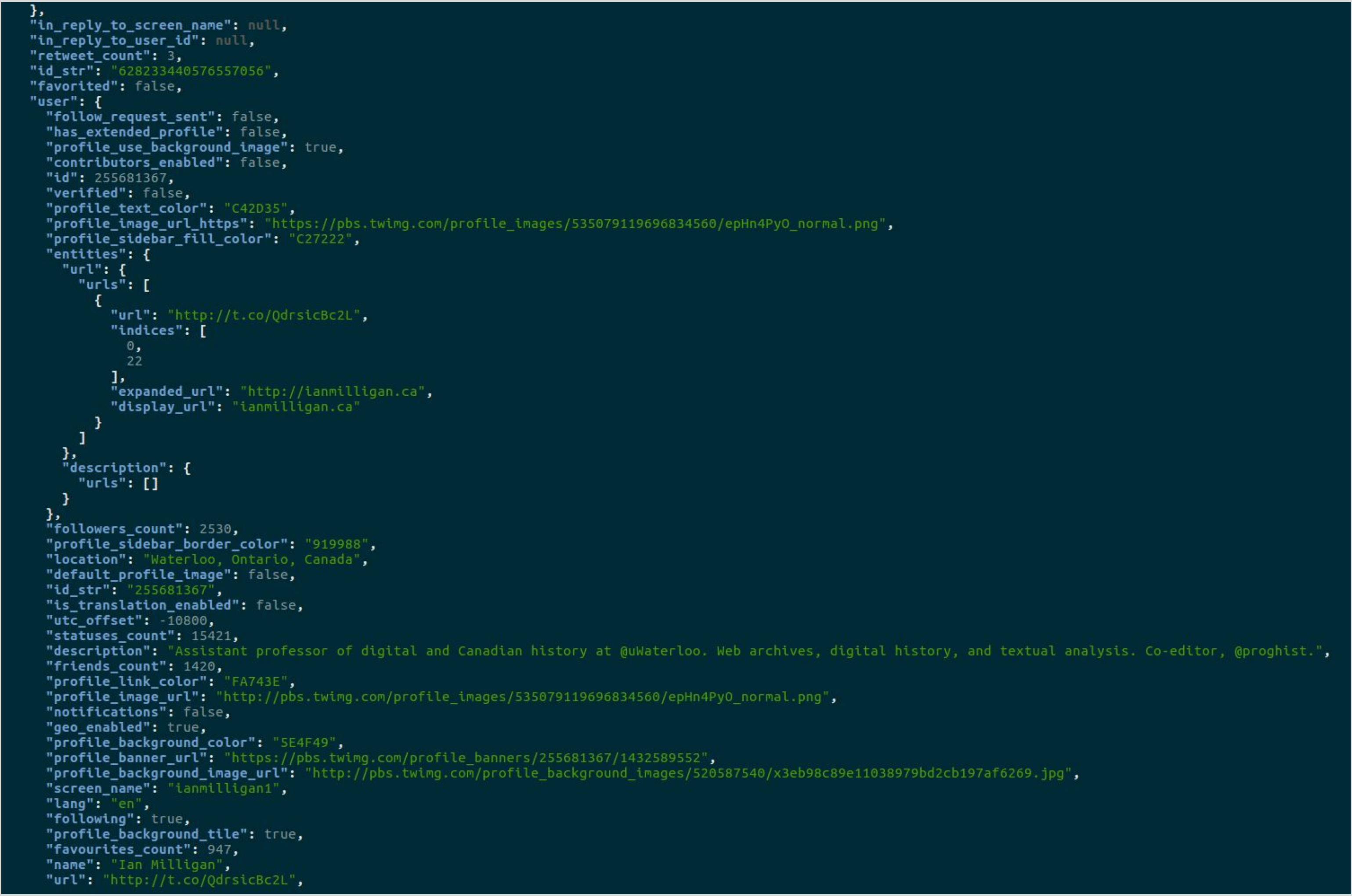
Connecting to the streaming API requires keeping a persistent HTTP connection open. In many cases this involves thinking about your application differently than if you were interacting with the REST API. For an example, consider a web



```
{"contributors": null, "truncated": false, "text": "RT @ianmilligan1: With #elxn42 underway, you can look backwards to 2005 here. One query: \u201cElectoral Reform.\u201d http://t.co/gw
4UsR4RaC http://\u2026", "is_quote_status": false, "in_reply_to_status_id": null, "id": 628237684511797249, "favorite_count": 0, "source": "<a href=\"http://twitter.com/download/android
\" rel=\"nofollow\">Twitter for Android</a>", "retweeted": false, "coordinates": null, "entities": {"symbols": [], "user_mentions": [{"indices": [3, 16], "id_str": "255681367", "screen_
name": "ianmilligan1", "name": "Ian Milligan", "id": 255681367}], "hashtags": [{"indices": [23, 30], "text": "elxn42"}], "urls": [{"url": "http://t.co/gw4UsR4RaC", "indices": [109, 131]
, "expanded_url": "http://webarchives.ca/graph?query=%22electoral+reform%22&year_start=2005&year_end=2015&action=update", "display_url": "webarchives.ca/graph?query=%2\u2026"}], "media"
: [{"source_user_id": 255681367, "source_status_id_str": "628233440576557056", "expanded_url": "http://twitter.com/ianmilligan1/status/628233440576557056/photo/1", "display_url": "pic.t
witter.com/JZqDdBQ25d", "url": "http://t.co/JZqDdBQ25d", "media_url_https": "https://pbs.twimg.com/media/CLfu-_uWcAA-Fwo.png", "source_user_id_str": "255681367", "source_status_id": 628
233440576557056, "id_str": "628233439066615808", "sizes": {"small": {"h": 265, "resize": "fit", "w": 340}, "large": {"h": 799, "resize": "fit", "w": 1024}, "medium": {"h": 468, "resize"
: "fit", "w": 600}, "thumb": {"h": 150, "resize": "crop", "w": 150}}, "indices": [139, 140], "type": "photo", "id": 628233439066615808, "media_url": "http://pbs.twimg.com/media/CLfu-_uW
cAA-Fwo.png"}]}, "in_reply_to_screen_name": null, "in_reply_to_user_id": null, "retweet_count": 3, "id_str": "628237684511797249", "favorited": false, "retweeted_status": {"contributors
": null, "truncated": false, "text": "With #elxn42 underway, you can look backwards to 2005 here. One query: \u201cElectoral Reform.\u201d http://t.co/gw4UsR4RaC http://t.co/JZqDdBQ25d"
, "is_quote_status": false, "in_reply_to_status_id": null, "id": 628233440576557056, "favorite_count": 1, "source": "<a href=\"http://tapbots.com/software/tweetbot/mac\" rel=\"nofollow\
">Tweetbot for Mac</a>", "retweeted": false, "coordinates": null, "entities": {"symbols": [], "user_mentions": [], "hashtags": [{"indices": [5, 12], "text": "elxn42"}], "urls": [{"url":
 "http://t.co/gw4UsR4RaC", "indices": [91, 113], "expanded_url": "http://webarchives.ca/graph?query=%22electoral+reform%22&year_start=2005&year_end=2015&action=update", "display_url": "
webarchives.ca/graph?query=%2\u2026"}], "media": [{"expanded_url": "http://twitter.com/ianmilligan1/status/628233440576557056/photo/1", "sizes": {"small": {"h": 265, "resize": "fit", "w
": 340}, "large": {"h": 799, "resize": "fit", "w": 1024}, "medium": {"h": 468, "resize": "fit", "w": 600}, "thumb": {"h": 150, "resize": "crop", "w": 150}}, "url": "http://t.co/JZqDdBQ2
5d", "media_url_https": "https://pbs.twimg.com/media/CLfu-_uWcAA-Fwo.png", "display_url": "pic.twitter.com/JZqDdBQ25d", "id_str": "628233439066615808", "indices": [114, 136], "type": "p
hoto", "id": 628233439066615808, "media_url": "http://pbs.twimg.com/media/CLfu-_uWcAA-Fwo.png"}]}, "in_reply_to_screen_name": null, "in_reply_to_user_id": null, "retweet_count": 3, "id_
str": "628233440576557056", "favorited": false, "user": {"follow_request_sent": false, "has_extended_profile": false, "profile_use_background_image": true, "contributors_enabled": false
, "id": 255681367, "verified": false, "profile_text_color": "C42D35", "profile_image_url_https": "https://pbs.twimg.com/profile_images/535079119696834560/epHn4PyO_normal.png", "profile_
sidebar_fill_color": "C27222", "entities": {"url": {"urls": [{"url": "http://t.co/QdrsicBc2L", "indices": [0, 22], "expanded_url": "http://ianmilligan.ca", "display_url": "ianmilligan.c
a"}]}, "description": {"urls": []}}, "followers_count": 2530, "profile_sidebar_border_color": "919988", "location": "Waterloo, Ontario, Canada", "default_profile_image": false, "id_str"
: "255681367", "is_translation_enabled": false, "utc_offset": -10800, "statuses_count": 15421, "description": "Assistant professor of digital and Canadian history at @uWaterloo. Web arc
hives, digital history, and textual analysis. Co-editor, @proghist.", "friends_count": 1420, "profile_link_color": "FA743E", "profile_image_url": "http://pbs.twimg.com/profile_images/53
5079119696834560/epHn4PyO_normal.png", "notifications": false, "geo_enabled": true, "profile_background_color": "5E4F49", "profile_banner_url": "https://pbs.twimg.com/profile_banners/25
5681367/1432589552", "profile_background_image_url": "http://pbs.twimg.com/profile_background_images/520587540/x3eb98c89e11038979bd2cb197af6269.jpg", "screen_name": "ianmilligan1", "lan
g": "en", "following": true, "profile_background_tile": true, "favourites_count": 947, "name": "Ian Milligan", "url": "http://t.co/QdrsicBc2L", "created_at": "Mon Feb 21 21:14:13 +0000 
2011", "profile_background_image_url_https": "https://pbs.twimg.com/profile_background_images/520587540/x3eb98c89e11038979bd2cb197af6269.jpg", "time_zone": "Atlantic Time (Canada)", "pr
otected": false, "default_profile": false, "is_translator": false, "listed_count": 165}, "geo": null, "in_reply_to_user_id_str": null, "possibly_sensitive": false, "lang": "en", "create
d_at": "Mon Aug 03 15:58:18 +0000 2015", "in_reply_to_status_id_str": null, "place": {"full_name": "Waterloo, Ontario", "url": "https://api.twitter.com/1.1/geo/id/07e68f760b5b6e71.json"
, "country": "Canada", "place_type": "admin", "bounding_box": {"type": "Polygon", "coordinates": [[[-80.8690828, 43.266799], [-80.187623, 43.266799], [-80.187623, 43.689689], [-80.86908
28, 43.689689]]]}, "contained_within": [], "country_code": "CA", "attributes": {}, "id": "07e68f760b5b6e71", "name": "Waterloo"}, "metadata": {"iso_language_code": "en", "result_type": 
"recent"}}, "user": {"follow_request_sent": false, "has_extended_profile": false, "profile_use_background_image": false, "contributors_enabled": false, "id": 24092133, "verified": false
, "profile_text_color": "333333", "profile_image_url_https": "https://pbs.twimg.com/profile_images/619133663708495872/liaSusI6_normal.jpg", "profile_sidebar_fill_color": "EFEFEF", "enti
ties": {"url": {"urls": [{"url": "http://t.co/gZofaBHLdw", "indices": [0, 22], "expanded_url": "http://ruebot.net", "display_url": "ruebot.net"}]}, "description": {"urls": []}}, "follow
ers_count": 1302, "profile_sidebar_border_color": "FFFFFF", "location": "Toronto", "default_profile_image": false, "id_str": "24092133", "is_translation_enabled": false, "utc_offset": -
14400, "statuses_count": 26185, "description": "", "friends_count": 870, "profile_link_color": "050505", "profile_image_url": "http://pbs.twimg.com/profile_images/619133663708495872/lia
SusI6_normal.jpg", "notifications": false, "geo_enabled": true, "profile_background_color": "F0F0F0", "profile_banner_url": "https://pbs.twimg.com/profile_banners/24092133/1401031715", 
"profile_background_image_url": "http://abs.twimg.com/images/themes/theme19/bg.gif", "screen_name": "ruebot", "lang": "en", "following": false, "profile_background_tile": false, "favour
ites_count": 4571, "name": "nick ruest", "url": "http://t.co/gZofaBHLdw", "created_at": "Fri Mar 13 01:11:04 +0000 2009", "profile_background_image_url_https": "https://abs.twimg.com/im
ages/themes/theme19/bg.gif", "time_zone": "Eastern Time (US & Canada)", "protected": false, "default_profile": false, "is_translator": false, "listed_count": 115}, "geo": null, "in_repl
y_to_user_id_str": null, "possibly_sensitive": false, "lang": "en", "created_at": "Mon Aug 03 16:15:10 +0000 2015", "in_reply_to_status_id_str": null, "place": null, "metadata": {"iso_l
anguage_code": "en", "result_type": "recent"}}
```

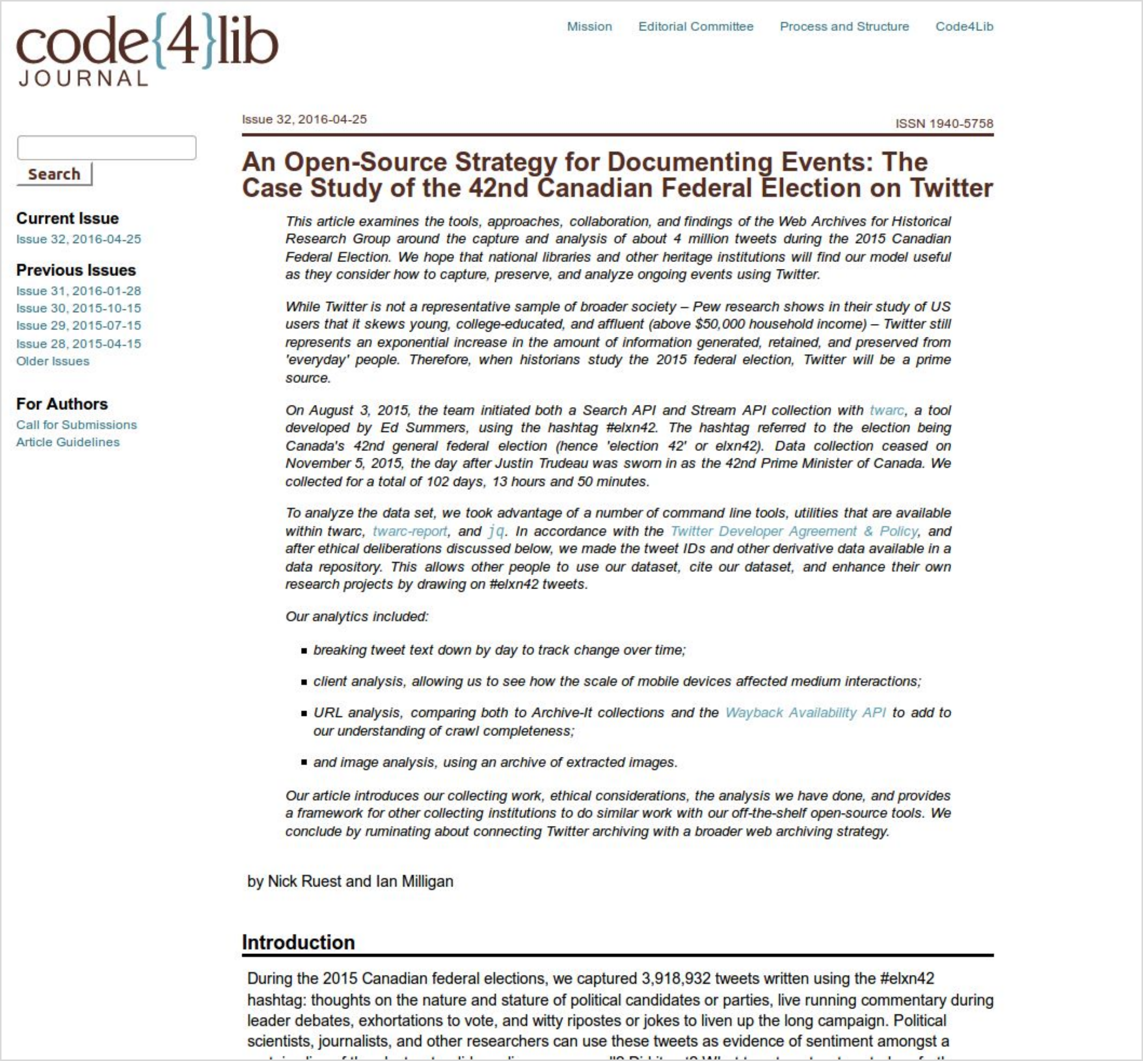


```
{
  "contributors": null,
  "truncated": false,
  "text": "RT @ianmilligan1: With #elxn42 underway, you can look backwards to 2005 here. One query: “Electoral Reform.” http://t.co/gw4UsR4RaC http://…",
  "is_quote_status": false,
  "in_reply_to_status_id": null,
  "id": 628237684511797200,
  "favorite_count": 0,
  "source": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
  "retweeted": false,
  "coordinates": null,
  "entities": {
    "symbols": [],
    "user_mentions": [
      {
        "indices": [
          3,
          16
        ],
        "id_str": "255681367",
        "screen_name": "ianmilligan1",
        "name": "Ian Milligan",
        "id": 255681367
      }
    ],
    "hashtags": [
      {
        "indices": [
          23,
          30
        ],
        "text": "elxn42"
      }
    ],
    "urls": [
      {
        "url": "http://t.co/gw4UsR4RaC",
        "indices": [
          109,
          131
        ],
        "expanded_url": "http://webarchives.ca/graph?query=%22electoral+reform%22&year_start=2005&year_end=2015&action=update",
        "display_url": "webarchives.ca/graph?query=%2…"
      }
    ],
    "media": [
      {
        "source_user_id": 255681367,
        "source_status_id_str": "628233440576557056",
        "expanded_url": "http://twitter.com/ianmilligan1/status/628233440576557056/photo/1",
        "display_url": "pic.twitter.com/JZqDdBQ25d",
        "url": "http://t.co/JZqDdBQ25d",
        "media_url_https": "https://pbs.twimg.com/media/CLfu-_uWcAA-Fwo.png",
        "source_user_id_str": "255681367",
        "source_status_id": 628233440576557000,
        "id_str": "628233439066615808",
        "sizes": {
```



```
  },
  "in_reply_to_screen_name": null,
  "in_reply_to_user_id": null,
  "retweet_count": 3,
  "id_str": "628233440576557056",
  "favorited": false,
  "user": {
    "follow_request_sent": false,
    "has_extended_profile": false,
    "profile_use_background_image": true,
    "contributors_enabled": false,
    "id": 255681367,
    "verified": false,
    "profile_text_color": "C42D35",
    "profile_image_url_https": "https://pbs.twimg.com/profile_images/535079119696834560/epHn4PyO_normal.png",
    "profile_sidebar_fill_color": "C27222",
    "entities": {
      "url": {
        "urls": [
          {
            "url": "http://t.co/QdrsicBc2L",
            "indices": [
              0,
              22
            ],
            "expanded_url": "http://ianmilligan.ca",
            "display_url": "ianmilligan.ca"
          }
        ]
      },
      "description": {
        "urls": []
      }
    },
    "followers_count": 2530,
    "profile_sidebar_border_color": "919988",
    "location": "Waterloo, Ontario, Canada",
    "default_profile_image": false,
    "id_str": "255681367",
    "is_translation_enabled": false,
    "utc_offset": -10800,
    "statuses_count": 15421,
    "description": "Assistant professor of digital and Canadian history at @uWaterloo. Web archives, digital history, and textual analysis. Co-editor, @proghist.",
    "friends_count": 1420,
    "profile_link_color": "FA743E",
    "profile_image_url": "http://pbs.twimg.com/profile_images/535079119696834560/epHn4PyO_normal.png",
    "notifications": false,
    "geo_enabled": true,
    "profile_background_color": "5E4F49",
    "profile_banner_url": "https://pbs.twimg.com/profile_banners/255681367/1432589552",
    "profile_background_image_url": "http://pbs.twimg.com/profile_background_images/520587540/x3eb98c89e11038979bd2cb197af6269.jpg",
    "screen_name": "ianmilligan1",
    "lang": "en",
    "following": true,
    "profile_background_tile": true,
    "favourites_count": 947,
    "name": "Ian Milligan",
    "url": "http://t.co/QdrsicBc2L",
```

#elxn42



code{4}lib JOURNAL

Issue 32, 2016-04-25 ISSN 1940-5758

An Open-Source Strategy for Documenting Events: The Case Study of the 42nd Canadian Federal Election on Twitter

This article examines the tools, approaches, collaboration, and findings of the Web Archives for Historical Research Group around the capture and analysis of about 4 million tweets during the 2015 Canadian Federal Election. We hope that national libraries and other heritage institutions will find our model useful as they consider how to capture, preserve, and analyze ongoing events using Twitter.

While Twitter is not a representative sample of broader society – Pew research shows in their study of US users that it skews young, college-educated, and affluent (above $50,000 household income) – Twitter still represents an exponential increase in the amount of information generated, retained, and preserved from 'everyday' people. Therefore, when historians study the 2015 federal election, Twitter will be a prime source.

On August 3, 2015, the team initiated both a Search API and Stream API collection with twarc, a tool developed by Ed Summers, using the hashtag #elxn42. The hashtag referred to the election being Canada's 42nd general federal election (hence 'election 42' or elxn42). Data collection ceased on November 5, 2015, the day after Justin Trudeau was sworn in as the 42nd Prime Minister of Canada. We collected for a total of 102 days, 13 hours and 50 minutes.

To analyze the data set, we took advantage of a number of command line tools, utilities that are available within twarc, twarc-report, and `jq`. In accordance with the Twitter Developer Agreement & Policy, and after ethical deliberations discussed below, we made the tweet IDs and other derivative data available in a data repository. This allows other people to use our dataset, cite our dataset, and enhance their own research projects by drawing on #elxn42 tweets.

Our analytics included:

- *breaking tweet text down by day to track change over time;*
- *client analysis, allowing us to see how the scale of mobile devices affected medium interactions;*
- *URL analysis, comparing both to Archive-It collections and the Wayback Availability API to add to our understanding of crawl completeness;*
- *and image analysis, using an archive of extracted images.*

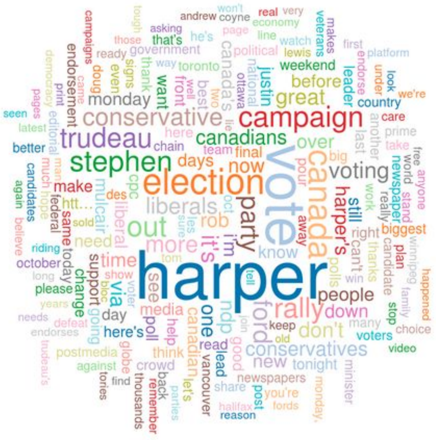
Our article introduces our collecting work, ethical considerations, the analysis we have done, and provides a framework for other collecting institutions to do similar work with our off-the-shelf open-source tools. We conclude by ruminating about connecting Twitter archiving with a broader web archiving strategy.

by Nick Ruest and Ian Milligan

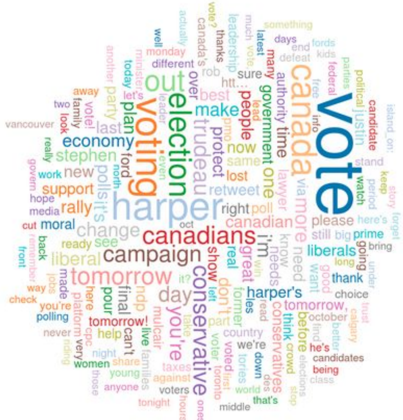
Introduction

During the 2015 Canadian federal elections, we captured 3,918,932 tweets written using the #elxn42 hashtag: thoughts on the nature and stature of political candidates or parties, live running commentary during leader debates, exhortations to vote, and witty ripostes or jokes to liven up the long campaign. Political scientists, journalists, and other researchers can use these tweets as evidence of sentiment amongst a

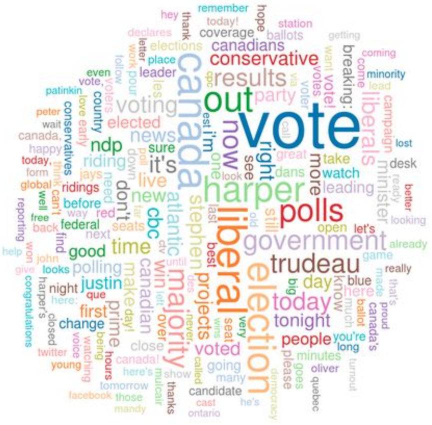
October 17, 2015



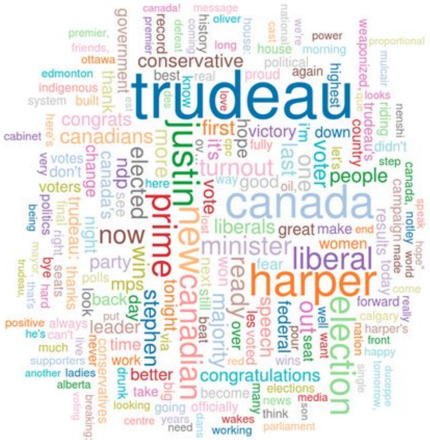
October 18, 2015

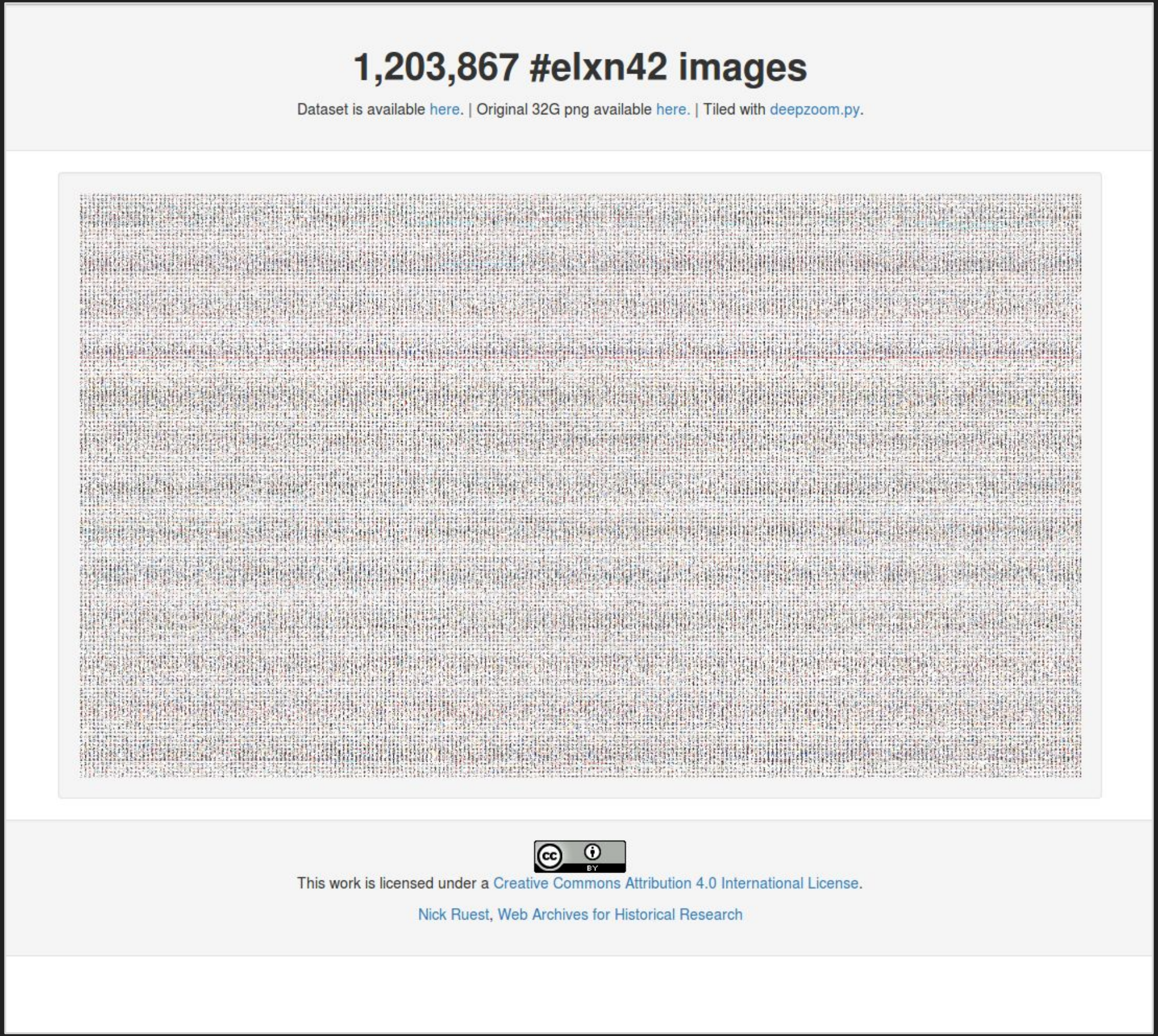


October 19, 2015



October 20, 2015

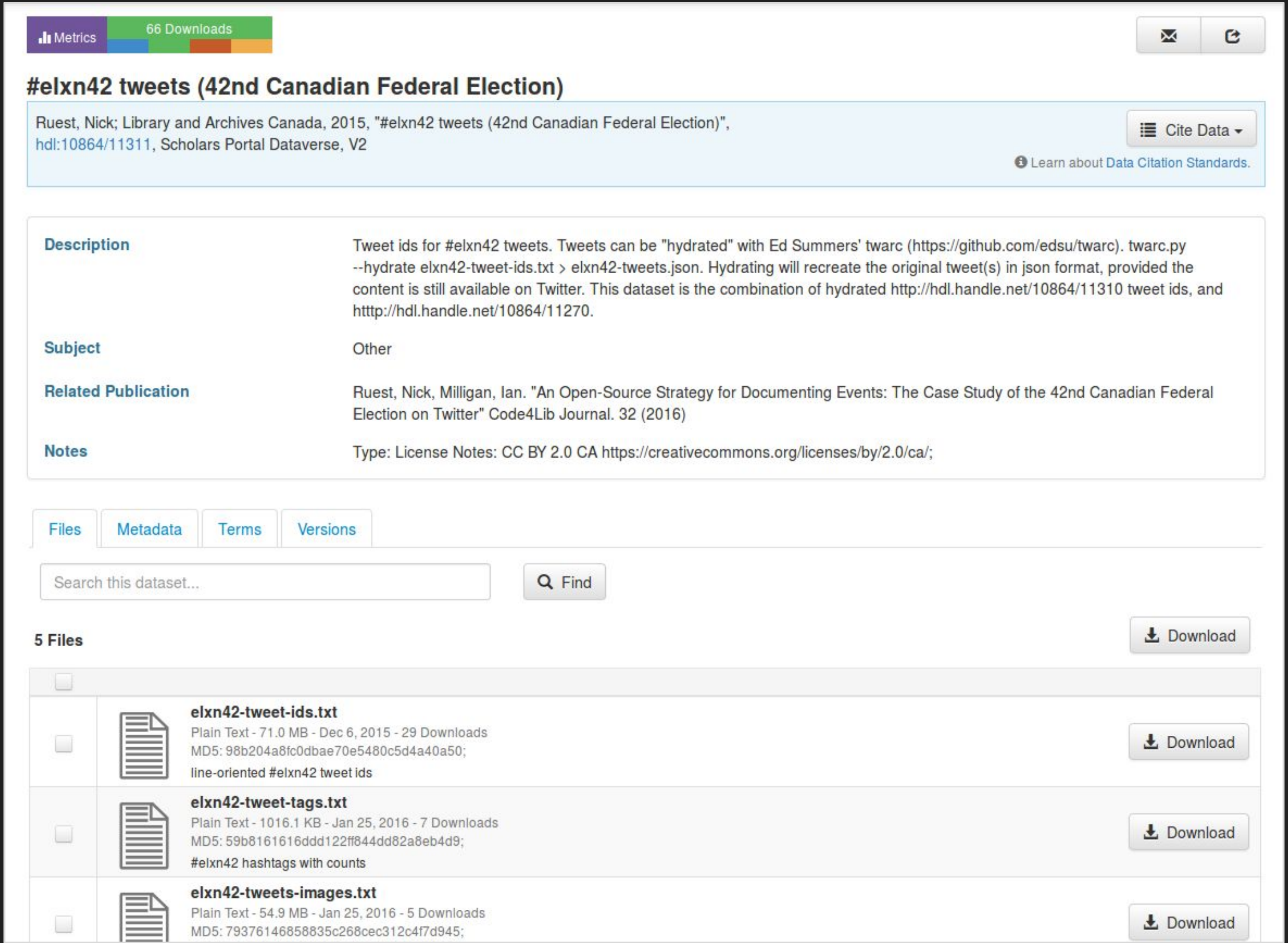




1,203,867 #elxn42 images

Dataset is available here. | Original 32G png available here. | Tiled with deepzoom.py.

Nick Ruest, Web Archives for Historical Research



#elxn42 tweets (42nd Canadian Federal Election)

Ruest, Nick; Library and Archives Canada, 2015, "#elxn42 tweets (42nd Canadian Federal Election)", hdl:10864/11311, Scholars Portal Dataverse, V2

Description Tweet ids for #elxn42 tweets. Tweets can be "hydrated" with Ed Summers' twarc (https://github.com/edsu/twarc). twarc.py --hydrate elxn42-tweet-ids.txt > elxn42-tweets.json. Hydrating will recreate the original tweet(s) in json format, provided the content is still available on Twitter. This dataset is the combination of hydrated http://hdl.handle.net/10864/11310 tweet ids, and htttp://hdl.handle.net/10864/11270.

Subject Other

Related Publication Ruest, Nick, Milligan, Ian. "An Open-Source Strategy for Documenting Events: The Case Study of the 42nd Canadian Federal Election on Twitter" Code4Lib Journal. 32 (2016)

Notes Type: License Notes: CC BY 2.0 CA https://creativecommons.org/licenses/by/2.0/ca/;

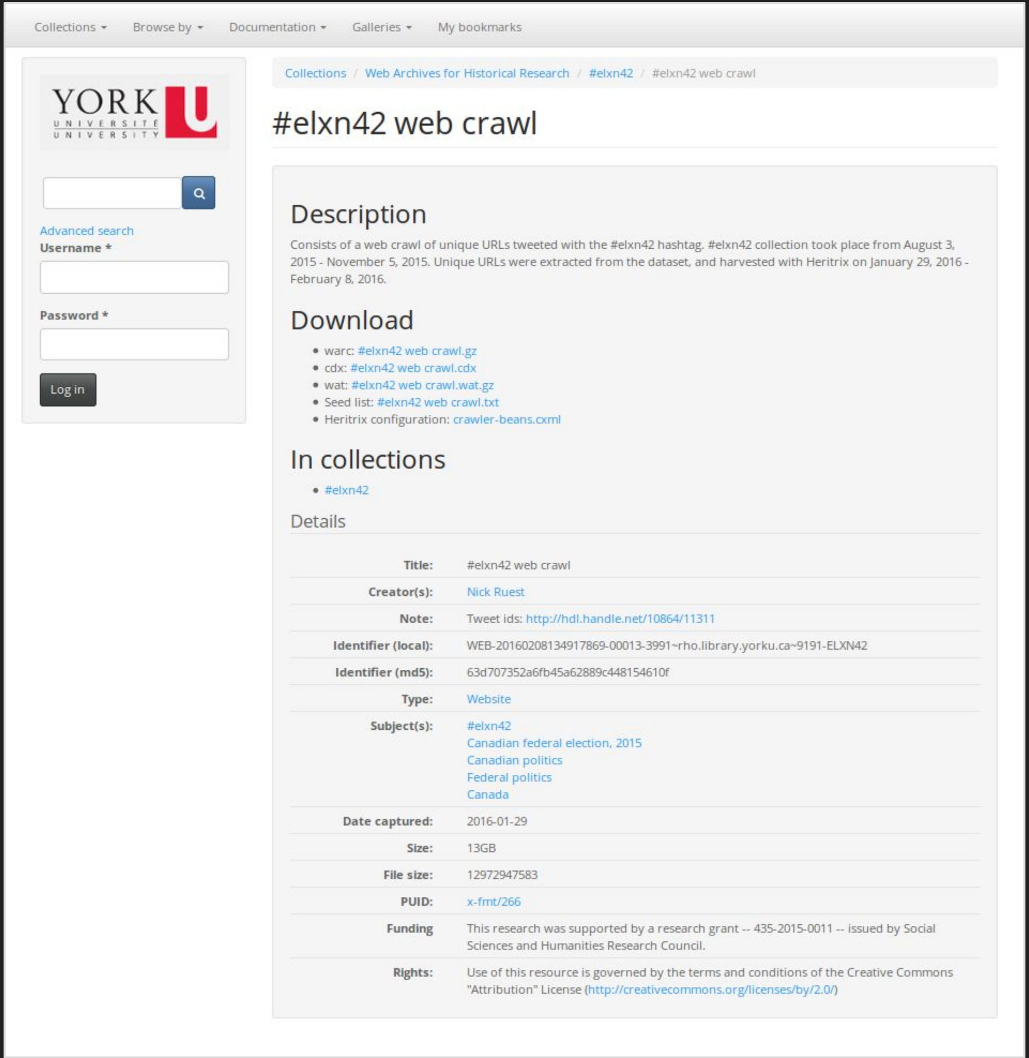
5 Files

□

□ **elxn42-tweet-ids.txt**
Plain Text - 71.0 MB - Dec 6, 2015 - 29 Downloads
MD5: 98b204a8fc0dbae70e5480c5d4a40a50;
line-oriented #elxn42 tweet ids

□ **elxn42-tweet-tags.txt**
Plain Text - 1016.1 KB - Jan 25, 2016 - 7 Downloads
MD5: 59b8161616ddd122ff844dd82a8eb4d9;
#elxn42 hashtags with counts

□ **elxn42-tweets-images.txt**
Plain Text - 54.9 MB - Jan 25, 2016 - 5 Downloads
MD5: 79376146858835c268cec312c4f7d945;



Collections / Web Archives for Historical Research / #elxn42 / #elxn42 web crawl

#elxn42 web crawl

Description

Consists of a web crawl of unique URLs tweeted with the #elxn42 hashtag. #elxn42 collection took place from August 3, 2015 - November 5, 2015. Unique URLs were extracted from the dataset, and harvested with Heritrix on January 29, 2016 - February 8, 2016.

Download

- warc: #elxn42 web crawl.gz
- cdx: #elxn42 web crawl.cdx
- wat: #elxn42 web crawl.wat.gz
- Seed list: #elxn42 web crawl.txt
- Heritrix configuration: crawler-beans.cxml

In collections

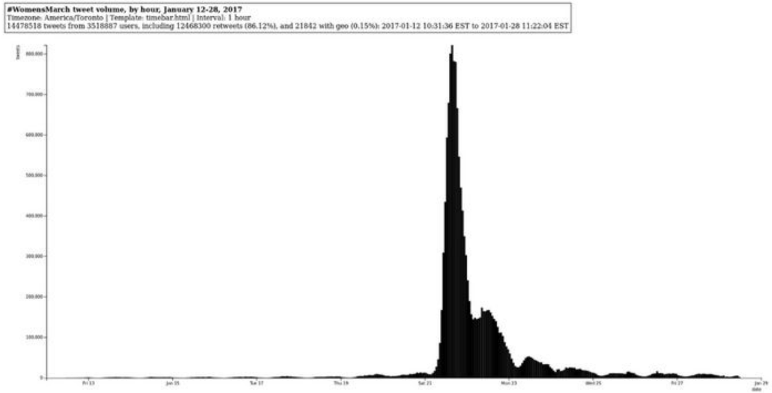
- #elxn42

Details

Title:	#elxn42 web crawl
Creator(s):	Nick Ruest
Note:	Tweet ids: http://hdl.handle.net/10864/11311
Identifier (local):	WEB-20160208134917869-00013-3991~rho.library.yorku.ca~9191-ELXN42
Identifier (md5):	63d707352a6fb45a62889c448154610f
Type:	Website
Subject(s):	#elxn42 Canadian federal election, 2015 Canadian politics Federal politics Canada
Date captured:	2016-01-29
Size:	13GB
File size:	12972947583
PUID:	x-fmt/266
Funding	This research was supported by a research grant -- 435-2015-0011 -- issued by Social Sciences and Humanities Research Council.
Rights:	Use of this resource is governed by the terms and conditions of the Creative Commons "Attribution" License (http://creativecommons.org/licenses/by/2.0/)

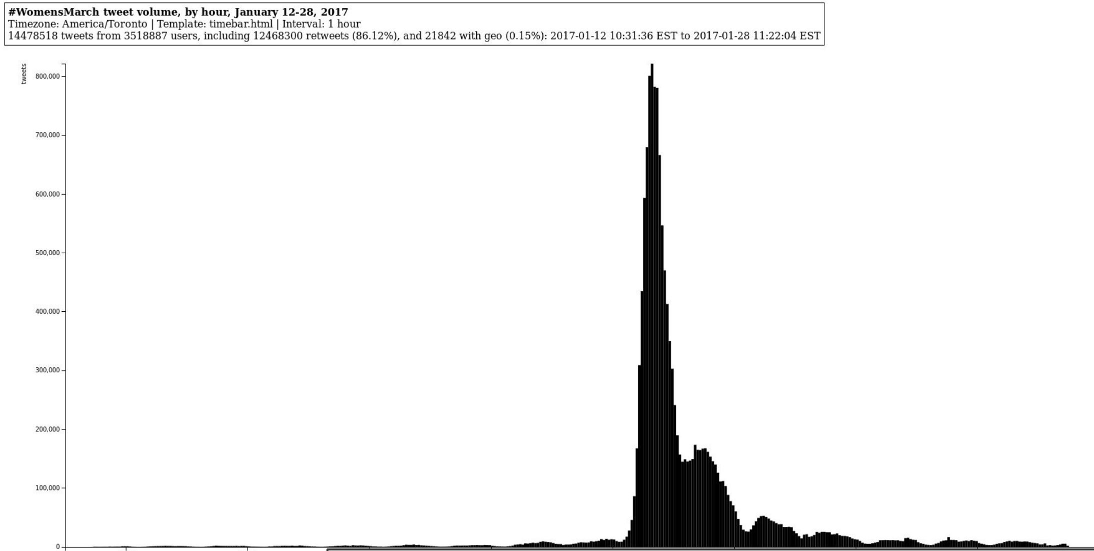
#WomensMarch

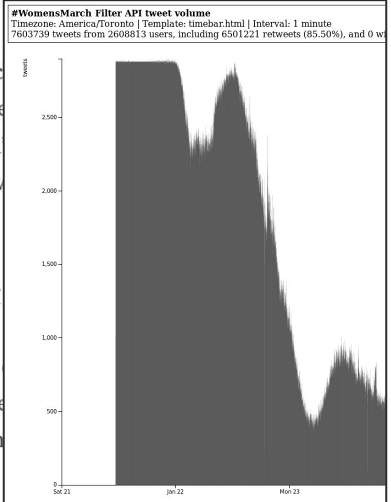
the time I stopped collecting a week later, we'd amassed 14,478,518 unique tweet ids from 3,582,495 unique users, and at one point hit around 1 million tweets in a single hour.

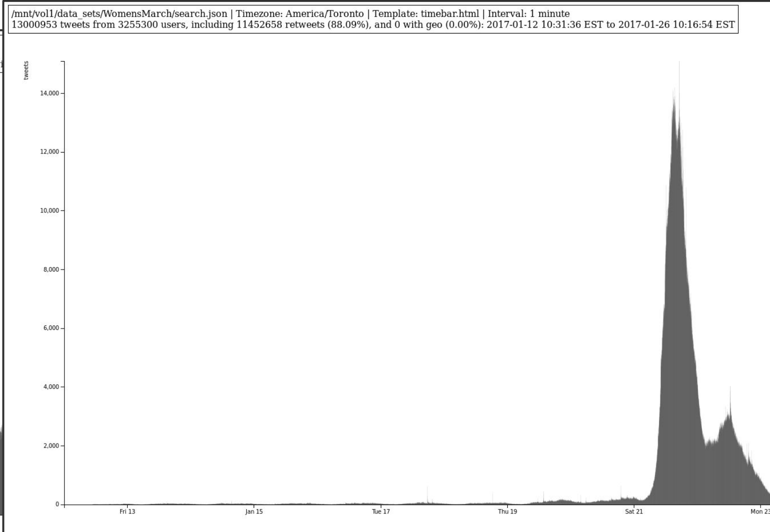

#WomensMarch tweet volume, by hour, January 12-28, 2017
Timezone: America/Toronto | Template: timebar.html | Interval: 1 hour
14478518 tweets from 3518887 users, including 12468300 retweets (86.12%), and 21842 with geo (0.15%): 2017-01-12 10:31:36 EST to 2017-01-28 11:22:04 EST

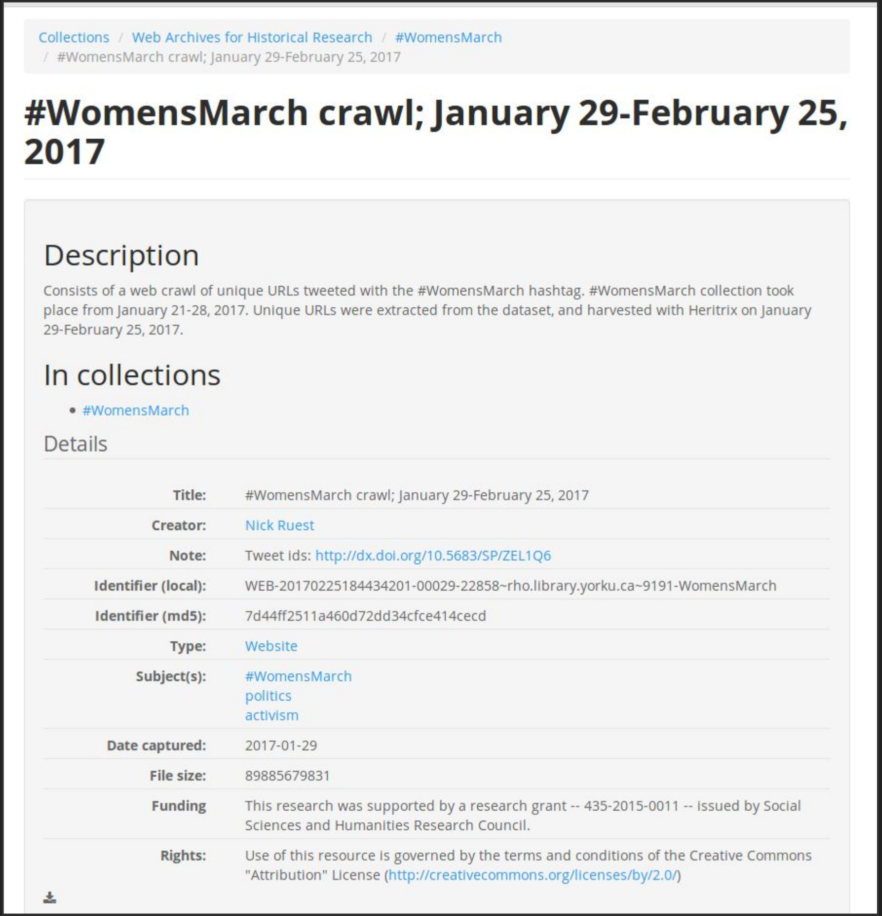
(Generated with Peter Binkley's twarc-report)

This put #WomensMarch well over 1% of the overall Twitter stream, whic causes dropped tweets if you're collecting from the Filter API, so I used the strategy of using the both the Filter and Search APIs for collection. (If you' curious about learning more about this, check out Kevin Driscoll and Shaw Walker's "Big Data, Big Questions | Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data", and Jiaul H. Paik and Jimmy Lin's "Do Multiple Listeners to the Public Twitter Sample Stream Receive the Same Tweets?). I've included the search and filter logs in the dataset. If you `grep "WARNING" WomensMarch_filter.log` `grep "WARNING" WomensMarch_filter.log | wc -l` you'll get a sense of the scale of dropped tweets. For a number of hours on January 22, I was seein around 1.6 million cumulative dropped tweets!


#WomensMarch tweet volume, by hour, January 12-28, 2017
Timezone: America/Toronto | Template: timebar.html | Interval: 1 hour
14478518 tweets from 3518887 users, including 12468300 retweets (86.12%), and 21842 with geo (0.15%): 2017-01-12 10:31:36 EST to 2017-01-28 11:22:04 EST


#WomensMarch Filter API tweet volume
Timezone: America/Toronto | Template: timebar.html | Interval: 1 minute
7603739 tweets from 2608813 users, including 6501221 retweets (85.50%), and 0 wi


/mnt/vol1/data_sets/WomensMarch/search.json | Timezone: America/Toronto | Template: timebar.html | Interval: 1 minute
13000953 tweets from 3255300 users, including 11452658 retweets (88.09%), and 0 with geo (0.00%): 2017-01-12 10:31:36 EST to 2017-01-26 10:16:54 EST



Collections / Web Archives for Historical Research / #WomensMarch / #WomensMarch crawl; January 29-February 25, 2017

#WomensMarch crawl; January 29-February 25, 2017

Description

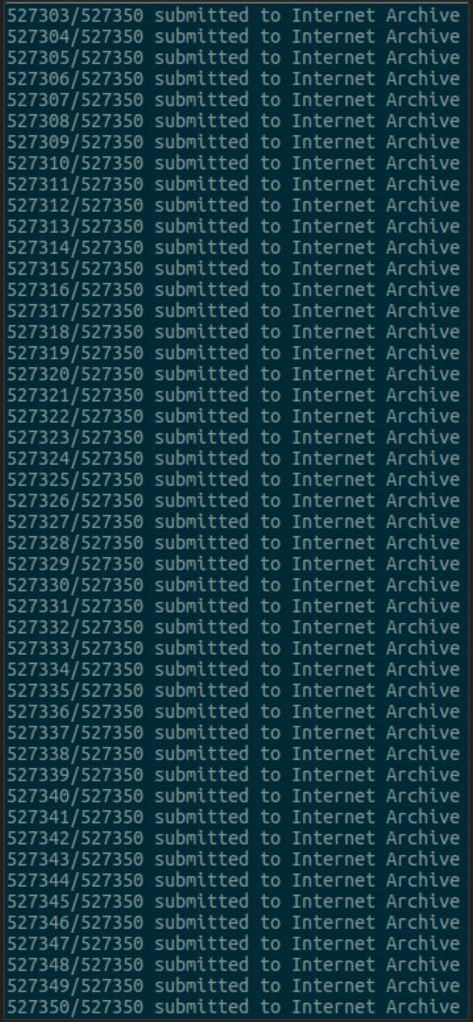
Consists of a web crawl of unique URLs tweeted with the #WomensMarch hashtag. #WomensMarch collection took place from January 21-28, 2017. Unique URLs were extracted from the dataset, and harvested with Heritrix on January 29-February 25, 2017.

In collections

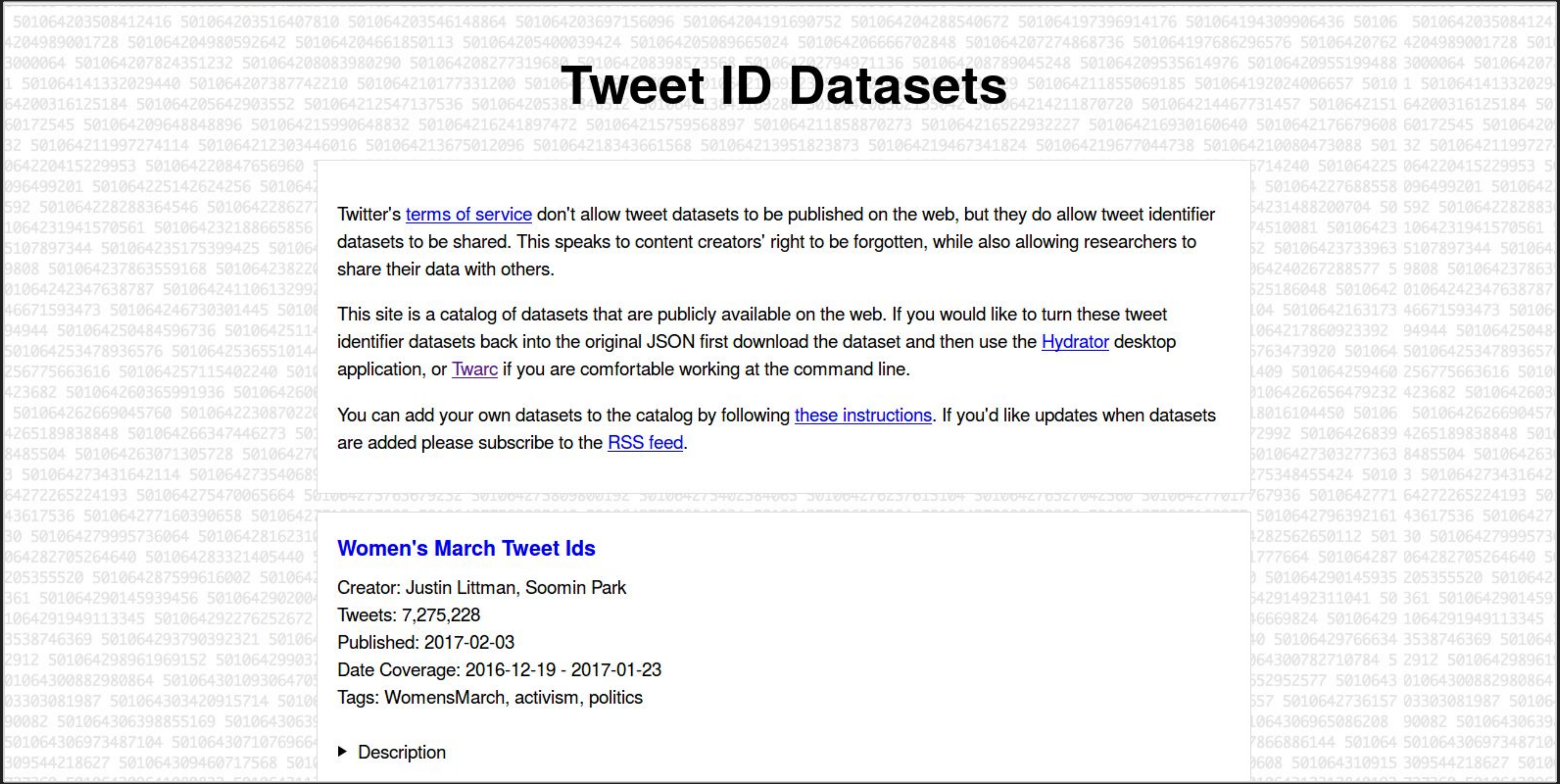
- #WomensMarch

Details

Title:	#WomensMarch crawl; January 29-February 25, 2017
Creator:	Nick Ruest
Note:	Tweet ids: http://dx.doi.org/10.5683/SP/ZEL1Q6
Identifier (local):	WEB-20170225184434201-00029-22858~rho.library.yorku.ca~9191-WomensMarch
Identifier (md5):	7d44ff2511a460d72dd34cfce414cecd
Type:	Website
Subject(s):	#WomensMarch politics activism
Date captured:	2017-01-29
File size:	89885679831
Funding	This research was supported by a research grant -- 435-2015-0011 -- issued by Social Sciences and Humanities Research Council.
Rights:	Use of this resource is governed by the terms and conditions of the Creative Commons "Attribution" License (http://creativecommons.org/licenses/by/2.0/)



```
527303/527350 submitted to Internet Archive
527304/527350 submitted to Internet Archive
527305/527350 submitted to Internet Archive
527306/527350 submitted to Internet Archive
527307/527350 submitted to Internet Archive
527308/527350 submitted to Internet Archive
527309/527350 submitted to Internet Archive
527310/527350 submitted to Internet Archive
527311/527350 submitted to Internet Archive
527312/527350 submitted to Internet Archive
527313/527350 submitted to Internet Archive
527314/527350 submitted to Internet Archive
527315/527350 submitted to Internet Archive
527316/527350 submitted to Internet Archive
527317/527350 submitted to Internet Archive
527318/527350 submitted to Internet Archive
527319/527350 submitted to Internet Archive
527320/527350 submitted to Internet Archive
527321/527350 submitted to Internet Archive
527322/527350 submitted to Internet Archive
527323/527350 submitted to Internet Archive
527324/527350 submitted to Internet Archive
527325/527350 submitted to Internet Archive
527326/527350 submitted to Internet Archive
527327/527350 submitted to Internet Archive
527328/527350 submitted to Internet Archive
527329/527350 submitted to Internet Archive
527330/527350 submitted to Internet Archive
527331/527350 submitted to Internet Archive
527332/527350 submitted to Internet Archive
527333/527350 submitted to Internet Archive
527334/527350 submitted to Internet Archive
527335/527350 submitted to Internet Archive
527336/527350 submitted to Internet Archive
527337/527350 submitted to Internet Archive
527338/527350 submitted to Internet Archive
527339/527350 submitted to Internet Archive
527340/527350 submitted to Internet Archive
527341/527350 submitted to Internet Archive
527342/527350 submitted to Internet Archive
527343/527350 submitted to Internet Archive
527344/527350 submitted to Internet Archive
527345/527350 submitted to Internet Archive
527346/527350 submitted to Internet Archive
527347/527350 submitted to Internet Archive
527348/527350 submitted to Internet Archive
527349/527350 submitted to Internet Archive
527350/527350 submitted to Internet Archive
```



Tweet ID Datasets

Twitter's terms of service don't allow tweet datasets to be published on the web, but they do allow tweet identifier datasets to be shared. This speaks to content creators' right to be forgotten, while also allowing researchers to share their data with others.

This site is a catalog of datasets that are publicly available on the web. If you would like to turn these tweet identifier datasets back into the original JSON first download the dataset and then use the Hydrator desktop application, or Twarc if you are comfortable working at the command line.

You can add your own datasets to the catalog by following these instructions. If you'd like updates when datasets are added please subscribe to the RSS feed.

Women's March Tweet Ids

Creator: Justin Littman, Soomin Park
Tweets: 7,275,228
Published: 2017-02-03
Date Coverage: 2016-12-19 - 2017-01-23
Tags: WomensMarch, activism, politics

▸ Description

This *a* future of collection development

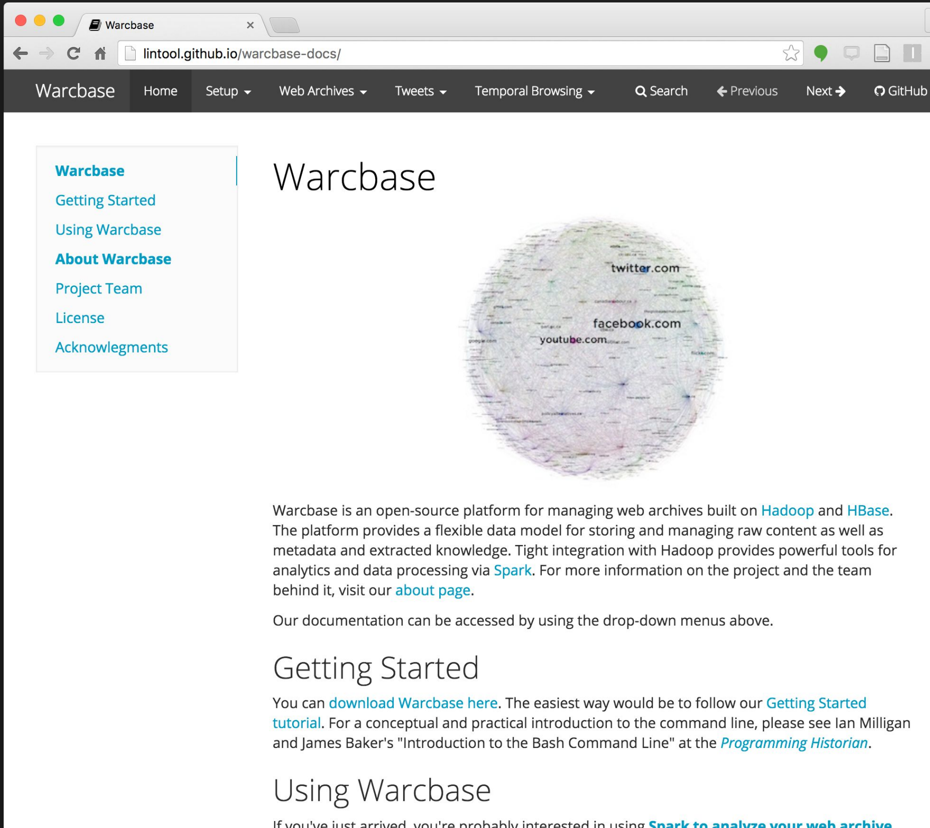
Institutional Collecting & Research Data

Warcbase

(Or, how can we work with web archives at scale?)

Warcbase

- Jimmy Lin (main developer, CS/lead), Ian Milligan (co-lead, history), Jeremy Wiebe (history/PhD), Alice Zhou (computer science, undergrad), Youngbin Kim (computer science, undergrad), Nick Ruest (librarian)
- Currently using it on the GeoCities and Canadian Politics web archives



Warcbase

Warcbase is an open-source platform for managing web archives built on Hadoop and HBase. The platform provides a flexible data model for storing and managing raw content as well as metadata and extracted knowledge. Tight integration with Hadoop provides powerful tools for analytics and data processing via Spark. For more information on the project and the team behind it, visit our about page.

Our documentation can be accessed by using the drop-down menus above.

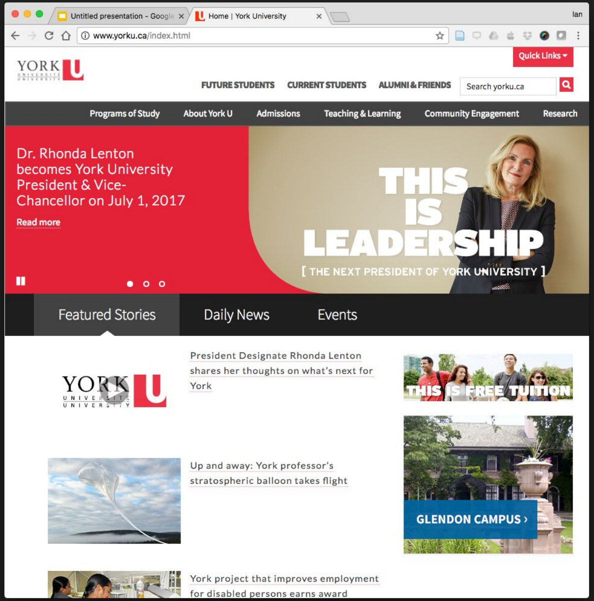
Getting Started

You can download Warcbase here. The easiest way would be to follow our Getting Started tutorial. For a conceptual and practical introduction to the command line, please see Ian Milligan and James Baker's "Introduction to the Bash Command Line" at the *Programming Historian*.

Using Warcbase

If you've just arrived, you're probably interested in using **Spark to analyze your web archive**

Why Warcbase



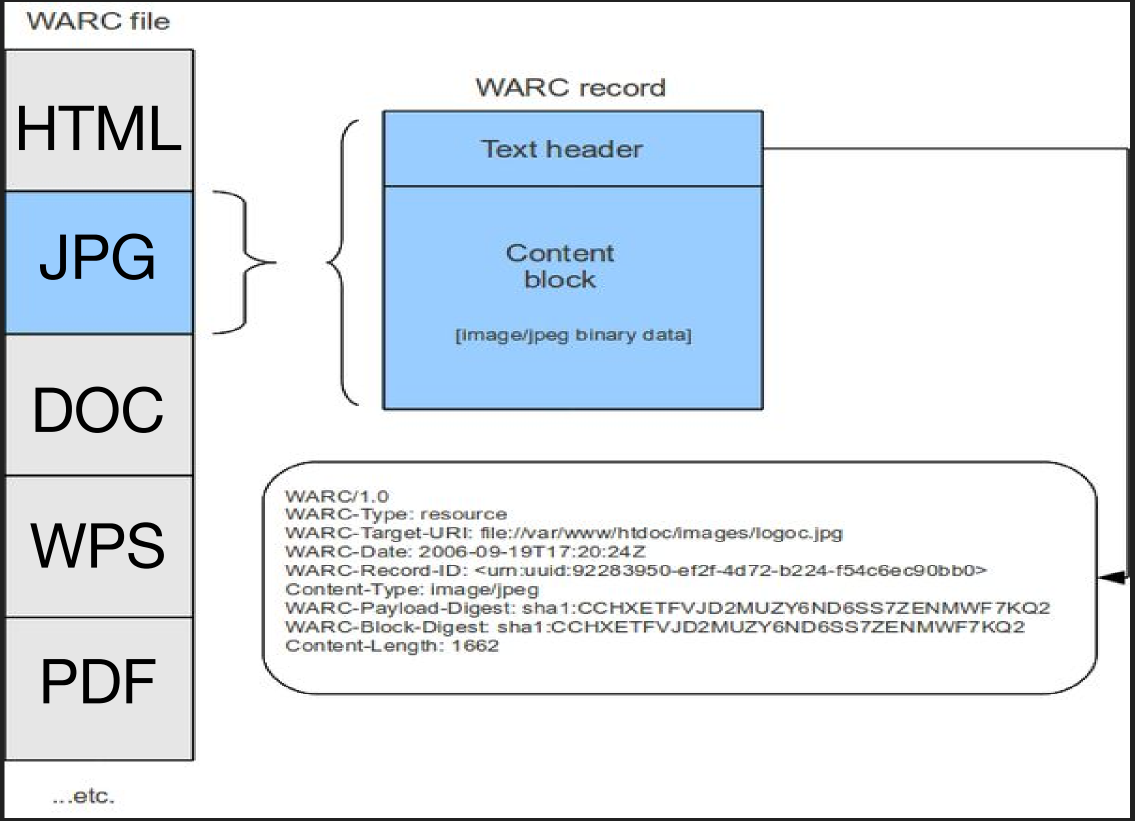
Dr. Rhonda Lenton becomes York University President & Vice-Chancellor on July 1, 2017

Read more

President Designate Rhonda Lenton shares her thoughts on what's next for York

Up and away: York professor's stratospheric balloon takes flight

York project that improves employment for disabled persons earns award



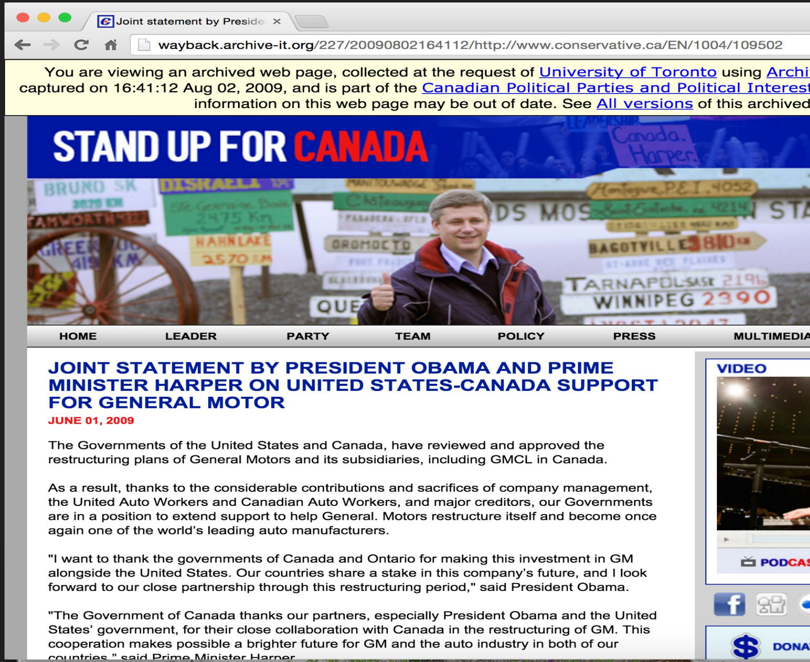
```
WARC/1.0
WARC-Type: resource
WARC-Target-URI: file://var/www/htdoc/images/logoc.jpg
WARC-Date: 2006-09-19T17:20:24Z
WARC-Record-ID: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>
Content-Type: image/jpeg
WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
WARC-Block-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
Content-Length: 1662
```

Warcbase (warcbase.org)

Two main facets

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

Funded by Mellon, SSHRC, NSERC, and Government of Ontario.



Joint statement by Preside
wayback.archive-it.org/227/20090802164112/http://www.conservative.ca/EN/1004/109502

You are viewing an archived web page, collected at the request of University of Toronto using Archi
captured on 16:41:12 Aug 02, 2009, and is part of the Canadian Political Parties and Political Interest
information on this web page may be out of date. See All versions of this archived

STAND UP FOR CANADA
HOME
LEADER
PARTY
TEAM
POLICY
PRESS
MULTIMEDIA
VIDEO
PODCAS
DONA

JOINT STATEMENT BY PRESIDENT OBAMA AND PRIME MINISTER HARPER ON UNITED STATES-CANADA SUPPORT FOR GENERAL MOTOR

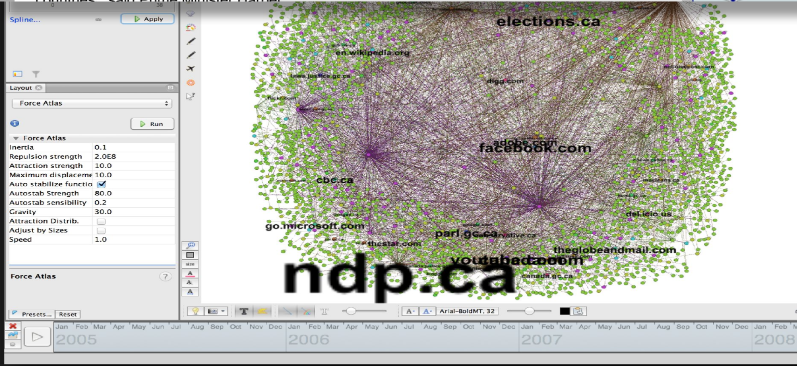
JUNE 01, 2009

The Governments of the United States and Canada, have reviewed and approved the restructuring plans of General Motors and its subsidiaries, including GMCL in Canada.

As a result, thanks to the considerable contributions and sacrifices of company management, the United Auto Workers and Canadian Auto Workers, and major creditors, our Governments are in a position to extend support to help General. Motors restructure itself and become once again one of the world's leading auto manufacturers.

"I want to thank the governments of Canada and Ontario for making this investment in GM alongside the United States. Our countries share a stake in this company's future, and I look forward to our close partnership through this restructuring period," said President Obama.

"The Government of Canada thanks our partners, especially President Obama and the United States' government, for their close collaboration with Canada in the restructuring of GM. This cooperation makes possible a brighter future for GM and the auto industry in both of our countries," said Prime Minister Harper



Force Atlas

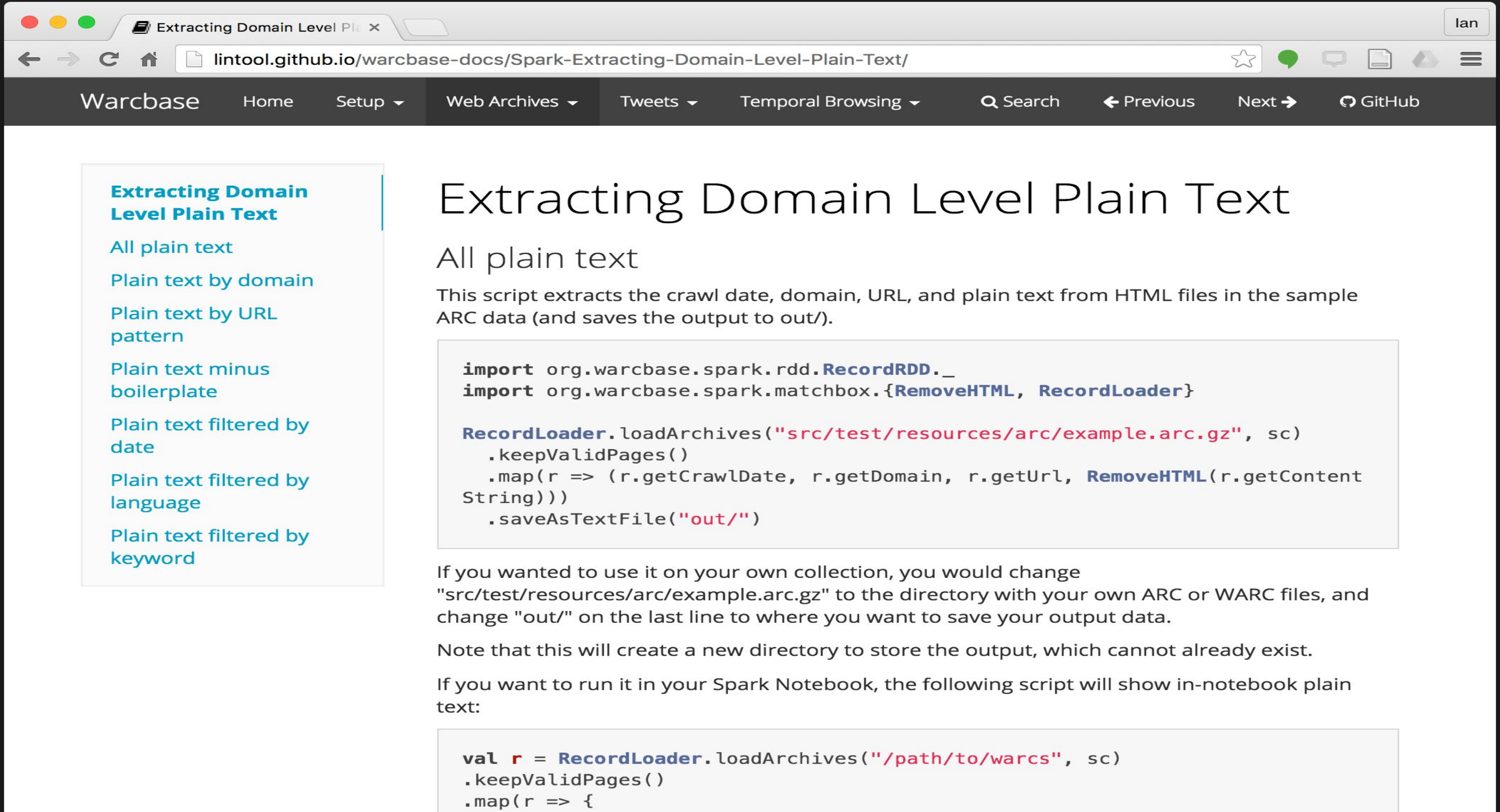
Force Atlas	
Inertia	0.1
Repulsion strength	2.0E8
Attraction strength	10.0
Maximum displaceme	10.0
Auto stabilize functio	☑
Autostab Strength	80.0
Autostab sensibility	0.2
Gravity	30.0
Attraction Distrib.	□
Adjust by Sizes	□
Speed	1.0

Force Atlas

Warcbase

- Scalable
  - From Raspberry Pi to Desktop Computer to Server to Cluster, **all with same scripts and commands**
- Potentially very powerful
  - **Trantor**: 1.2PB of disk, 25 compute nodes (each w/ 128GB memory, 2×6-core Intel Xeon E5 v3 = 3.2TB memory and 300 current-generation Intel cores)
- In active development, led by **Jimmy Lin**, collaborator at the University of Waterloo

docs.warcbase.org



Extracting Domain Level Pl
Ian
lintool.github.io/warcbase-docs/Spark-Extracting-Domain-Level-Plain-Text/
Warcbase
Home
Setup
Web Archives
Tweets
Temporal Browsing
Search
Previous
Next
GitHub
Extracting Domain Level Plain Text
All plain text
Plain text by domain
Plain text by URL pattern
Plain text minus boilerplate
Plain text filtered by date
Plain text filtered by language
Plain text filtered by keyword

Extracting Domain Level Plain Text

All plain text

This script extracts the crawl date, domain, URL, and plain text from HTML files in the sample ARC data (and saves the output to out/).

```
import org.warcbase.spark.rdd.RecordRDD._
import org.warcbase.spark.matchbox.{RemoveHTML, RecordLoader}

RecordLoader.loadArchives("src/test/resources/arc/example.arc.gz", sc)
  .keepValidPages()
  .map(r => (r.getCrawlDate, r.getDomain, r.getUrl, RemoveHTML(r.getContent
String)))
  .saveAsTextFile("out/")
```

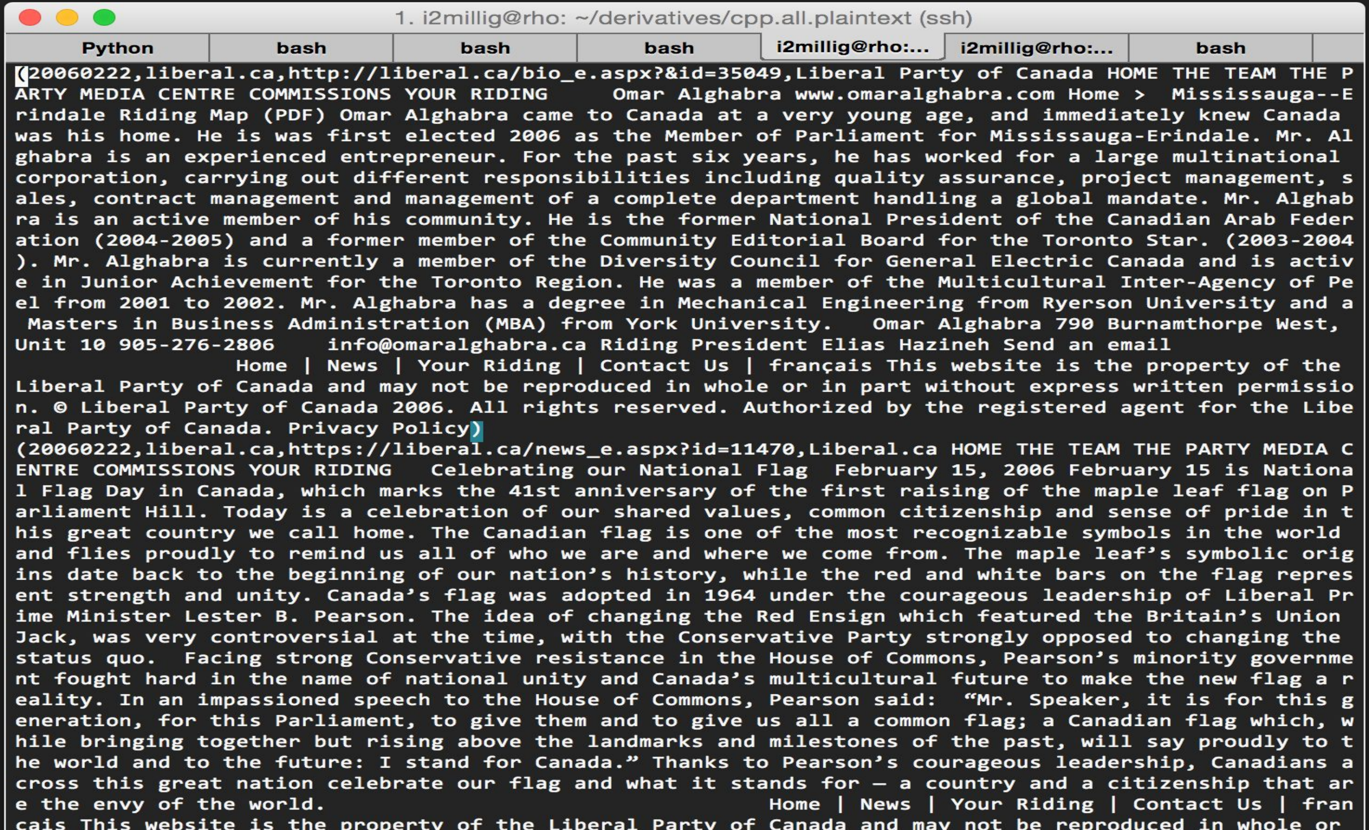
If you wanted to use it on your own collection, you would change "src/test/resources/arc/example.arc.gz" to the directory with your own ARC or WARC files, and change "out/" on the last line to where you want to save your output data.

Note that this will create a new directory to store the output, which cannot already exist.

If you want to run it in your Spark Notebook, the following script will show in-notebook plain text:

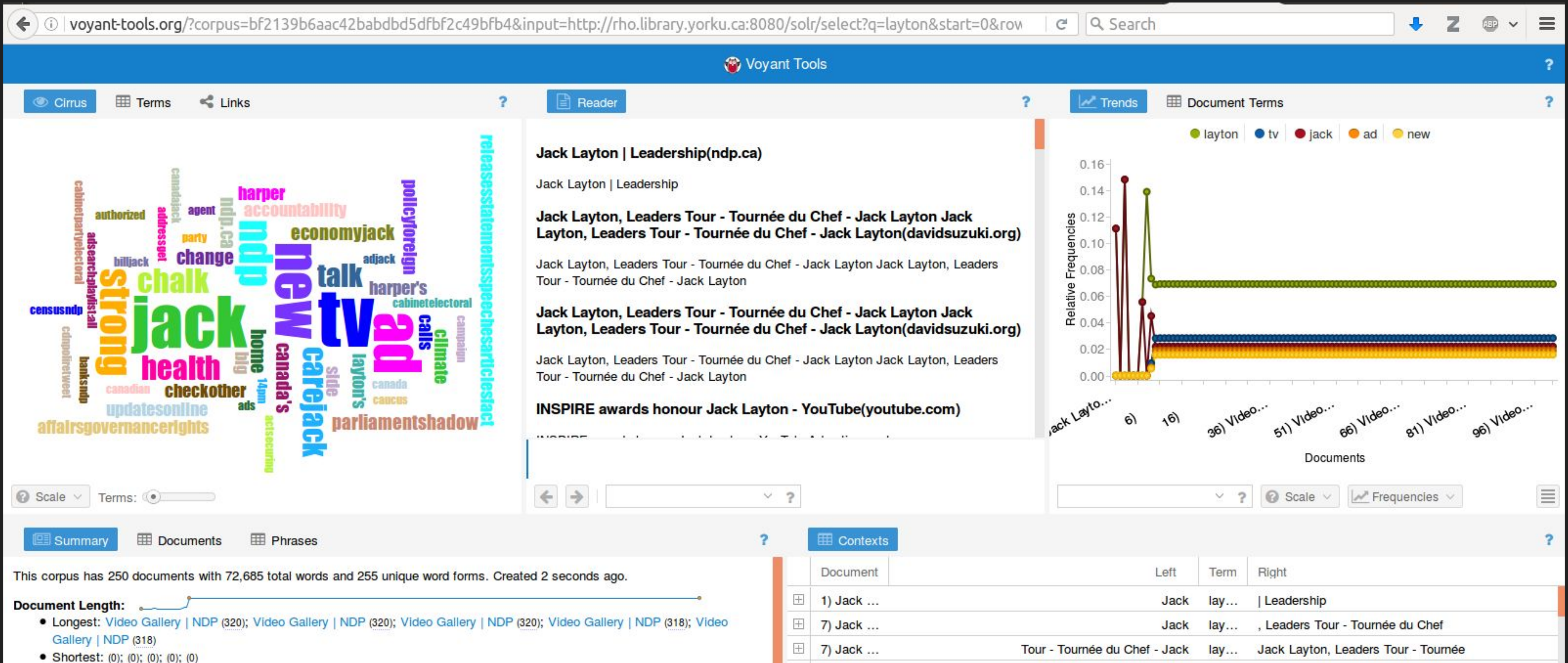
```
val r = RecordLoader.loadArchives("/path/to/warcs", sc)
.keepValidPages()
.map(r => {
```

Extract all Text



```
(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 is 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
el 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 permissio
n. © 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 all Text



Jack Layton | Leadership(ndp.ca)

Jack Layton | Leadership

Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton(davidsuzuki.org)

Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton

Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton(davidsuzuki.org)

Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton

INSPIRE awards honour Jack Layton - YouTube(youtube.com)

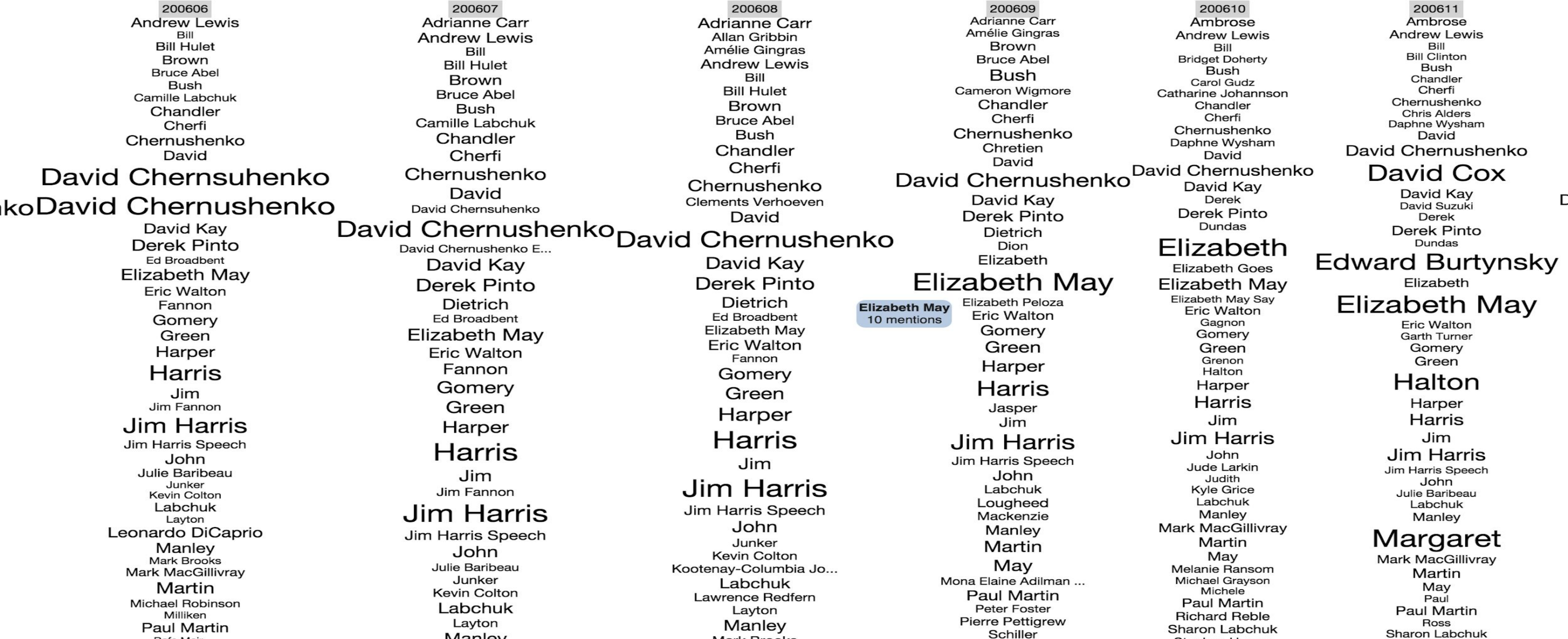
This corpus has 250 documents with 72,685 total words and 255 unique word forms. Created 2 seconds ago.

Document Length:

- Longest: Video Gallery | NDP (320); Video Gallery | NDP (320); Video Gallery | NDP (320); Video Gallery | NDP (318); Video Gallery | NDP (318)
- Shortest: (0); (0); (0); (0); (0)

	Document	Left	Term	Right
⊞	1) Jack …	Jack	lay…	| Leadership
⊞	7) Jack …	Jack	lay…	, Leaders Tour - Tournée du Chef
⊞	7) Jack …	Tour - Tournée du Chef - Jack	lay…	Jack Layton, Leaders Tour - Tournée

Extract Entities



200606
Andrew Lewis
Bill
Bill Hulet
Brown
Bruce Abel
Bush
Camille Labchuk
Chandler
Cherfi
Chernushenko
David
David Chernsuhenko
David Chernushenko
David Kay
Derek Pinto
Ed Broadbent
Elizabeth May
Eric Walton
Fannon
Gomery
Green
Harper
Harris
Jim
Jim Fannon
Jim Harris
Jim Harris Speech
John
Julie Baribeau
Junker
Kevin Colton
Labchuk
Layton
Leonardo DiCaprio
Manley
Mark Brooks
Mark MacGillivray
Martin
Michael Robinson
Milliken
Paul Martin

200607
Adrianne Carr
Andrew Lewis
Bill
Bill Hulet
Brown
Bruce Abel
Bush
Camille Labchuk
Chandler
Cherfi
Chernushenko
David
David Chernsuhenko
David Chernushenko
David Chernushenko E...
David Kay
Derek Pinto
Dietrich
Ed Broadbent
Elizabeth May
Eric Walton
Fannon
Gomery
Green
Harper
Harris
Jim
Jim Fannon
Jim Harris
Jim Harris Speech
John
Julie Baribeau
Junker
Kevin Colton
Labchuk
Layton
Manley

200608
Adrianne Carr
Allan Gribbin
Amélie Gingras
Andrew Lewis
Bill
Bill Hulet
Brown
Bruce Abel
Bush
Chandler
Cherfi
Chernushenko
Clements Verhoeven
David
David Chernushenko
David Kay
Derek Pinto
Dietrich
Ed Broadbent
Elizabeth May
Eric Walton
Fannon
Gomery
Green
Harper
Harris
Jim
Jim Harris
Jim Harris Speech
John
Junker
Kevin Colton
Kootenay-Columbia Jo...
Labchuk
Lawrence Redfern
Layton
Manley
Mark Brooks

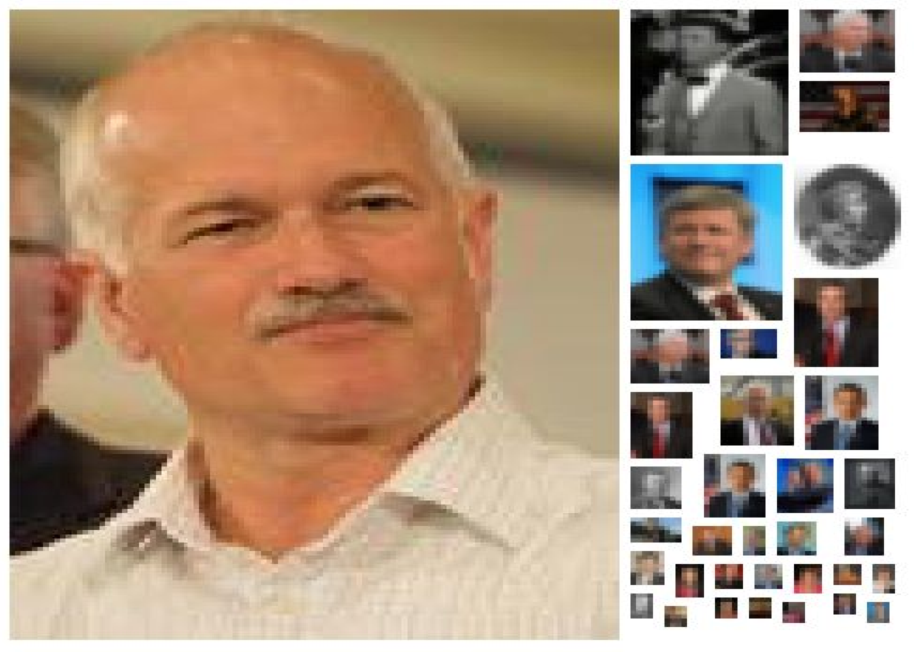
200609
Adrianne Carr
Amélie Gingras
Brown
Bruce Abel
Bush
Cameron Wigmore
Chandler
Cherfi
Chernushenko
Chretien
David
David Chernushenko
David Kay
Derek Pinto
Dietrich
Dion
Elizabeth
Elizabeth May
Elizabeth Peloza
Eric Walton
Gomery
Green
Harper
Harris
Jasper
Jim
Jim Harris
Jim Harris Speech
John
Labchuk
Lougheed
Mackenzie
Manley
Martin
May
Mona Elaine Adilman ...
Paul Martin
Peter Foster
Pierre Pettigrew
Schiller

Elizabeth May
10 mentions

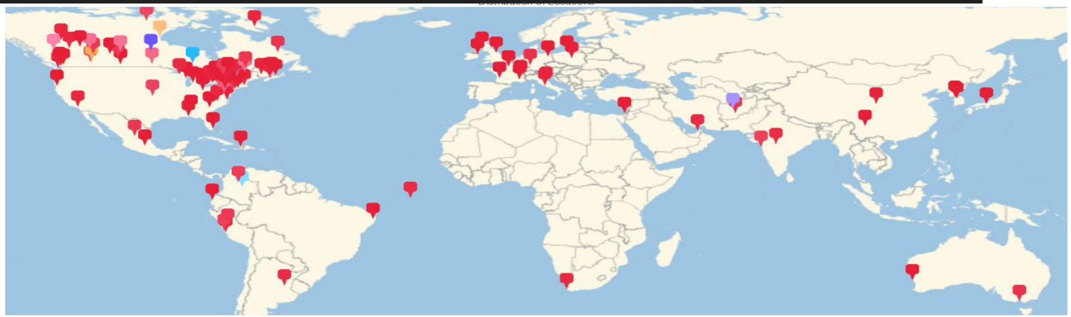
200610
Ambrose
Andrew Lewis
Bill
Bridget Doherty
Bush
Carol Gudz
Catharine Johannson
Chandler
Cherfi
Chernushenko
Daphne Wysham
David
David Chernushenko
David Kay
Derek
Derek Pinto
Dundas
Elizabeth
Elizabeth Goes
Elizabeth May
Elizabeth May Say
Eric Walton
Gagnon
Gomery
Green
Grenon
Halton
Harper
Harris
Jim
Jim Harris
John
Jude Larkin
Judith
Kyle Grice
Labchuk
Manley
Mark MacGillivray
Martin
May
Melanie Ransom
Michael Grayson
Michele
Paul Martin
Richard Reble
Sharon Labchuk

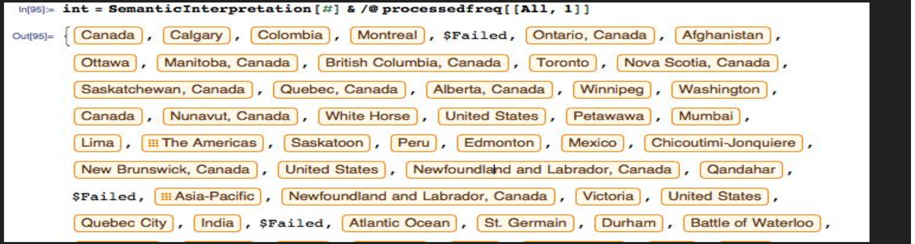
200611
Ambrose
Andrew Lewis
Bill
Bill Clinton
Bush
Chandler
Cherfi
Chernushenko
Chris Alders
Daphne Wysham
David
David Chernushenko
David Cox
David Kay
David Suzuki
Derek
Derek Pinto
Dundas
Edward Burtynsky
Elizabeth
Elizabeth May
Eric Walton
Garth Turner
Gomery
Green
Halton
Harper
Harris
Jim
Jim Harris
Jim Harris Speech
John
Julie Baribeau
Labchuk
Manley
Margaret
Mark MacGillivray
Martin
May
Paul
Paul Martin
Ross
Sharon Labchuk

Extract Entities

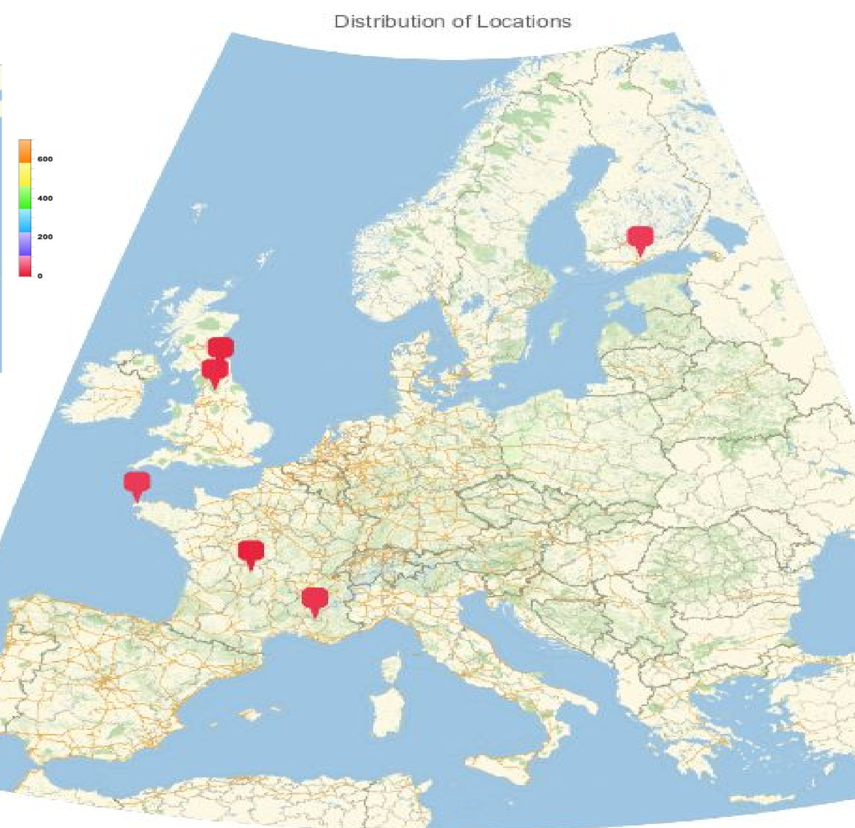


Extract Entities





```
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 ,
```



And a move away from content and towards **structured metadata**...

An Example

To: Tony Smith <tsmith@yorku.ca>

From: Ian Milligan
(i2millig@uwaterloo.ca)

Re: Lunch

Date: 6 March @ 1034 GMT

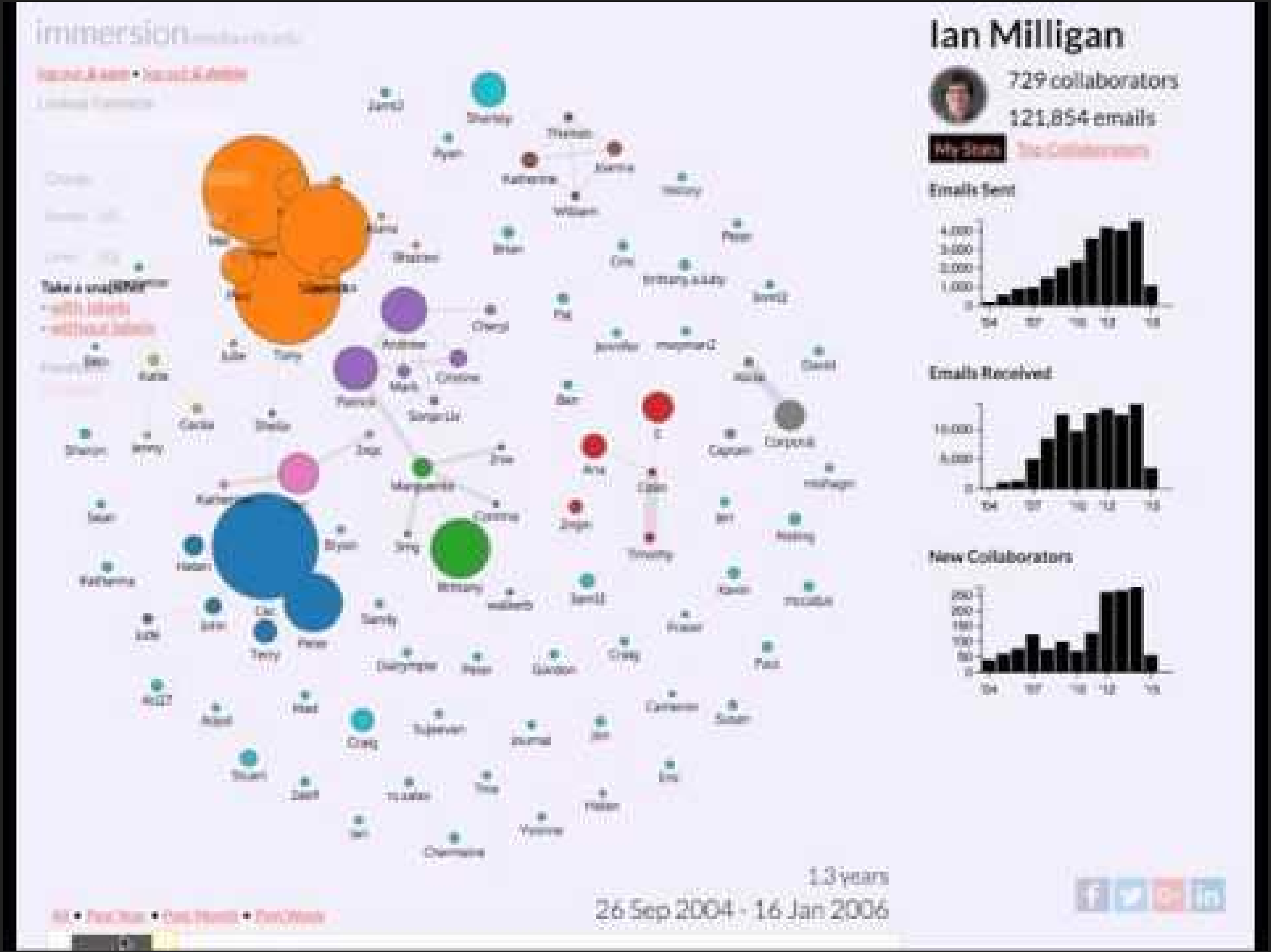
Hi Tony –

See you after class?

Ian

Tells you little!

But what if I emailed him every Monday? Or Friday? Or every day?



immersion

Take a snapshot
- with labels
- without labels

Ian Milligan

729 collaborators

121,854 emails

My Stats Top Collaborators

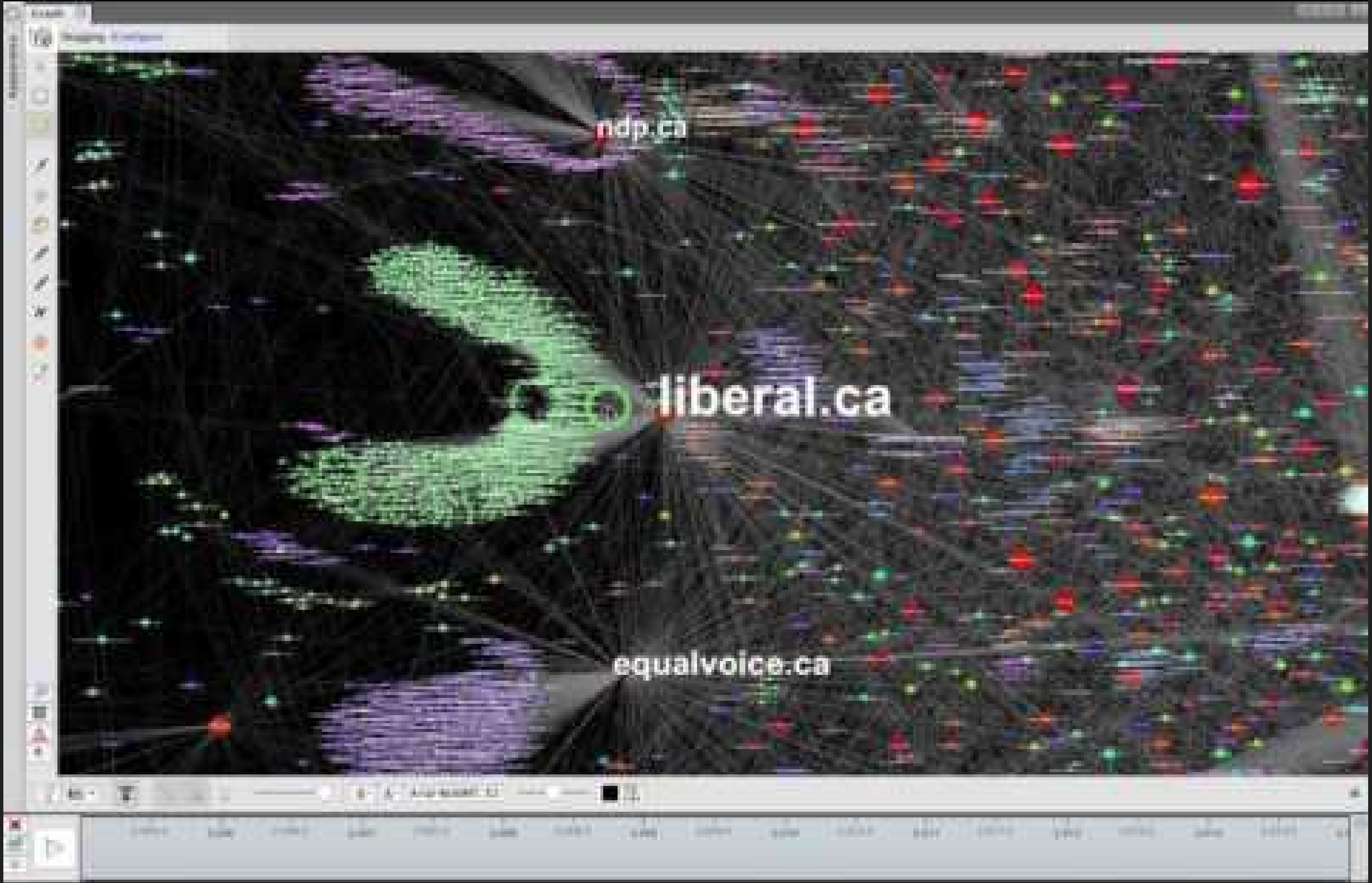
Emails Sent

Emails Received

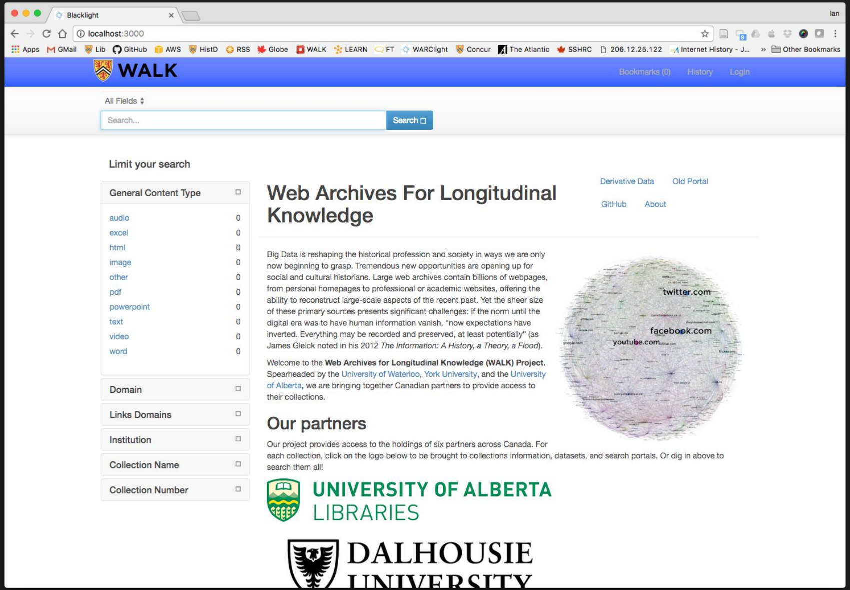
New Collaborators

1.3 years

26 Sep 2004 - 16 Jan 2006


ndp.ca
liberal.ca
equalvoice.ca

… and we can **create indexes** to provide search access to collections.



Blacklight
localhost:3000
Apps GMail Lib GitHub AWS HistD RSS Globe WALK LEARN FT WARClight Concur The Atlantic SSHRC 206.12.25.122 Internet History - J... Other Bookmarks
Ian
WALK
Bookmarks (0) History Login
All Fields
Search...
Search
Derivative Data Old Portal
GitHub About
twitter.com
facebook.com
youtube.com

Limit your search

General Content Type	□
audio	0
excel	0
html	0
image	0
other	0
pdf	0
powerpoint	0
text	0
video	0
word	0

Domain □

Links Domains □

Institution □

Collection Name □

Collection Number □

Web Archives For Longitudinal Knowledge

Big Data is reshaping the historical profession and society in ways we are only now beginning to grasp. Tremendous new opportunities are opening up for social and cultural historians. Large web archives contain billions of webpages, from personal homepages to professional or academic websites, offering the ability to reconstruct large-scale aspects of the recent past. Yet the sheer size of these primary sources presents significant challenges: if the norm until the digital era was to have human information vanish, "now expectations have inverted. Everything may be recorded and preserved, at least potentially" (as James Gleick noted in his 2012 *The Information: A History, a Theory, a Flood*).

Welcome to the **Web Archives for Longitudinal Knowledge (WALK) Project.** Spearheaded by the University of Waterloo, York University, and the University of Alberta, we are bringing together Canadian partners to provide access to their collections.

Our partners

Our project provides access to the holdings of six partners across Canada. For each collection, click on the logo below to be brought to collections information, datasets, and search portals. Or dig in above to search them all!

UNIVERSITY OF ALBERTA LIBRARIES

DALHOUSIE UNIVERSITY

So we have this tool…

how have we used it?

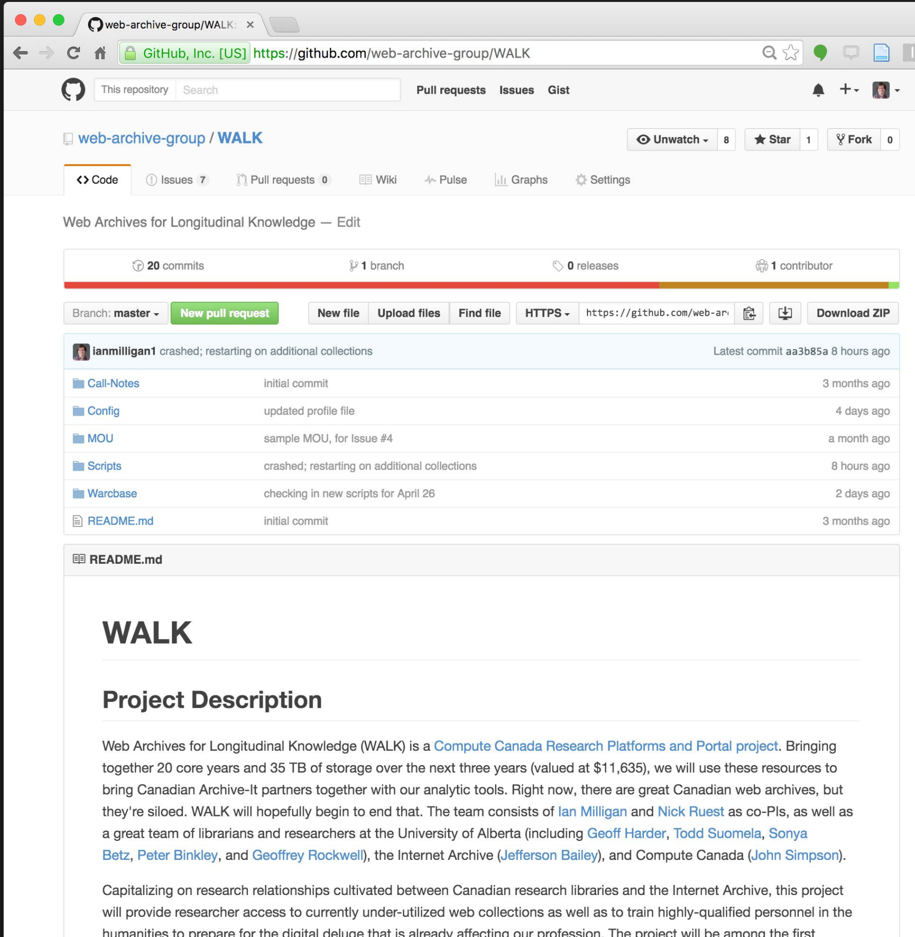
WALK Project

Web Archives for Longitudinal Knowledge

WALK

Web Archives for Longitudinal Knowledge (WALK)

Ian Milligan (Co-PI, UW) + Nick Ruest (Co-PI, York), w/ Geoff Harder, Todd Suomela, Sonya Betz, Peter Binkley, Geoffrey Rockwell (Alberta), Jefferson Bailey (Internet Archive), and John Simpson (Compute Canada).



Web Archives for Longitudinal Knowledge — Edit

ianmilligan1 crashed; restarting on additional collections — Latest commit aa3b85a 8 hours ago

Call-Notes	initial commit	3 months ago
Config	updated profile file	4 days ago
MOU	sample MOU, for Issue #4	a month ago
Scripts	crashed; restarting on additional collections	8 hours ago
Warcbase	checking in new scripts for April 26	2 days ago
README.md	initial commit	3 months ago

README.md

WALK

Project Description

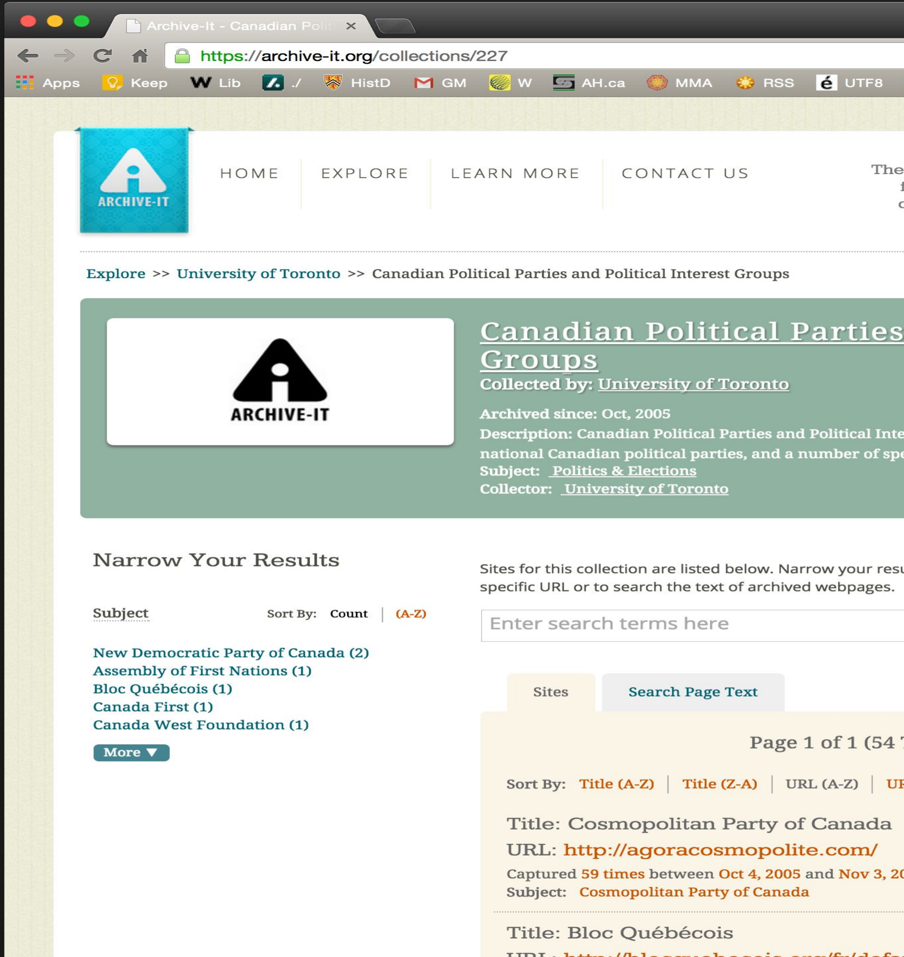
Web Archives for Longitudinal Knowledge (WALK) is a Compute Canada Research Platforms and Portal project. Bringing together 20 core years and 35 TB of storage over the next three years (valued at $11,635), we will use these resources to bring Canadian Archive-It partners together with our analytic tools. Right now, there are great Canadian web archives, but they're siloed. WALK will hopefully begin to end that. The team consists of Ian Milligan and Nick Ruest as co-PIs, as well as a great team of librarians and researchers at the University of Alberta (including Geoff Harder, Todd Suomela, Sonya Betz, Peter Binkley, and Geoffrey Rockwell), the Internet Archive (Jefferson Bailey), and Compute Canada (John Simpson).

Capitalizing on research relationships cultivated between Canadian research libraries and the Internet Archive, this project will provide researcher access to currently under-utilized web collections as well as to train highly-qualified personnel in the humanities to prepare for the digital deluge that is already affecting our profession. The project will be among the first

WALK

Currently ~ 25 Canadian partners who collect web archives, ~ 130 collections.

Archive-It is the back-end provider of web archiving services.



Explore >> University of Toronto >> Canadian Political Parties and Political Interest Groups

Canadian Political Parties
Groups

Collected by: University of Toronto

Archived since: Oct, 2005

Description: Canadian Political Parties and Political Inte
national Canadian political parties, and a number of spe

Subject: Politics & Elections

Collector: University of Toronto

Narrow Your Results

Subject | Sort By: Count | (A-Z)

New Democratic Party of Canada (2)
Assembly of First Nations (1)
Bloc Québécois (1)
Canada First (1)
Canada West Foundation (1)

Sites for this collection are listed below. Narrow your resu
specific URL or to search the text of archived webpages.

Enter search terms here

Page 1 of 1 (54

Sort By: Title (A-Z) | Title (Z-A) | URL (A-Z) | UR

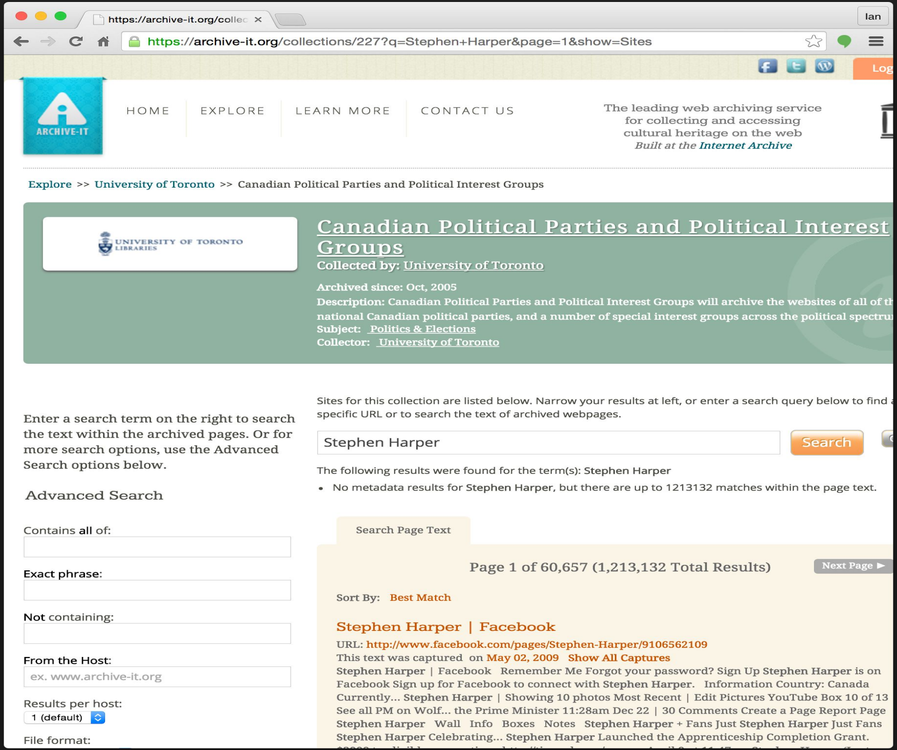
Title: Cosmopolitan Party of Canada
URL: http://agoracosmopolite.com/
Captured 59 times between Oct 4, 2005 and Nov 3, 20
Subject: Cosmopolitan Party of Canada

Title: Bloc Québécois

WALK

Current interfaces are **very limited** - simple search engine, some advanced options, no facets

Great collections, but nobody uses them.


https://archive-it.org/collec ×
Ian
https://archive-it.org/collections/227?q=Stephen+Harper&page=1&show=Sites
Log
ARCHIVE-IT
HOME
EXPLORE
LEARN MORE
CONTACT US
UNIVERSITY OF TORONTO LIBRARIES
Search
Search Page Text
Next Page ▶

The leading web archiving service
for collecting and accessing
cultural heritage on the web
Built at the Internet Archive

Explore >> University of Toronto >> Canadian Political Parties and Political Interest Groups

Canadian Political Parties and Political Interest Groups
Collected by: University of Toronto

Archived since: Oct, 2005
Description: Canadian Political Parties and Political Interest Groups will archive the websites of all of th
national Canadian political parties, and a number of special interest groups across the political spectrur
Subject: Politics & Elections
Collector: University of Toronto

Enter a search term on the right to search the text within the archived pages. Or for more search options, use the Advanced Search options below.

Advanced Search

Contains **all** of:

Exact phrase:

Not containing:

From the Host: ex. www.archive-it.org

Results per host: 1 (default)

File format:

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a
specific URL or to search the text of archived webpages.

Stephen Harper

The following results were found for the term(s): Stephen Harper

- No metadata results for Stephen Harper, but there are up to 1213132 matches within the page text.

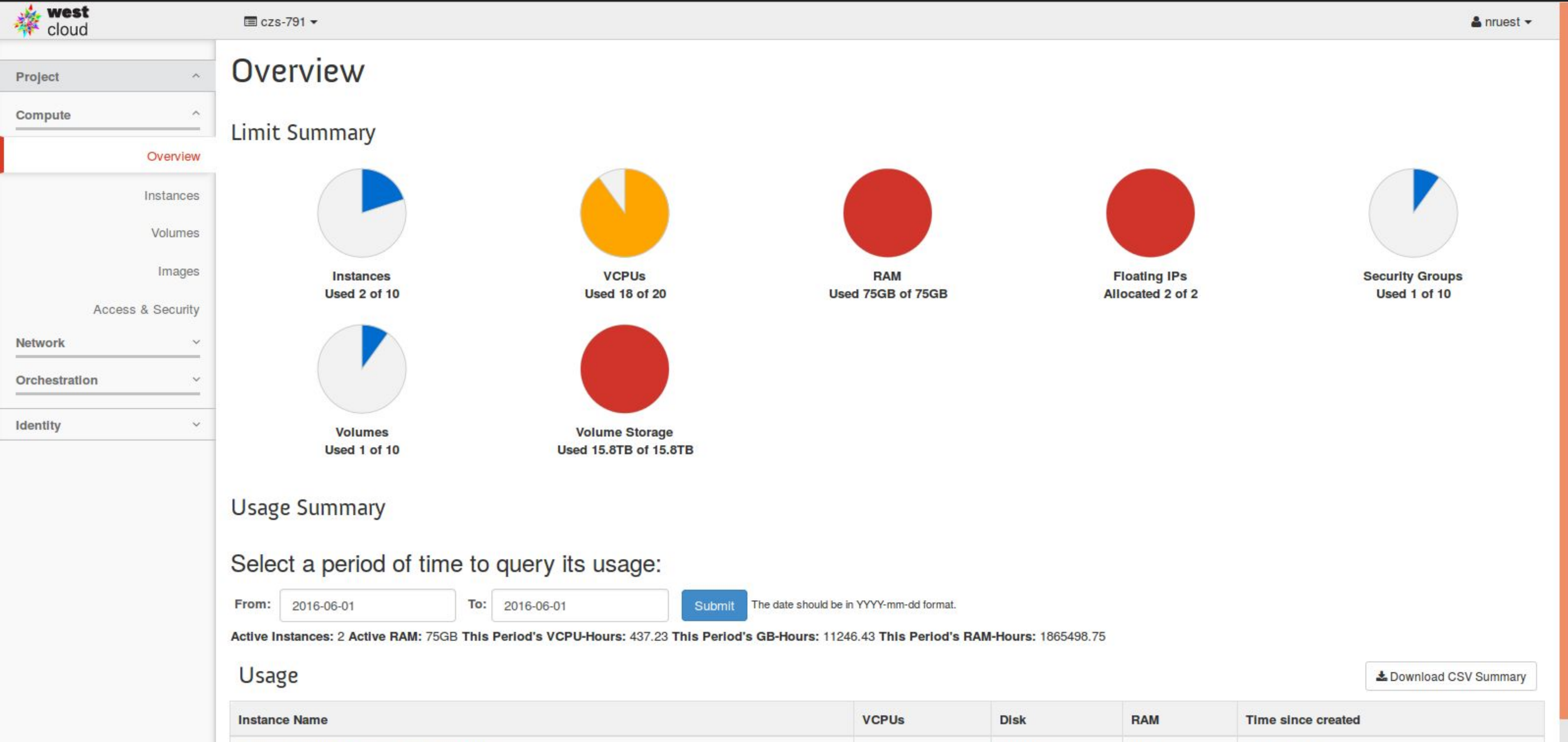
Page 1 of 60,657 (1,213,132 Total Results)

Sort By: Best Match

Stephen Harper | Facebook
URL: http://www.facebook.com/pages/Stephen-Harper/9106562109
This text was captured on May 02, 2009 Show All Captures
Stephen Harper | Facebook Remember Me Forgot your password? Sign Up Stephen Harper is on Facebook Sign up for Facebook to connect with Stephen Harper. Information Country: Canada Currently... Stephen Harper | Showing 10 photos Most Recent | Edit Pictures YouTube Box 10 of 13 See all PM on Wolf... the Prime Minister 11:28am Dec 22 | 30 Comments Create a Page Report Page Stephen Harper Wall Info Boxes Notes Stephen Harper + Fans Just Stephen Harper Just Fans Stephen Harper Celebrating... Stephen Harper Launched the Apprenticeship Completion Grant.

How could we build better access?

Preparing our dataset!



Overview

Limit Summary

Instances
Used 2 of 10

VCPUs
Used 18 of 20

RAM
Used 75GB of 75GB

Floating IPs
Allocated 2 of 2

Security Groups
Used 1 of 10

Volumes
Used 1 of 10

Volume Storage
Used 15.8TB of 15.8TB

Usage Summary

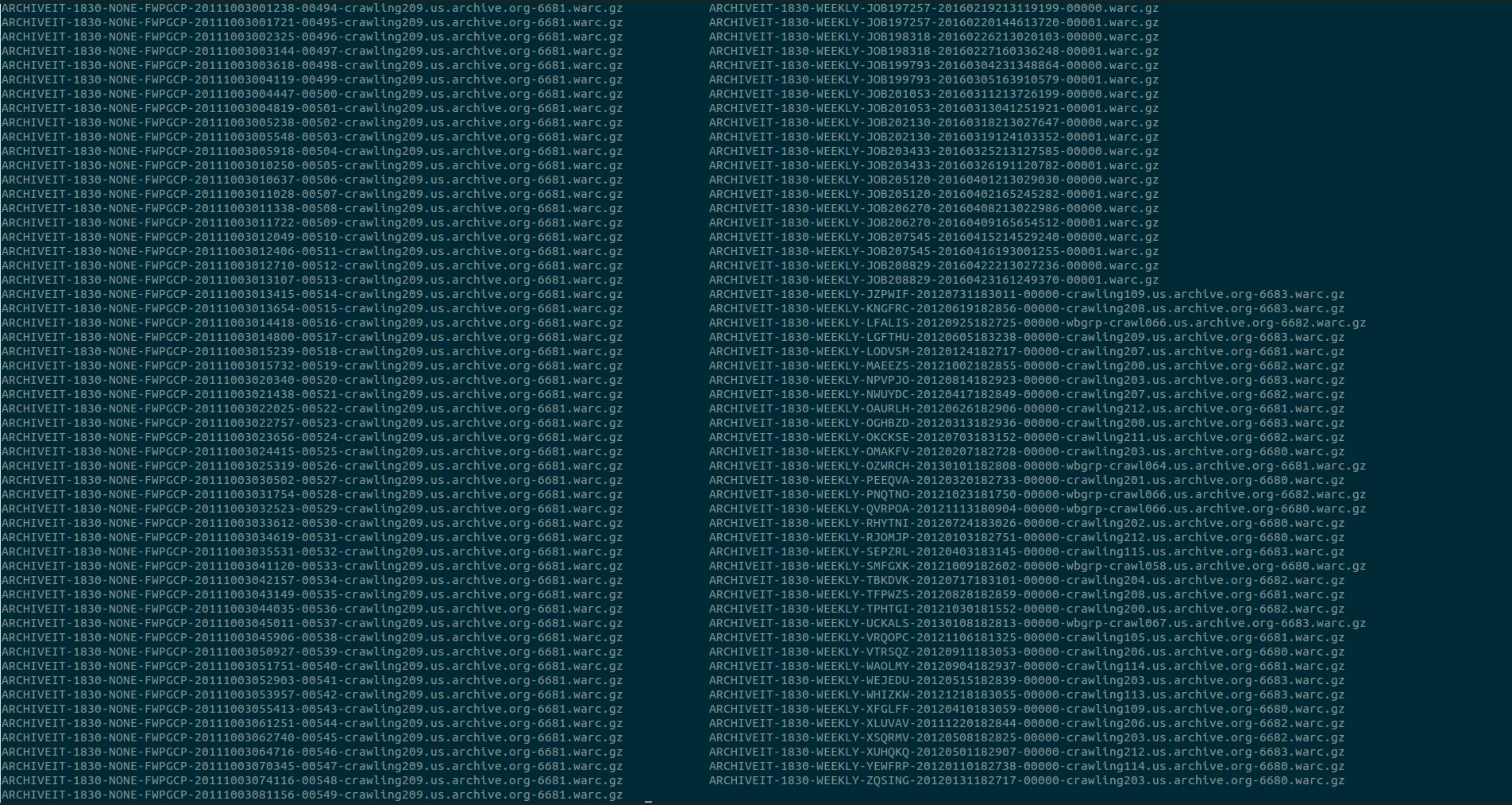
Select a period of time to query its usage:

From: 2016-06-01 **To:** 2016-06-01 Submit The date should be in YYYY-mm-dd format.

Active Instances: 2 **Active RAM:** 75GB **This Period's VCPU-Hours:** 437.23 **This Period's GB-Hours:** 11246.43 **This Period's RAM-Hours:** 1865498.75

Usage

Instance Name	VCPUs	Disk	RAM	Time since created



```
ARCHIVEIT-1830-NONE-FWPGCP-20111003001238-00494-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB197257-20160219213119199-00000.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003001721-00495-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB197257-20160220144613720-00001.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003002325-00496-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB198318-20160226213020103-00000.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003003144-00497-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB198318-20160227160336248-00001.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003003618-00498-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB199793-20160304231348864-00000.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003004119-00499-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB199793-20160305163910579-00001.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003004447-00500-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB201053-20160311213726199-00000.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003004819-00501-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB201053-20160313041251921-00001.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003005238-00502-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB202130-20160318213027647-00000.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003005548-00503-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB202130-20160319124103352-00001.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003005918-00504-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB203433-20160325213127585-00000.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003010250-00505-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB203433-20160326191120782-00001.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003010637-00506-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB205120-20160401213029030-00000.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003011028-00507-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB205120-20160402165245282-00001.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003011338-00508-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB206270-20160408213022986-00000.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003011722-00509-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB206270-20160409165654512-00001.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003012049-00510-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB207545-20160415214529240-00000.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003012406-00511-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB207545-20160416193001255-00001.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003012710-00512-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB208829-20160422213027236-00000.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003013107-00513-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JOB208829-20160423161249370-00001.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003013415-00514-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-JZPWIF-20120731183011-00000-crawling109.us.archive.org-6683.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003013654-00515-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-KNGFRC-20120619182856-00000-crawling208.us.archive.org-6683.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003014418-00516-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-LFALIS-20120925182725-00000-wbgrp-crawl066.us.archive.org-6682.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003014800-00517-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-LGFTHU-20120605183238-00000-crawling209.us.archive.org-6683.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003015239-00518-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-LODVSM-20120124182717-00000-crawling207.us.archive.org-6681.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003015732-00519-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-MAEEZS-20121002182855-00000-crawling200.us.archive.org-6682.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003020340-00520-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-NPVPJO-20120814182923-00000-crawling203.us.archive.org-6683.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003021438-00521-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-NWUYDC-20120417182849-00000-crawling207.us.archive.org-6682.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003022025-00522-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-OAURLH-20120626182906-00000-crawling212.us.archive.org-6681.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003022757-00523-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-OGHBZD-20120313182936-00000-crawling200.us.archive.org-6683.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003023656-00524-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-OKCKSE-20120703183152-00000-crawling211.us.archive.org-6682.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003024415-00525-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-OMAKFV-20120207182728-00000-crawling203.us.archive.org-6680.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003025319-00526-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-OZWRCH-20130101182808-00000-wbgrp-crawl064.us.archive.org-6681.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003030502-00527-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-PEEQVA-20120320182733-00000-crawling201.us.archive.org-6680.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003031754-00528-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-PNQTNO-20121023181750-00000-wbgrp-crawl066.us.archive.org-6682.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003032523-00529-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-QVRPOA-20121113180904-00000-wbgrp-crawl066.us.archive.org-6680.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003033612-00530-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-RHYTNI-20120724183026-00000-crawling202.us.archive.org-6680.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003034619-00531-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-RJOMJP-20120103182751-00000-crawling212.us.archive.org-6680.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003035531-00532-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-SEPZRL-20120403183145-00000-crawling115.us.archive.org-6683.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003041120-00533-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-SMFGXK-20121009182602-00000-wbgrp-crawl058.us.archive.org-6680.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003042157-00534-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-TBKDVK-20120717183101-00000-crawling204.us.archive.org-6682.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003043149-00535-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-TFPWZS-20120828182859-00000-crawling208.us.archive.org-6681.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003044035-00536-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-TPHTGI-20121030181552-00000-crawling200.us.archive.org-6682.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003045011-00537-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-UCKALS-20130108182813-00000-wbgrp-crawl067.us.archive.org-6683.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003045906-00538-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-VRQOPC-20121106181325-00000-crawling105.us.archive.org-6681.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003050927-00539-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-VTRSQZ-20120911183053-00000-crawling206.us.archive.org-6680.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003051751-00540-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-WAOLMY-20120904182937-00000-crawling114.us.archive.org-6681.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003052903-00541-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-WEJEDU-20120515182839-00000-crawling203.us.archive.org-6683.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003053957-00542-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-WHIZKW-20121218183055-00000-crawling113.us.archive.org-6683.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003055413-00543-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-XFGLFF-20120410183059-00000-crawling109.us.archive.org-6680.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003061251-00544-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-XLUVAV-20111220182844-00000-crawling206.us.archive.org-6682.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003062740-00545-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-XSQRMV-20120508182825-00000-crawling203.us.archive.org-6682.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003064716-00546-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-XUHQKQ-20120501182907-00000-crawling212.us.archive.org-6683.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003070345-00547-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-YEWFRP-20120110182738-00000-crawling114.us.archive.org-6680.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003074116-00548-crawling209.us.archive.org-6681.warc.gz            ARCHIVEIT-1830-WEEKLY-ZQSING-20120131182717-00000-crawling203.us.archive.org-6680.warc.gz
ARCHIVEIT-1830-NONE-FWPGCP-20111003081156-00549-crawling209.us.archive.org-6681.warc.gz
```

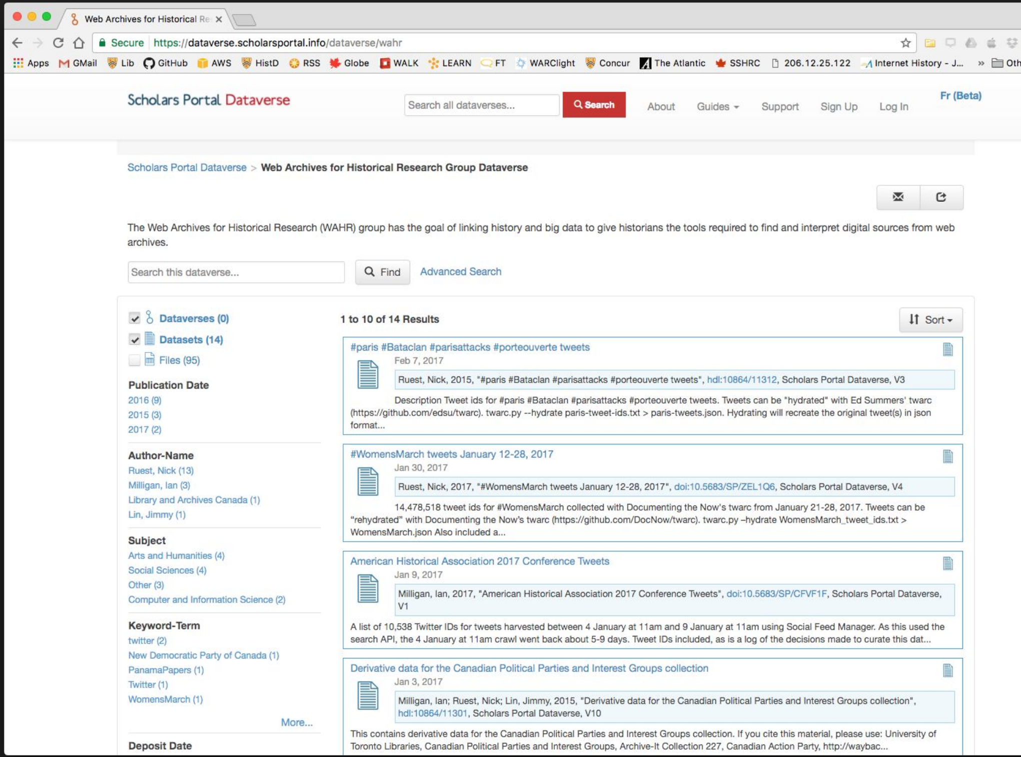
Public Access

Providing Access to Derivative Datasets

Want to do network diagrams like Ian showed? All collections have data.

List of URLs, domains

(As well as social media data, etc.)



Scholars Portal Dataverse > **Web Archives for Historical Research Group Dataverse**

The Web Archives for Historical Research (WAHR) group has the goal of linking history and big data to give historians the tools required to find and interpret digital sources from web archives.

☑ Dataverses (0)

☑ Datasets (14)

□ Files (95)

Publication Date

2016 (9)

2015 (3)

2017 (2)

Author-Name

Ruest, Nick (13)

Milligan, Ian (3)

Library and Archives Canada (1)

Lin, Jimmy (1)

Subject

Arts and Humanities (4)

Social Sciences (4)

Other (3)

Computer and Information Science (2)

Keyword-Term

twitter (2)

New Democratic Party of Canada (1)

PanamaPapers (1)

Twitter (1)

WomensMarch (1)

More...

Deposit Date

1 to 10 of 14 Results

#paris #Bataclan #parisattacks #porteouverte tweets

Feb 7, 2017

Ruest, Nick, 2015, "#paris #Bataclan #parisattacks #porteouverte tweets", hdl:10864/11312, Scholars Portal Dataverse, V3

Description Tweet ids for #paris #Bataclan #parisattacks #porteouverte tweets. Tweets can be "hydrated" with Ed Summers' twarc (https://github.com/edsu/twarc). twarc.py --hydrate paris-tweet-ids.txt > paris-tweets.json. Hydrating will recreate the original tweet(s) in json format...

#WomensMarch tweets January 12-28, 2017

Jan 30, 2017

Ruest, Nick, 2017, "#WomensMarch tweets January 12-28, 2017", doi:10.5683/SP/ZEL1Q6, Scholars Portal Dataverse, V4

14,478,518 tweet ids for #WomensMarch collected with Documenting the Now's twarc from January 21-28, 2017. Tweets can be "rehydrated" with Documenting the Now's twarc (https://github.com/DocNow/twarc). twarc.py –hydrate WomensMarch_tweet_ids.txt > WomensMarch.json Also included a...

American Historical Association 2017 Conference Tweets

Jan 9, 2017

Milligan, Ian, 2017, "American Historical Association 2017 Conference Tweets", doi:10.5683/SP/CFVF1F, Scholars Portal Dataverse, V1

A list of 10,538 Twitter IDs for tweets harvested between 4 January at 11am and 9 January at 11am using Social Feed Manager. As this used the search API, the 4 January at 11am crawl went back about 5-9 days. Tweet IDs included, as is a log of the decisions made to curate this dat...

Derivative data for the Canadian Political Parties and Interest Groups collection

Jan 3, 2017

Milligan, Ian; Ruest, Nick; Lin, Jimmy, 2015, "Derivative data for the Canadian Political Parties and Interest Groups collection", hdl:10864/11301, Scholars Portal Dataverse, V10

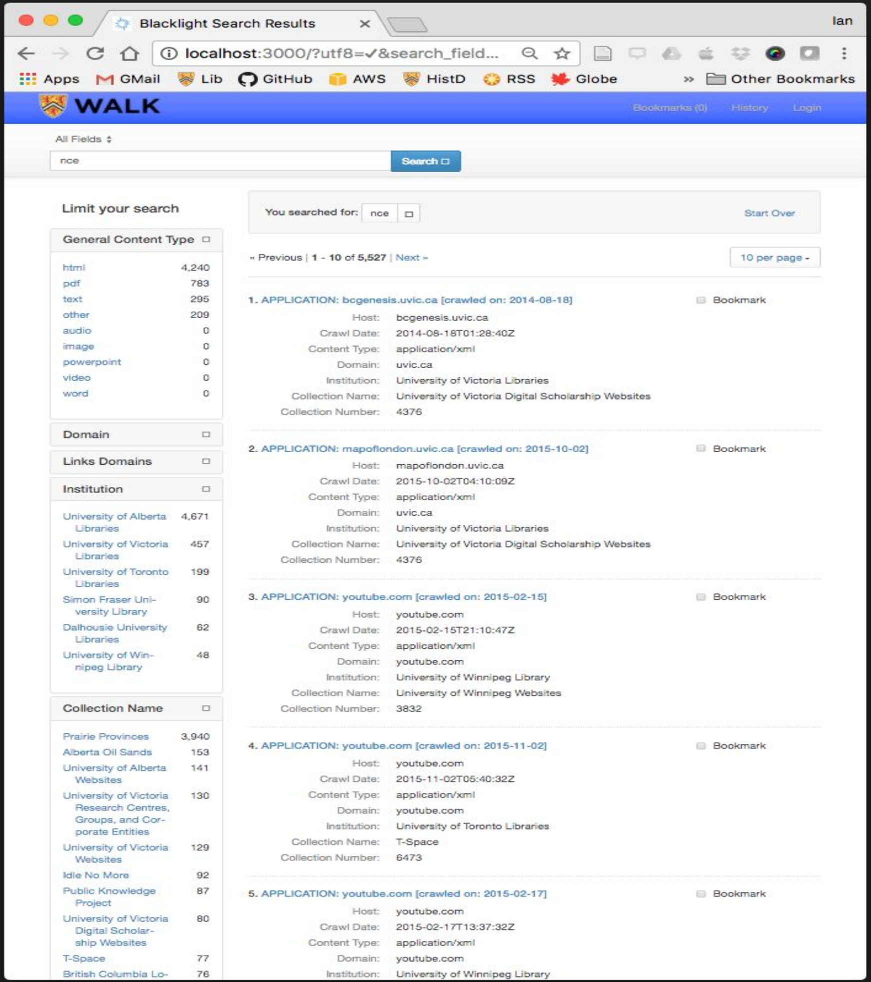
This contains derivative data for the Canadian Political Parties and Interest Groups collection. If you cite this material, please use: University of Toronto Libraries, Canadian Political Parties and Interest Groups, Archive-It Collection 227, Canadian Action Party, http://waybac...

~~Shine~~

Blacklight

WebArchives.ca

- Warcbase with a GUI front end
- Archivists/Librarians bring their collections, generate derivatives
- Hosted with Compute Canada



nce

Limit your search

General Content Type □

html	4,240
pdf	783
text	295
other	209
audio	0
image	0
powerpoint	0
video	0
word	0

Domain □

Links Domains □

Institution □

University of Alberta Libraries	4,671
University of Victoria Libraries	457
University of Toronto Libraries	199
Simon Fraser University Library	90
Dalhousie University Libraries	62
University of Winnipeg Library	48

Collection Name □

Prairie Provinces	3,940
Alberta Oil Sands	153
University of Alberta Websites	141
University of Victoria Research Centres, Groups, and Corporate Entities	130
University of Victoria Websites	129
Idle No More	92
Public Knowledge Project	87
University of Victoria Digital Scholarship Websites	80
T-Space	77
British Columbia Lo-	76

You searched for: nce □ Start Over

« Previous | 1 - 10 of 5,527 | Next » 10 per page -

1. APPLICATION: bcgenesis.uvic.ca [crawled on: 2014-08-18] □ Bookmark

Host: bcgenesis.uvic.ca
Crawl Date: 2014-08-18T01:28:40Z
Content Type: application/xml
Domain: uvic.ca
Institution: University of Victoria Libraries
Collection Name: University of Victoria Digital Scholarship Websites
Collection Number: 4376

2. APPLICATION: mapoflondon.uvic.ca [crawled on: 2015-10-02] □ Bookmark

Host: mapoflondon.uvic.ca
Crawl Date: 2015-10-02T04:10:09Z
Content Type: application/xml
Domain: uvic.ca
Institution: University of Victoria Libraries
Collection Name: University of Victoria Digital Scholarship Websites
Collection Number: 4376

3. APPLICATION: youtube.com [crawled on: 2015-02-15] □ Bookmark

Host: youtube.com
Crawl Date: 2015-02-15T21:10:47Z
Content Type: application/xml
Domain: youtube.com
Institution: University of Winnipeg Library
Collection Name: University of Winnipeg Websites
Collection Number: 3832

4. APPLICATION: youtube.com [crawled on: 2015-11-02] □ Bookmark

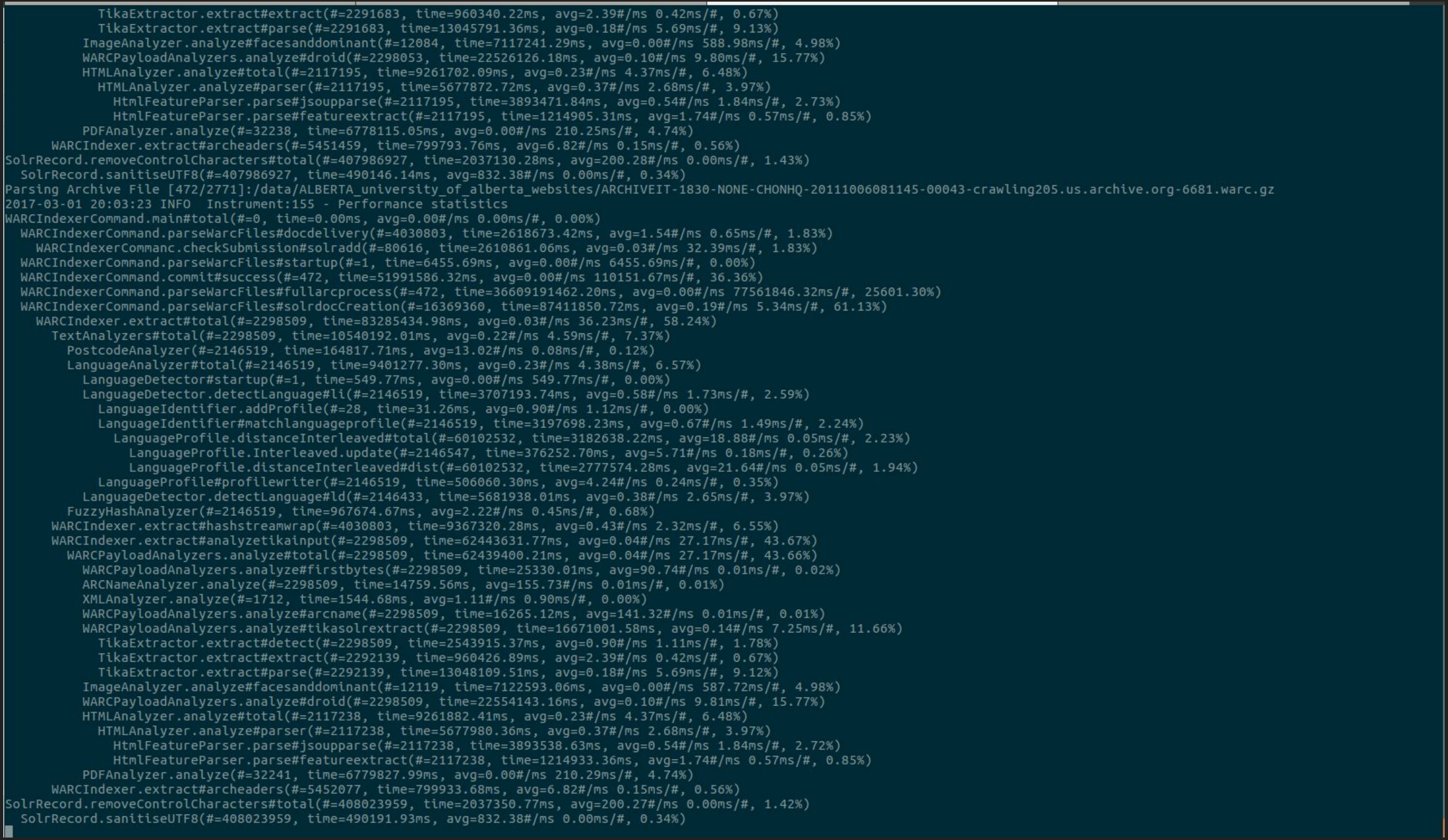
Host: youtube.com
Crawl Date: 2015-11-02T05:40:32Z
Content Type: application/xml
Domain: youtube.com
Institution: University of Toronto Libraries
Collection Name: T-Space
Collection Number: 6473

5. APPLICATION: youtube.com [crawled on: 2015-02-17] □ Bookmark

Host: youtube.com
Crawl Date: 2015-02-17T13:37:32Z
Content Type: application/xml
Domain: youtube.com
Institution: University of Winnipeg Library

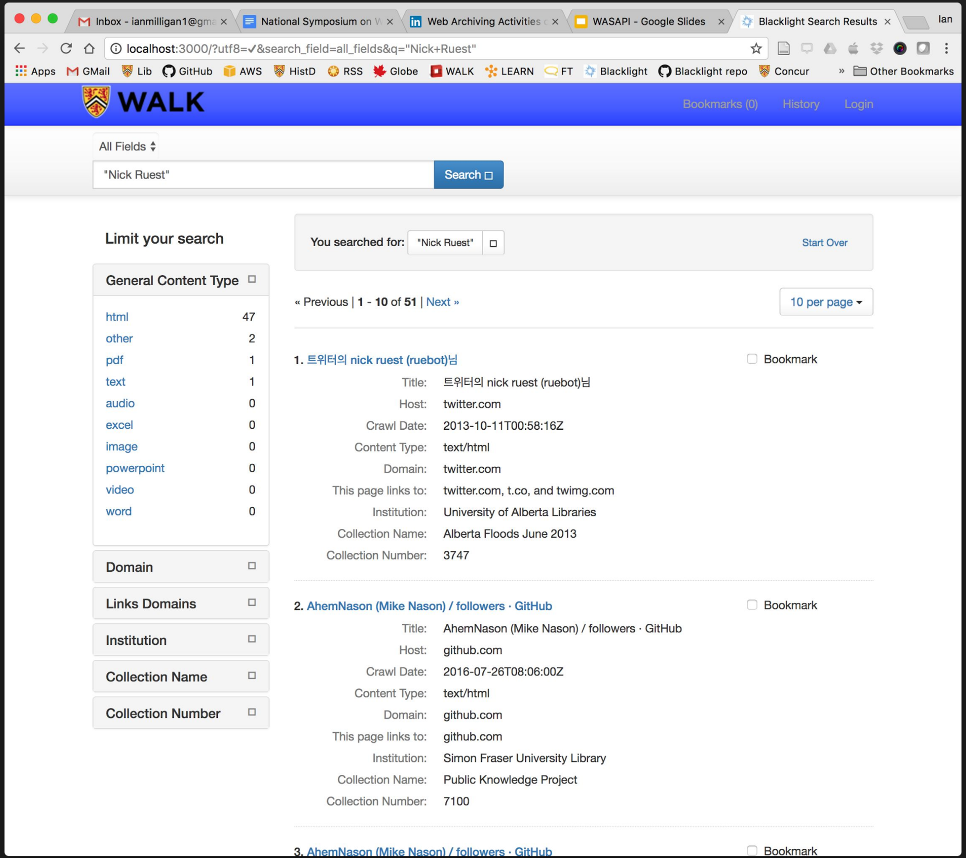
~~14 Million Solr docs!~~

100+ Million Solr docs!



```
      TikaExtractor.extract#extract(#=2291683, time=960340.22ms, avg=2.39#/ms 0.42ms/#, 0.67%)
      TikaExtractor.extract#parse(#=2291683, time=13045791.36ms, avg=0.18#/ms 5.69ms/#, 9.13%)
     ImageAnalyzer.analyze#facesanddominant(#=12084, time=7117241.29ms, avg=0.00#/ms 588.98ms/#, 4.98%)
     WARCPayloadAnalyzers.analyze#droid(#=2298053, time=22526126.18ms, avg=0.10#/ms 9.80ms/#, 15.77%)
     HTMLAnalyzer.analyze#total(#=2117195, time=9261702.09ms, avg=0.23#/ms 4.37ms/#, 6.48%)
      HTMLAnalyzer.analyze#parser(#=2117195, time=5677872.72ms, avg=0.37#/ms 2.68ms/#, 3.97%)
       HtmlFeatureParser.parse#jsoupparse(#=2117195, time=3893471.84ms, avg=0.54#/ms 1.84ms/#, 2.73%)
       HtmlFeatureParser.parse#featureextract(#=2117195, time=1214905.31ms, avg=1.74#/ms 0.57ms/#, 0.85%)
     PDFAnalyzer.analyze(#=32238, time=6778115.05ms, avg=0.00#/ms 210.25ms/#, 4.74%)
   WARCIndexer.extract#archeaders(#=5451459, time=799793.76ms, avg=6.82#/ms 0.15ms/#, 0.56%)
SolrRecord.removeControlCharacters#total(#=407986927, time=2037130.28ms, avg=200.28#/ms 0.00ms/#, 1.43%)
 SolrRecord.sanitiseUTF8(#=407986927, time=490146.14ms, avg=832.38#/ms 0.00ms/#, 0.34%)
Parsing Archive File [472/2771]:/data/ALBERTA_university_of_alberta_websites/ARCHIVEIT-1830-NONE-CHONHQ-20111006081145-00043-crawling205.us.archive.org-6681.warc.gz
2017-03-01 20:03:23 INFO  Instrument:155 - Performance statistics
WARCIndexerCommand.main#total(#=0, time=0.00ms, avg=0.00#/ms 0.00ms/#, 0.00%)
 WARCIndexerCommand.parseWarcFiles#docdelivery(#=4030803, time=2618673.42ms, avg=1.54#/ms 0.65ms/#, 1.83%)
  WARCIndexerCommanc.checkSubmission#solradd(#=80616, time=2610861.06ms, avg=0.03#/ms 32.39ms/#, 1.83%)
 WARCIndexerCommand.parseWarcFiles#startup(#=1, time=6455.69ms, avg=0.00#/ms 6455.69ms/#, 0.00%)
 WARCIndexerCommand.commit#success(#=472, time=51991586.32ms, avg=0.00#/ms 110151.67ms/#, 36.36%)
 WARCIndexerCommand.parseWarcFiles#fullarcprocess(#=472, time=36609191462.20ms, avg=0.00#/ms 77561846.32ms/#, 25601.30%)
 WARCIndexerCommand.parseWarcFiles#solrdocCreation(#=16369360, time=87411850.72ms, avg=0.19#/ms 5.34ms/#, 61.13%)
  WARCIndexer.extract#total(#=2298509, time=83285434.98ms, avg=0.03#/ms 36.23ms/#, 58.24%)
   TextAnalyzers#total(#=2298509, time=10540192.01ms, avg=0.22#/ms 4.59ms/#, 7.37%)
    PostcodeAnalyzer(#=2146519, time=164817.71ms, avg=13.02#/ms 0.08ms/#, 0.12%)
    LanguageAnalyzer#total(#=2146519, time=9401277.30ms, avg=0.23#/ms 4.38ms/#, 6.57%)
     LanguageDetector#startup(#=1, time=549.77ms, avg=0.00#/ms 549.77ms/#, 0.00%)
     LanguageDetector.detectLanguage#li(#=2146519, time=3707193.74ms, avg=0.58#/ms 1.73ms/#, 2.59%)
      LanguageIdentifier.addProfile(#=28, time=31.26ms, avg=0.90#/ms 1.12ms/#, 0.00%)
      LanguageIdentifier#matchlanguageprofile(#=2146519, time=3197698.23ms, avg=0.67#/ms 1.49ms/#, 2.24%)
       LanguageProfile.distanceInterleaved#total(#=60102532, time=3182638.22ms, avg=18.88#/ms 0.05ms/#, 2.23%)
        LanguageProfile.Interleaved.update(#=2146547, time=376252.70ms, avg=5.71#/ms 0.18ms/#, 0.26%)
        LanguageProfile.distanceInterleaved#dist(#=60102532, time=2777574.28ms, avg=21.64#/ms 0.05ms/#, 1.94%)
      LanguageProfile#profilewriter(#=2146519, time=506060.30ms, avg=4.24#/ms 0.24ms/#, 0.35%)
     LanguageDetector.detectLanguage#ld(#=2146433, time=5681938.01ms, avg=0.38#/ms 2.65ms/#, 3.97%)
    FuzzyHashAnalyzer(#=2146519, time=967674.67ms, avg=2.22#/ms 0.45ms/#, 0.68%)
   WARCIndexer.extract#hashstreamwrap(#=4030803, time=9367320.28ms, avg=0.43#/ms 2.32ms/#, 6.55%)
   WARCIndexer.extract#analyzetikainput(#=2298509, time=62443631.77ms, avg=0.04#/ms 27.17ms/#, 43.67%)
    WARCPayloadAnalyzers.analyze#total(#=2298509, time=62439400.21ms, avg=0.04#/ms 27.17ms/#, 43.66%)
     WARCPayloadAnalyzers.analyze#firstbytes(#=2298509, time=25330.01ms, avg=90.74#/ms 0.01ms/#, 0.02%)
     ARCNameAnalyzer.analyze(#=2298509, time=14759.56ms, avg=155.73#/ms 0.01ms/#, 0.01%)
     XMLAnalyzer.analyze(#=1712, time=1544.68ms, avg=1.11#/ms 0.90ms/#, 0.00%)
     WARCPayloadAnalyzers.analyze#arcname(#=2298509, time=16265.12ms, avg=141.32#/ms 0.01ms/#, 0.01%)
     WARCPayloadAnalyzers.analyze#tikasolrextract(#=2298509, time=16671001.58ms, avg=0.14#/ms 7.25ms/#, 11.66%)
      TikaExtractor.extract#detect(#=2298509, time=2543915.37ms, avg=0.90#/ms 1.11ms/#, 1.78%)
      TikaExtractor.extract#extract(#=2292139, time=960426.89ms, avg=2.39#/ms 0.42ms/#, 0.67%)
      TikaExtractor.extract#parse(#=2292139, time=13048109.51ms, avg=0.18#/ms 5.69ms/#, 9.12%)
     ImageAnalyzer.analyze#facesanddominant(#=12119, time=7122593.06ms, avg=0.00#/ms 587.72ms/#, 4.98%)
     WARCPayloadAnalyzers.analyze#droid(#=2298509, time=22554143.16ms, avg=0.10#/ms 9.81ms/#, 15.77%)
     HTMLAnalyzer.analyze#total(#=2117238, time=9261882.41ms, avg=0.23#/ms 4.37ms/#, 6.48%)
      HTMLAnalyzer.analyze#parser(#=2117238, time=5677980.36ms, avg=0.37#/ms 2.68ms/#, 3.97%)
       HtmlFeatureParser.parse#jsoupparse(#=2117238, time=3893538.63ms, avg=0.54#/ms 1.84ms/#, 2.72%)
       HtmlFeatureParser.parse#featureextract(#=2117238, time=1214933.36ms, avg=1.74#/ms 0.57ms/#, 0.85%)
     PDFAnalyzer.analyze(#=32241, time=6779827.99ms, avg=0.00#/ms 210.29ms/#, 4.74%)
   WARCIndexer.extract#archeaders(#=5452077, time=799933.68ms, avg=6.82#/ms 0.15ms/#, 0.56%)
SolrRecord.removeControlCharacters#total(#=408023959, time=2037350.77ms, avg=200.27#/ms 0.00ms/#, 1.42%)
 SolrRecord.sanitiseUTF8(#=408023959, time=490191.93ms, avg=832.38#/ms 0.00ms/#, 0.34%)
```

WARClight prototype



localhost:3000/?utf8=✓&search_field=all_fields&q="Nick+Ruest"

"Nick Ruest"

Limit your search

You searched for: "Nick Ruest"

« Previous | 1 - 10 of 51 | Next »

General Content Type

html	47
other	2
pdf	1
text	1
audio	0
excel	0
image	0
powerpoint	0
video	0
word	0

Domain

Links Domains

Institution

Collection Name

Collection Number

1. 트위터의 nick ruest (ruebot)님

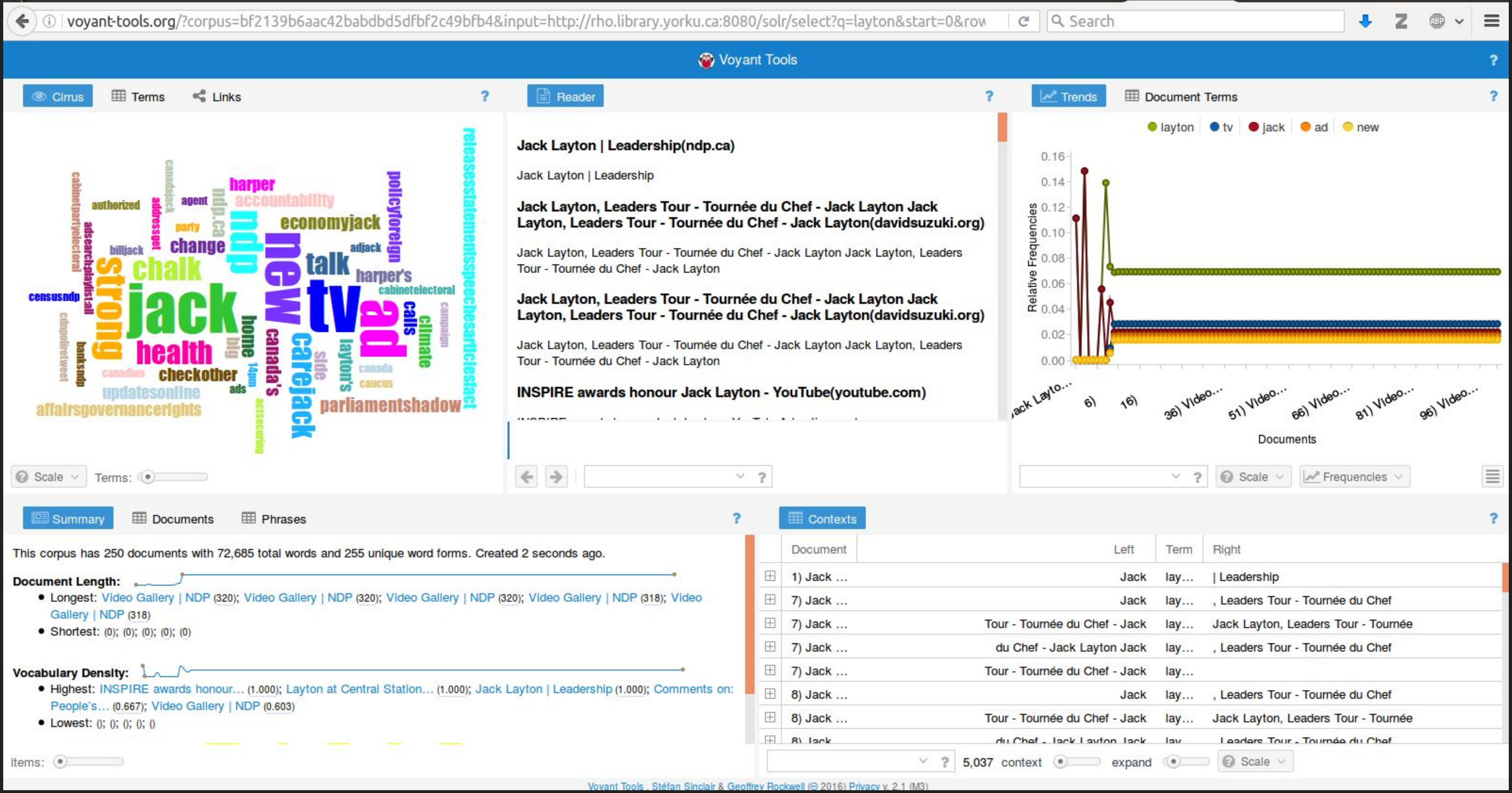
Title:	트위터의 nick ruest (ruebot)님
Host:	twitter.com
Crawl Date:	2013-10-11T00:58:16Z
Content Type:	text/html
Domain:	twitter.com
This page links to:	twitter.com, t.co, and twimg.com
Institution:	University of Alberta Libraries
Collection Name:	Alberta Floods June 2013
Collection Number:	3747

2. AhemNason (Mike Nason) / followers · GitHub

Title:	AhemNason (Mike Nason) / followers · GitHub
Host:	github.com
Crawl Date:	2016-07-26T08:06:00Z
Content Type:	text/html
Domain:	github.com
This page links to:	github.com
Institution:	Simon Fraser University Library
Collection Name:	Public Knowledge Project
Collection Number:	7100

3. AhemNason (Mike Nason) / followers · GitHub

Finding famous people in WARClight, by Ian Milligan



Jack Layton | Leadership(ndp.ca)

Jack Layton | Leadership

Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton(davidsuzuki.org)

Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton

Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton(davidsuzuki.org)

Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton Jack Layton, Leaders Tour - Tournée du Chef - Jack Layton

INSPIRE awards honour Jack Layton - YouTube(youtube.com)

This corpus has 250 documents with 72,685 total words and 255 unique word forms. Created 2 seconds ago.

Document Length:

- Longest: Video Gallery | NDP (320); Video Gallery | NDP (320); Video Gallery | NDP (320); Video Gallery | NDP (318); Video Gallery | NDP (318)
- Shortest: (0); (0); (0); (0); (0)

Vocabulary Density:

- Highest: INSPIRE awards honour… (1.000); Layton at Central Station… (1.000); Jack Layton | Leadership (1.000); Comments on: People's… (0.667); Video Gallery | NDP (0.603)
- Lowest: (); (); (); (); ()

Document	Left	Term	Right
1) Jack …	Jack	lay…	| Leadership
7) Jack …	Jack	lay…	, Leaders Tour - Tournée du Chef
7) Jack …	Tour - Tournée du Chef - Jack	lay…	Jack Layton, Leaders Tour - Tournée
7) Jack …	du Chef - Jack Layton Jack	lay…	, Leaders Tour - Tournée du Chef
7) Jack …	Tour - Tournée du Chef - Jack	lay…	
8) Jack …	Jack	lay…	, Leaders Tour - Tournée du Chef
8) Jack …	Tour - Tournée du Chef - Jack	lay…	Jack Layton, Leaders Tour - Tournée
8) Jack …	du Chef - Jack Layton Jack	lay…	, Leaders Tour - Tournée du Chef

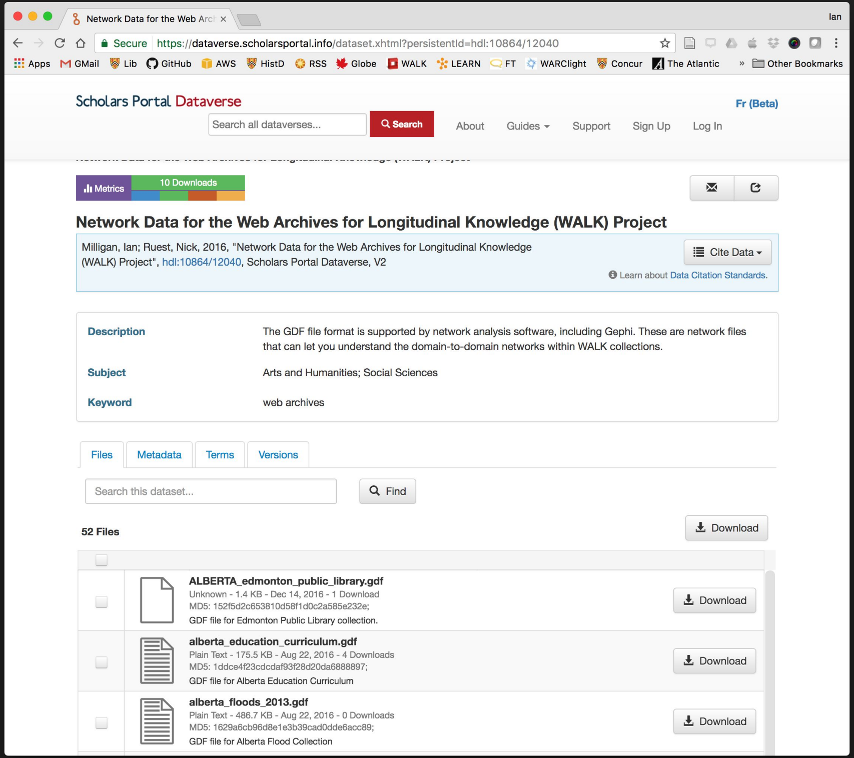
Voyant Tools , Stéfan Sinclair & Geoffrey Rockwell (© 2016) Privacy v. 2.1 (M3)

warcbase

Again :-)

WALK

- Mass Ingest Script to take each collection and:
  - Extract hyperlinks and generate gephi files;
  - Extract all URLs and domain counts;
  - Extract plain text;



Network Data for the Web Arc
Secure https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=hdl:10864/12040
Apps GMail Lib GitHub AWS HistD RSS Globe WALK LEARN FT WARClight Concur The Atlantic Other Bookmarks
Scholars Portal Dataverse
Search all dataverses...
Search
About
Guides
Support
Sign Up
Log In
Fr (Beta)
Metrics
10 Downloads
Cite Data
Files
Metadata
Terms
Versions
Search this dataset...
Find
Download

Network Data for the Web Archives for Longitudinal Knowledge (WALK) Project

Milligan, Ian; Ruest, Nick, 2016, "Network Data for the Web Archives for Longitudinal Knowledge (WALK) Project", hdl:10864/12040, Scholars Portal Dataverse, V2

Learn about Data Citation Standards.

Description	The GDF file format is supported by network analysis software, including Gephi. These are network files that can let you understand the domain-to-domain networks within WALK collections.
Subject	Arts and Humanities; Social Sciences
Keyword	web archives

52 Files

ALBERTA_edmonton_public_library.gdf
Unknown - 1.4 KB - Dec 14, 2016 - 1 Download
MD5: 152f5d2c653810d58f1d0c2a585e232e;
GDF file for Edmonton Public Library collection.

alberta_education_curriculum.gdf
Plain Text - 175.5 KB - Aug 22, 2016 - 4 Downloads
MD5: 1ddce4f23cdcdaf93f28d20da6888897;
GDF file for Alberta Education Curriculum

alberta_floods_2013.gdf
Plain Text - 486.7 KB - Aug 22, 2016 - 0 Downloads
MD5: 1629a6cb96d8e1e3b39cad0dde6acc89;
GDF file for Alberta Flood Collection

Links!

- https://uwaterloo.ca/web-archive-group/
- https://github.com/web-archive-group/
- https://github.com/ianmilligan1/
- https://github.com/ruebot
- http://dataverse.scholarsportal.info/dvn/dv/wahr

Contact

Nick Ruest: @ruebot

ruestn@yorku.ca

Ian Milligan: @ianmilligan1

i2milligan@uwaterloo.ca

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Ontario

compute canada | calcul canada

UNIVERSITY OF WATERLOO

YORK
UNIVERSITÉ
UNIVERSITY
U

Thanks very much!