

# Content Selection and Curation for Web Archiving

The Gatekeepers vs the Masses

UNIVERSITY OF  
**WATERLOO**

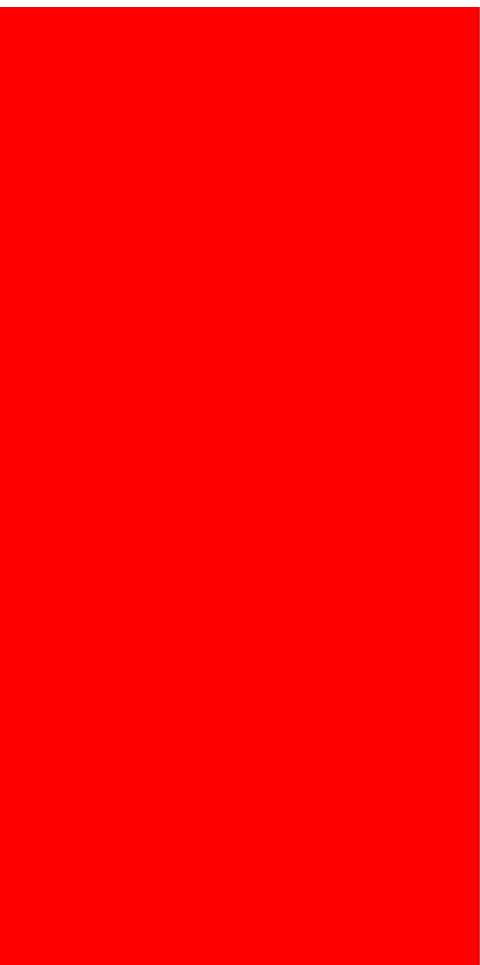


Ian Milligan (@ianmilligan1)

Nick Ruest (@ruebot)

Jimmy Lin (@lintool)





A tale of three collections

...or...

# Gatekeepers vs. the Masses

# 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

The screenshot shows the Liberal Party of Canada website. At the top is a navigation bar with links: HOME, THE TEAM, THE PARTY, ISSUES, MEDIA CENTRE, YOUR RIDING, and DONATE. Below this is a large banner image of Prime Minister Paul Martin and Adrienne Clarkson, with the text "ADDRESS BY PRIME MINISTER PAUL MARTIN". Under the banner are three smaller images with captions: "Volunteer", "Join Today!", and "Donate Now". The main content area has a heading "Your Excellencies, Honourable Members, Ladies and Gentlemen:" followed by a paragraph of text. Below this is another paragraph of text. To the right of the main content is a red sidebar with the heading "Stay Informed" and a "GO" button. Below this is a "Top Stories" section with three items: "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", and "September 27, 2005 Address by Prime Minister Paul Martin at the installation of the new Governor General". Below the stories is a "Complete List of Stories" link. Further down is a "Commissions" section with four items: "Young Liberals of Canada", "National Women's Liberal Commission", "Aboriginal Peoples' Commission", and "Senior Liberals Commission". At the bottom of the page is a red footer bar with links: Home | News | Your Riding | Issues | Contact Us | français. Below the links is a small disclaimer: "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".

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

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](#)

**Stay Informed**

**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](#)

# Access

- **ArchiveIT** - simple search engine, some advanced options; no facets
- **wget/curl** - ~350GB

HOMEEXPLORELEARN MORECONTACT US

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:

Results per host:

File format:

Capture date range:  
From:  To:

[Help with Search](#)

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.

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.

Search 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](#)

# warcbase

[warcbase.org](https://warcbase.org) | [docs.warcbase.org](https://docs.warcbase.org)



#elxn42

# 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](#) or using the [Twitter advanced search query builder](#). 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](#)**

# The Streaming APIs

## Overview

The Streaming APIs give developers low latency access to Twitter's global stream of Tweet data. A proper implementation of 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 streaming endpoints, each customized to certain use cases.

<a href="#">Public streams</a>	Streams of the public data flowing through Twitter. Suitable for following specific users or topics, and data mining.
<a href="#">User streams</a>	Single-user streams, containing roughly all of the data corresponding with a single user's view of Twitter.
<a href="#">Site streams</a>	The multi-user version of user streams. Site streams are intended for servers which must connect to Twitter on behalf of many users.

## 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 application

# twarc

<https://github.com/edsu/twarc>

# twarc

build `passing` `gitter` `join chat` DOI `10.5281/zenodo.48797`

twarc is a command line tool and Python library for archiving Twitter JSON data. Each tweet is represented as a JSON object that is exactly what was returned from the Twitter API. Tweets are stored as line-oriented JSON. Twarc runs in three modes: search, filter stream and hydrate. When running in each mode twarc will stop and resume activity in order to work within the Twitter API's [rate limits](#).

## Install

1. install [Python](#) (2 or 3)
2. `pip install twarc`

## Twitter API Keys

Before using twarc you will need to register an application at [apps.twitter.com](https://apps.twitter.com). Once you've created your application, note down the consumer key, consumer secret and then click to generate an access token and access token secret. With these four variables in hand you are ready to start using twarc.

The first time you run twarc it will prompt you for these keys and store them in a `.twarc` file in your home directory. Sometimes it can be handy to store multiple authorization keys for different Twitter accounts in your config file. So if you can have multiple profiles to your `.twarc` file, for example:

```
[main]
consumer_key=lksdf1jklksdjf
consumer_secret=lkjsdf1kjksdlkfj
```



**3,918,932 tweets**

eh?

**318,176 unique URLs**

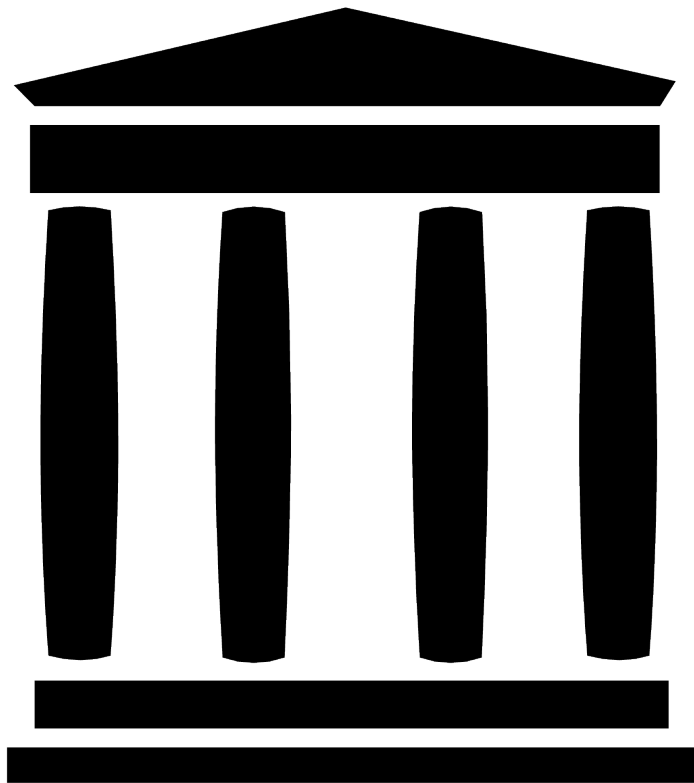
(1,988,693 URLs total)

# unshrtn

<https://github.com/edsu/unshrtn>



INTERNET ARCHIVE



# Intersection Analysis

# Top Domains

## Twitter

twitter.com **615,421**

cbc.ca **143,941**

youtube.com

**66,886**

huffingtonpost.ca **66,758**

theglobeandmail.com **63,401**

thestar.com **53,051**

ctvnews.ca **49,295**

globalnews.ca **46,488**

twimg.com **39,989**

macleans.ca **35,280**

## CPP (Aug - Nov 2015)

liberal.ca

**55,536**

greenparty.ca **45,788**

policyalternatives.ca **37,810**

socialist.ca **26,856**

davidsuzuki.org **25,487**

canadians.org **24,424**

ccrweb.ca **19,521**

afn.ca **15,879**

blocquebecois.org **10,899**

egale.ca **7,837**

	CPP	Twitter	Wayback
CPP		0.341%	74.30%
Twitter	0.269%		10.06%
Wayback	N/A	N/A	

# Included Domains (Internet Archive)

## Included

cbc.ca	3,035
youtube.com	2,639
thestar.com	1,665
theglobeandmail.com	1,644
huffingtonpost.ca	1,561
twitter.com	1,550
ctvnews.ca	1,423
nationalpost.com	1,262
globalnews.ca	1,062
ottawacitizen.com	836

## Excluded

twitter.com	173,931
linkis.com	11,071
youtube.com	6,026
instagram.com	5,302
globalnews.ca	4,709
cbc.ca	4,529
facebook.com	4,282
rabble.ca	3,859
huffingtonpost.ca	3,762
fw.to	3,284

# Conclusions



Advanced search

Username \*

Password \*

Log in

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.xml](#)

### In collections

- [#elxn42](#)

#### Details

Title:	#elxn42 web crawl
Creator(s):	<a href="#">Nick Ruest</a>
Note:	Tweet ids: <a href="http://hdl.handle.net/10864/11311">http://hdl.handle.net/10864/11311</a>
Identifier (local):	WEB-20160208134917869-00013-3991--rho.library.yorku.ca-9191-ELXN42
Identifier (md5):	63d707352a6fb45a62889c448154610f
Type:	<a href="#">Website</a>
Subject(s):	<a href="#">#elxn42</a> <a href="#">Canadian federal election, 2015</a> <a href="#">Canadian politics</a> <a href="#">Federal politics</a> <a href="#">Canada</a>
Date captured:	2016-01-29
Size:	13GB
File size:	12972947583
PUID:	<a href="#">x-fmt/266</a>
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 ( <a href="http://creativecommons.org/licenses/by/2.0/">http://creativecommons.org/licenses/by/2.0/</a> )



Advanced search

Username \*

Password \*

Log in

Collections / Web Archives for Historical Research / #panamapapers / #panamapapers crawl; May 1-7, 2016

## #panamapapers crawl; May 1-7, 2016

### Description

Consists of a web crawl of unique URLs tweeted with the #panamapapers hashtag. #panamapapers collection took place from April 4-29, 2016. Unique URLs were extracted from the dataset, and harvested with Heritrix on May 1-7, 2016.

### Download

- warc: [#panamapapers crawl; May 1-7, 2016.gz](#)
- cdx: [#panamapapers crawl; May 1-7, 2016.cdx](#)
- wat: [#panamapapers crawl; May 1-7, 2016.wat.gz](#)
- Seed list: [#panamapapers crawl; May 1-7, 2016.txt](#)
- Heritrix configuration: [crawler-beans.xml](#)

### In collections

- [#panamapapers](#)

#### Details

Title:	#panamapapers crawl; May 1-7, 2016
Creator(s):	<a href="#">Nick Ruest</a>
Note:	Tweet ids: <a href="http://hdl.handle.net/10864/11592">http://hdl.handle.net/10864/11592</a>
Identifier (local):	WEB-20160505182658813-00010-9949--rho.library.yorku.ca-9191-PANAMAPAPERS
Identifier (md5):	e43eb940bf56e340a85ab3897567edfa
Type:	<a href="#">Website</a>
Subject(s):	<a href="#">#panamapapers</a> <a href="#">politics</a> <a href="#">Panama Papers</a> <a href="#">Mossack Fonseca</a> <a href="#">International Consortium of Investigative Journalists</a> <a href="#">tax havens</a> <a href="#">web archiving</a> <a href="#">twitter</a>
Date captured:	2016-05-01
Size:	9.9GB
File size:	10547578814
PUID:	<a href="#">x-fmt/266</a>
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 ( <a href="http://creativecommons.org/licenses/by/2.0/">http://creativecommons.org/licenses/by/2.0/</a> )



# Questions

[ruestn@yorku.ca](mailto:ruestn@yorku.ca)

@ruebot