Enabling Access to Old Wu-Tang Clan Fan Sites

Facilitating Interdisciplinary Web Archive Collaboration

UNIVERSITY OF WATERLOO

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

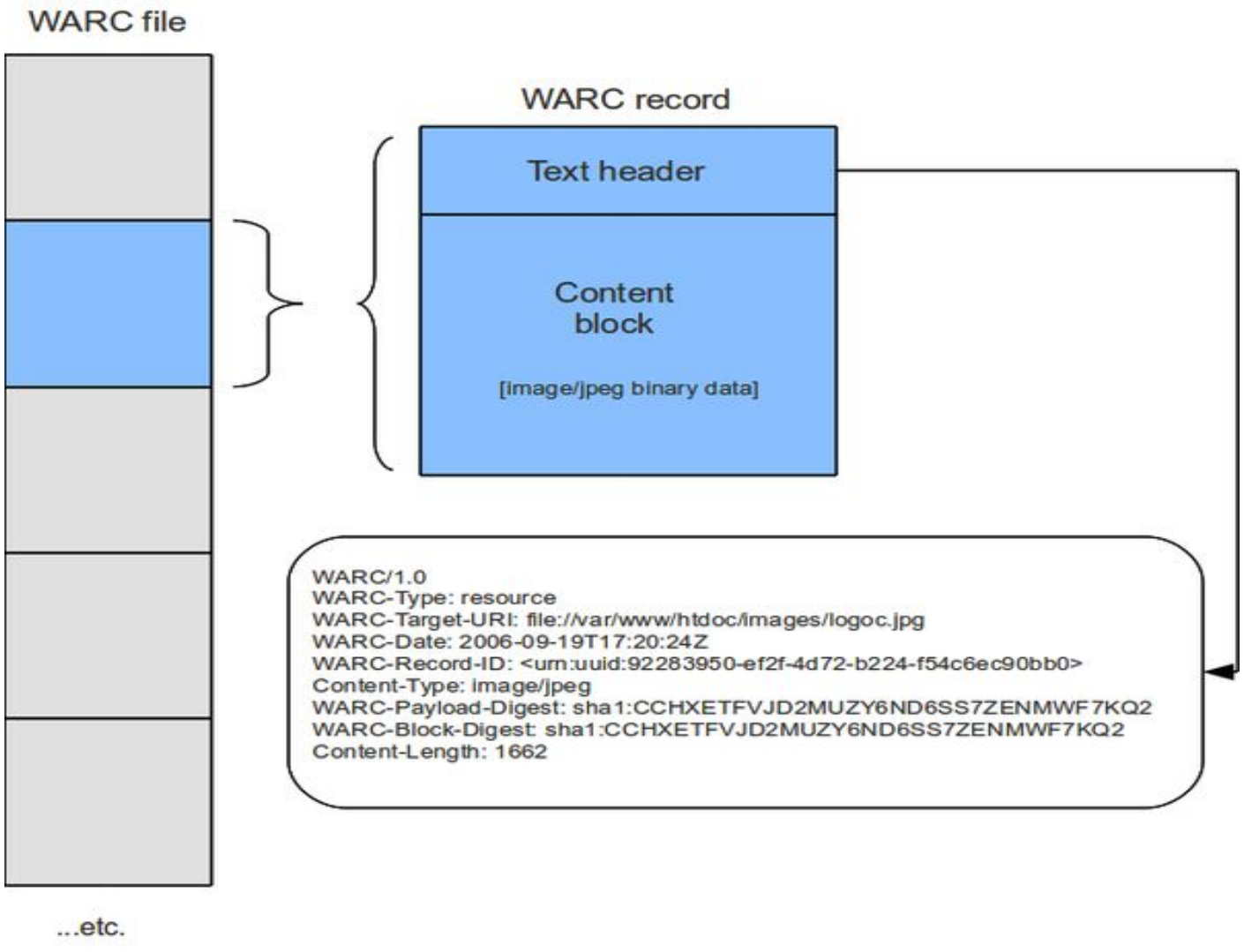
YORK UNIVERSITÉ UNIVERSITY


Why should we even care about web archives?

First, more data than ever before is being preserved...

Second, it'll be saved and delivered to us in very different ways

WARC (ISO 28500:2009)


WARC file
WARC record
Text header
Content block
[image/jpeg binary data]
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
...etc.

OccupyWallStreet

The revolution continues worldwide!

Farmers Join Occupy Wall Street, Calling for Food Justice

Posted 5 hours ago on Dec. 2, 2011, 6:21 p.m. EST by OccupyWallSt

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

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

Read More...

30 Comments

Occupy Wall Street Goes Home

Posted 1 day ago on Dec. 1, 2011, 3:04 p.m. EST by OccupyWallSt

On December 6th Occupy Wall Street will join in solidarity with a Brooklyn community to re-occupy a foreclosed home. The day of action marks a national kick-off for a new frontier for the

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

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

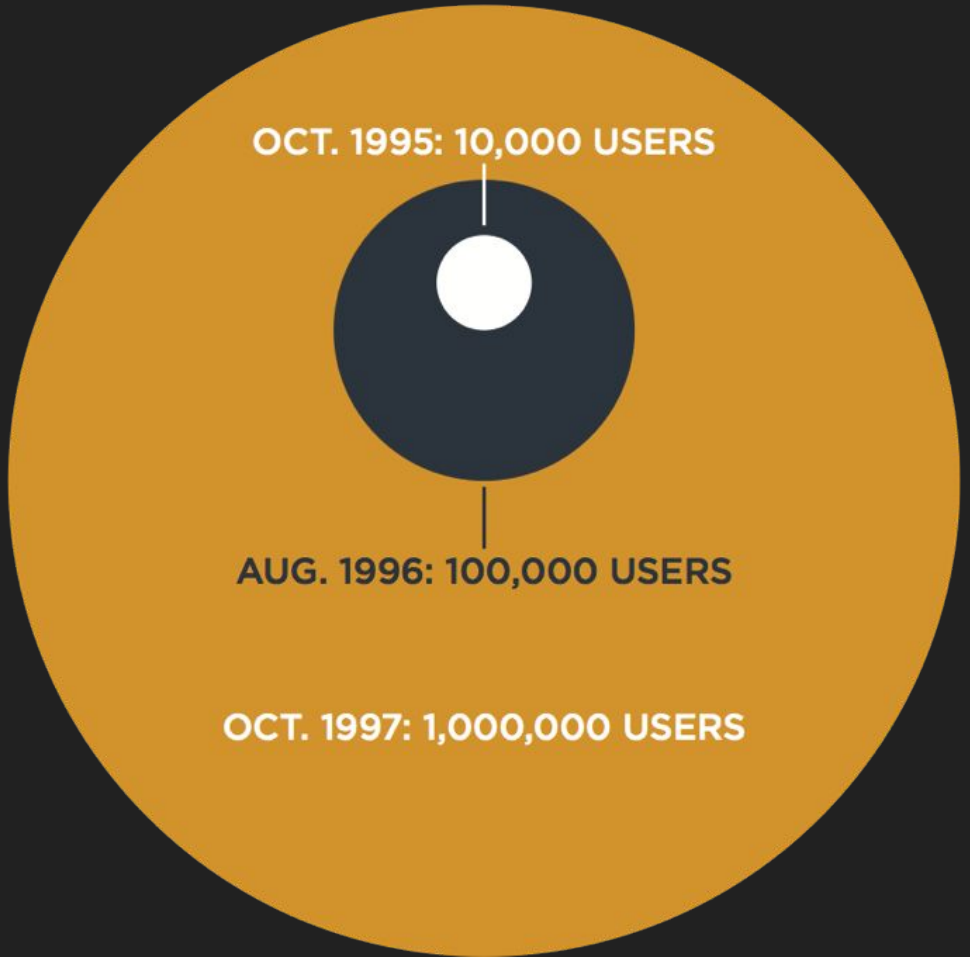
This #ows movement empowers real people to create real change from the bottom up. We want to see a general assembly in every backyard, on every street corner because we don't need Wall Street and we don't need politicians to build a better society.

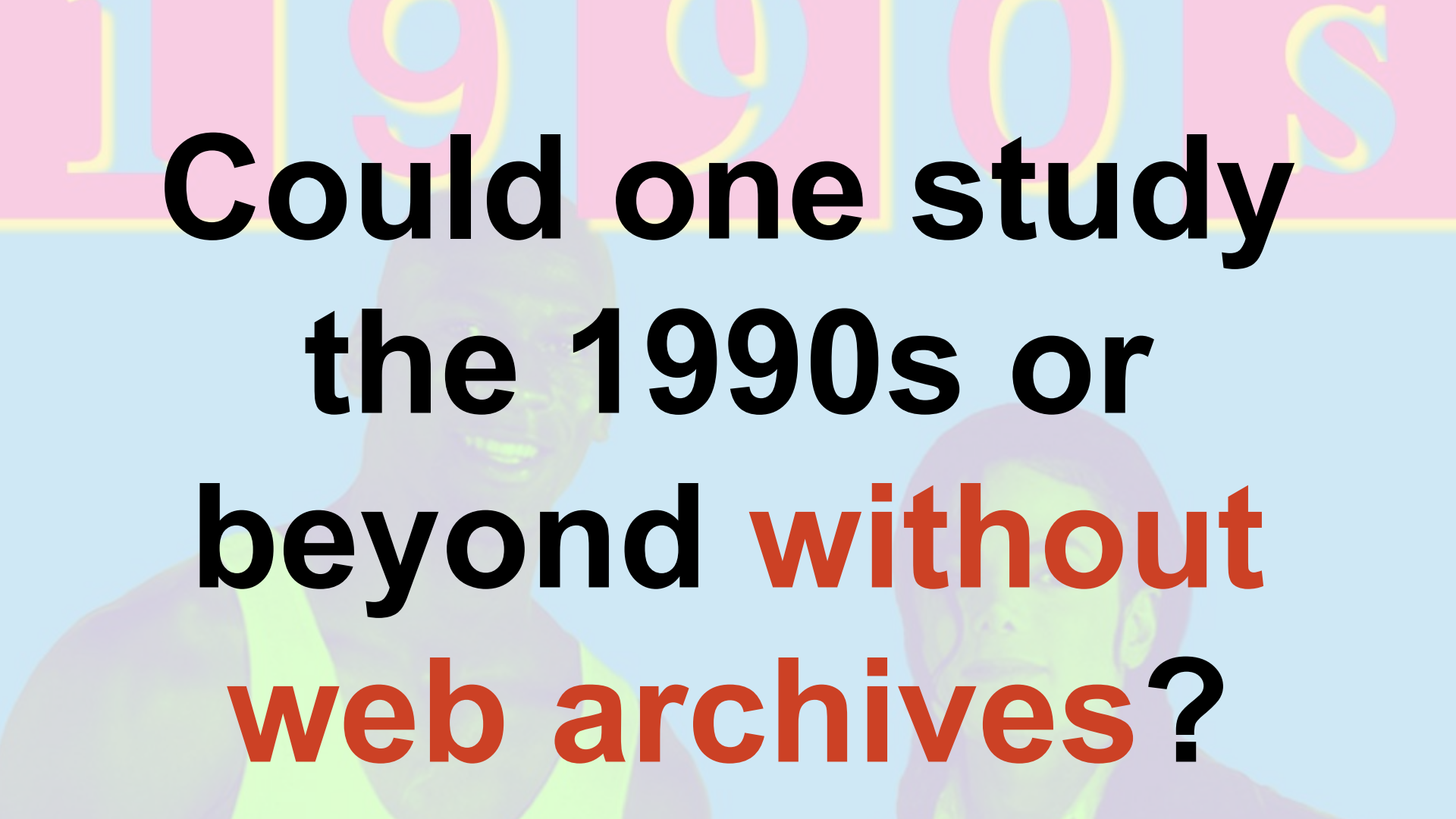
the only solution is WorldRevolution

Click here for NYCGA committee meeting times.


~~Scarcity~~
Abundance

GEOCITIES USERS:




1990s
Could one study the 1990s or beyond without web archives?

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

But right now you have to use the Wayback Machine - requiring you know the URL!

Internet Archive Wayback

web.archive.org/web/*/geocities.com

Sign In

INTERNET ARCHIVE WayBackMachine

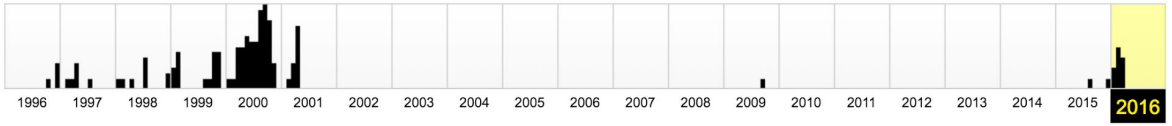
http://geocities.com

BROWSE HISTORY

http://geocities.com

Saved **1,675 times** between October 22, 1996 and March 5, 2016.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.



JAN

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	**22**	23
24	25	26	**27**	28	**29**	30
31						

FEB

	1	2	**3**	**4**	**5**	6
7	**8**	**9**	**10**	11	**12**	**13**
14	15	**16**	**17**	18	**19**	20
21	22	23	**24**	25	**26**	**27**
28	29					

MAR

		1	**2**	**3**	**4**	**5**
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

APR

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

MAY

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28

JUN

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25

JUL

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23

AUG

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27

Microsoft Visual Basic™ 5.0 Site Builder network
CONTROL CREATION EDITION

Visual Basic 5.0 Control Creation Edition Free!

GEOCITIES YOUR HOME ON THE WEB

- ENTER HERE
- INFORMATION
- NEIGHBORHOODS
- WHAT'S NEW
- WHAT'S COOL
- WHAT IS GEOCITIES?

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

Audio DIALweb Update

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

HOT GEOSTORE

Geotoon of the week

Happy Holidays from all of us at Geocities!

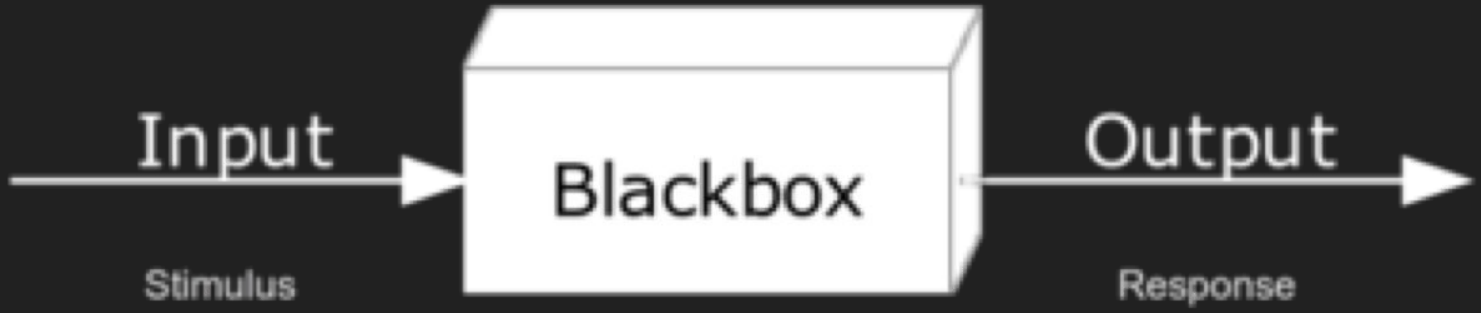
A message from our CEO

Today's Cool Homestead
WallStreet1456
Beginning investors and speculators won't want to miss the Working Class Investor Newsletter.

GeoCities News of the Day - 12/25/96

GEOCITIES LIVE CHRISTMAS TREE! ON CAMERA!

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 im our virtual holiday tree!

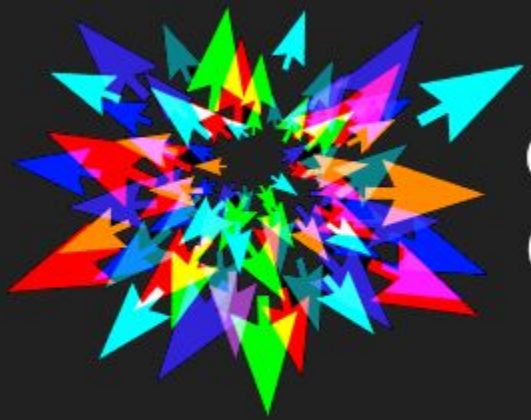

Input
Blackbox
Output
Stimulus
Response

We need interdisciplinary collaboration to tackle this problem!

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada



compute canada | calcul canada

Team(s)

We form like Voltron

WARCS
RULE
EVERYTHING
AROUND
ME (US!)

Ian Milligan

History Faculty Member

Jimmy Lin

Computer Science Faculty Member

Jeremy Wiebe

History PhD Candidate

Alice Zhou

Computer Science Undergraduate

Nick Ruest

Digital Assets Librarian

Collaboration

My beats travel like a vortex, through your spine
to the top of your cerebrum cortex

#Slack & GitHub

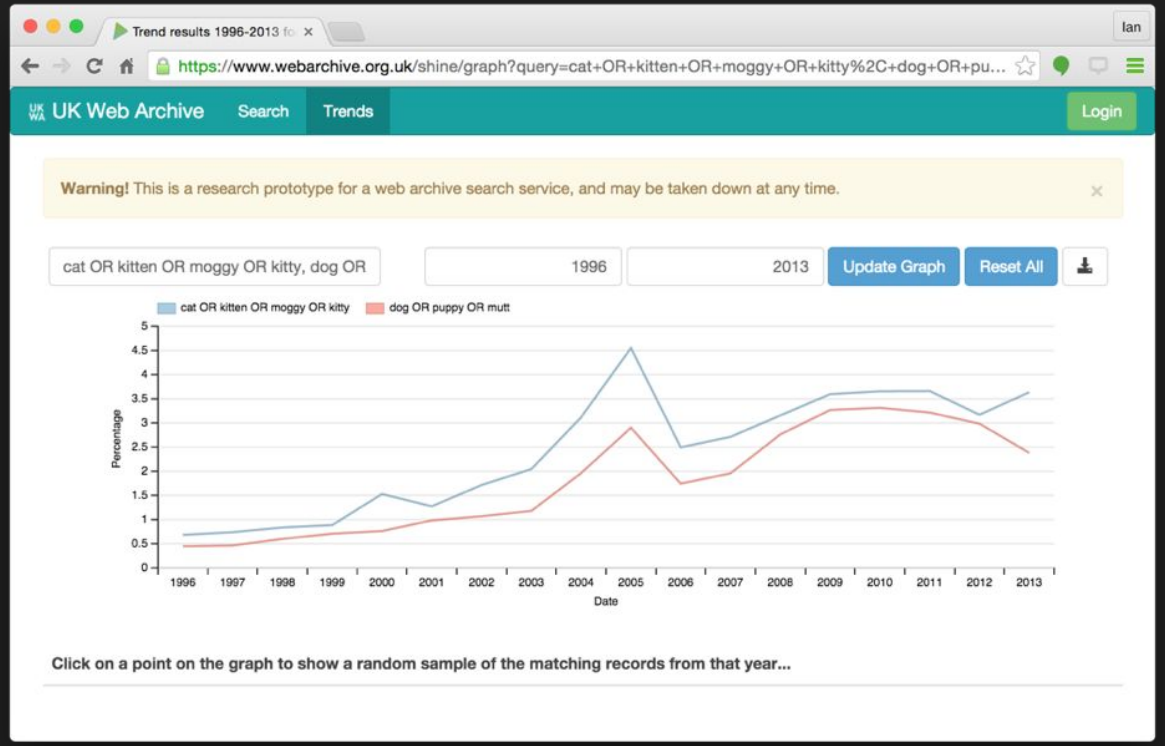
Platforms

Every time the horn blows, the Wu's signal's back on
Transform, pack form a whole another platform

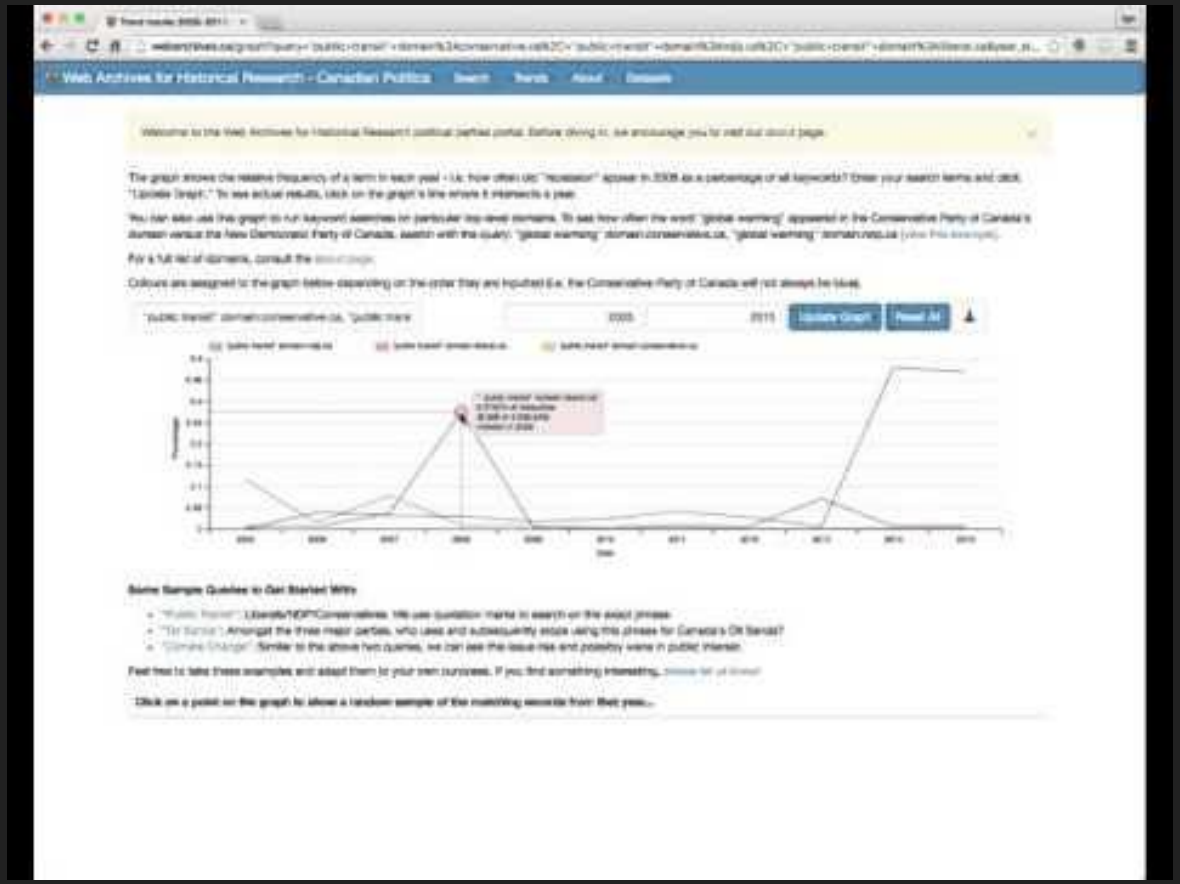
Shine

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

Shine


Trend results 1996-2013
Ian
https://www.webarchive.org.uk/shine/graph?query=cat+OR+kitten+OR+moggy+OR+kitty%2C+dog+OR+pu...
UK Web Archive
Search
Trends
Login
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
Update Graph
Reset All
cat OR kitten OR moggy OR kitty
dog OR puppy OR mutt
Percentage
Date
Click on a point on the graph to show a random sample of the matching records from that year...

webarchives.ca



CLI tools

awk, sed, grep, parallel, sort, uniq, wc, jq

Geocities

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.

- NEIGHBORHOODS
- WHAT'S NEW
- WHAT'S COOL
- WHAT IS GEOCITIES?

* Free Home Pages & Free Member Email Advertiser Information

Today's Cool Homestead GeoCities Daily Audio Update

HotSprings 1837

So you hit the snooze bar ten times every morning. You might be lazy. But then again, you might have a sleep disorder. Find out here.

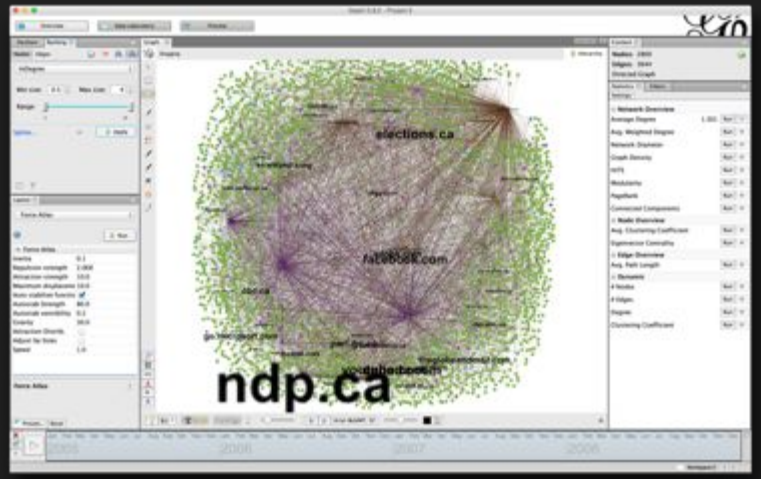
GeoCities News of the Day - 10/22/96

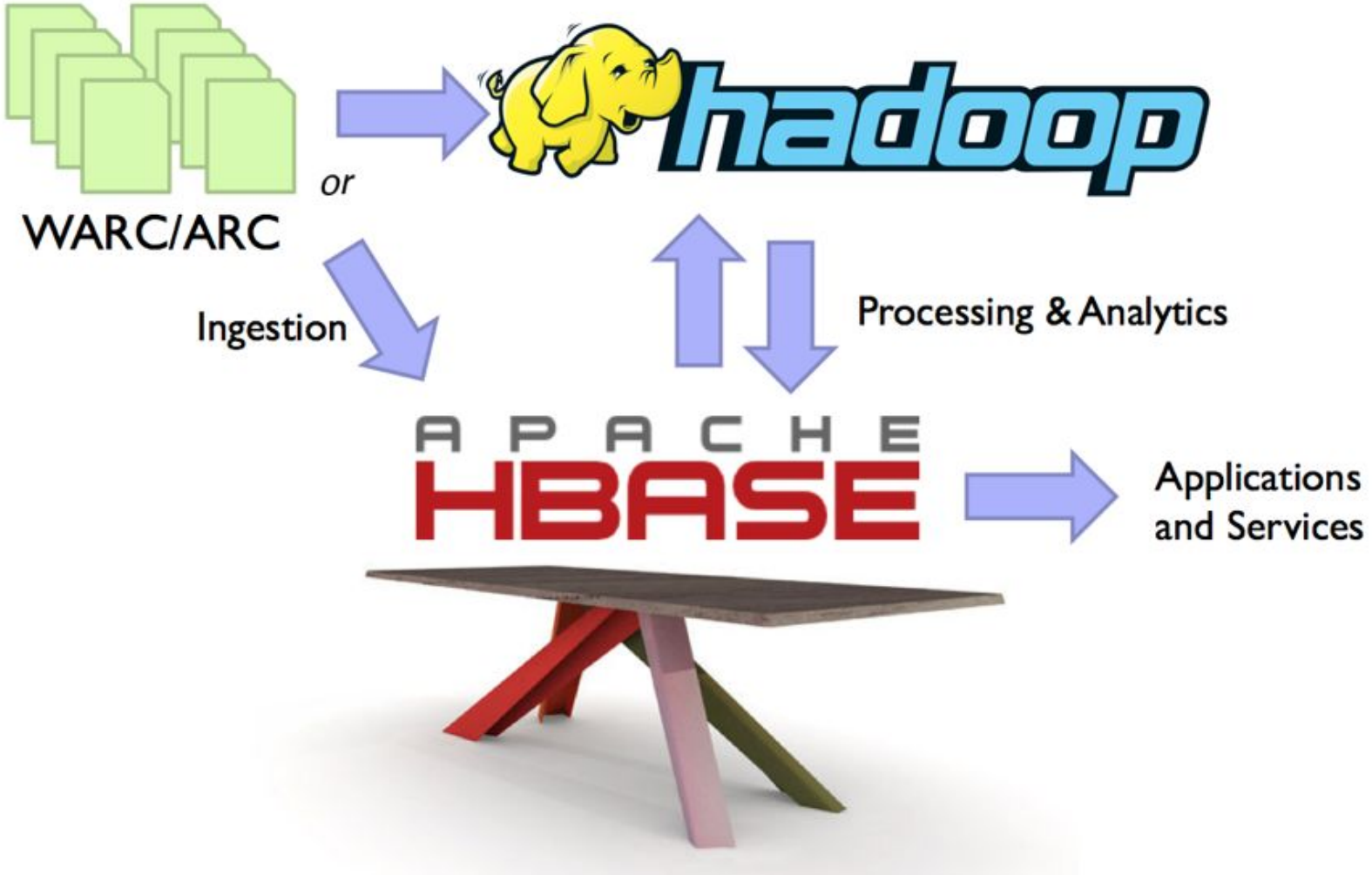
Warcbase

Warcbase

- An open-source platform for managing web archives

- Two main components
  - A flexible data store: your own Wayback Machine
  - Scriptable analytics and data processing

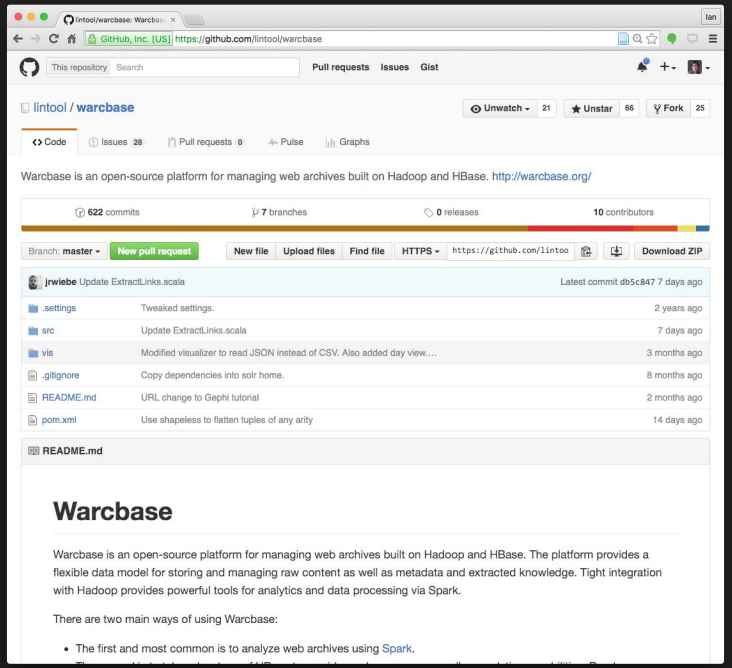



WARC/ARC
or
hadoop
Ingestion
Processing & Analytics
APACHE
HBASE
Applications
and Services

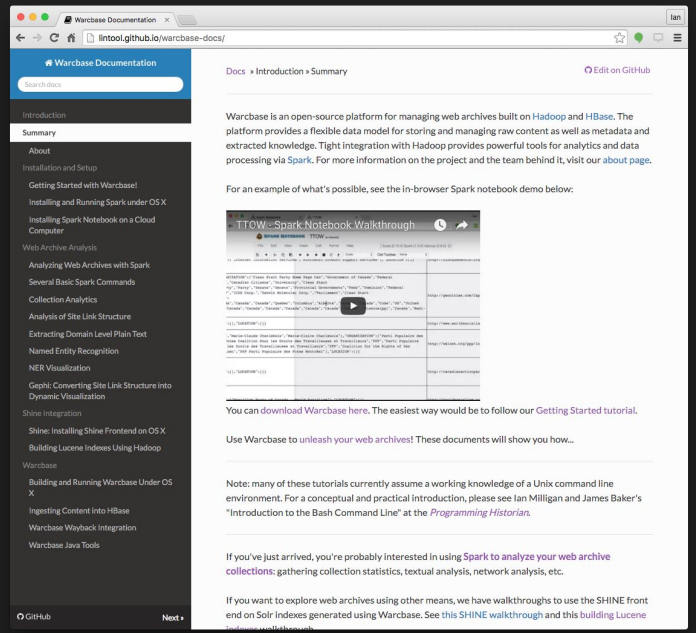
Warcbase

- 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 with Web Archives Historical Research Group

You can Warcbase Too! (...and Twarcbase soon!)



warcbase.org



docs.warcbase.org

Let’s do a quick walkthrough of how we’ve used it on GeoCities

1. i2millig@rho: /mnt/vol1/data_sets/geocities/warcs (ssh)

bash | bash | i2millig@rho: /mnt/vol1/data...

```
GEOCITIES-20091029114236-00191-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029115416-00171-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029123034-00172-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029130439-00173-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029134536-00174-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029140344-00192-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029141553-00193-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029141726-00175-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029144445-00176-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029152151-00177-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029160824-00178-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029164941-00179-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029165037-00194-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029170431-00195-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029171605-00180-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029174154-00181-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029180818-00182-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029182725-00183-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029185858-00184-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029193728-00185-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029194541-00196-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029195911-00197-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029202041-00186-crawling08.us.archive.org.warc.gz
GEOCITIES-20091029221340-00198-ia400110.us.archive.org.warc.gz
GEOCITIES-20091029222459-00199-ia400110.us.archive.org.warc.gz
GEOCITIES-20091030021147-00197-ia400103.us.archive.org.warc.gz
GEOCITIES-20091030021444-00198-ia400103.us.archive.org.warc.gz
GEOCITIES-20091030022413-00171-ia400104.us.archive.org.warc.gz
i2millig@rho:/mnt/vol1/data_sets/geocities/warcs$ du -h
4.1T    .
i2millig@rho:/mnt/vol1/data_sets/geocities/warcs$
```

```
ianmilligan1@Ians-MacBook-Pro:~$ rho
i2millig@rho.library.yorku.ca's password:
Welcome to Ubuntu 14.04.2 LTS (GNU/Linux 3.13.0-32-generic x86_64)

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

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

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

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

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

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

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

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

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

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

// Exiting paste mode, now interpreting.

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

scala>
```

Spark Notebook Demo

localhost:9000/notebooks/Spark%20Notebook%20Demo.snb#tab314587536-2

SPARK NOTEBOOK Spark Notebook Demo (unsaved changes)

File Edit View Insert Cell Kernel Help | Scala [2.10.4] Spark [1.3.0] Hadoop [2.6.0]

Code Cell Toolbar: None

C4L Hackathon Demo, March 2016

This is a notebook to demo how we're forseeing the rapid prototyping of work with web archives.

Note that we can begin to intersperse text with the code that we're writing, to enable the sharing of notebooks and research ideas.

In [ ]:
```
:cp /Users/ianmilligan1/dropbox/warcbase/target/warcbase-0.1.0-SNAPSHOT-fatjar.jar
```

...

In [ ]:
```
import org.warcbase.spark.matchbox._
import org.warcbase.spark.rdd.RecordRDD._
```

...

In [ ]:
```
var arc="/Users/ianmilligan1/Dropbox/warcs-workshop/227-20051004191331-00000-crawling015.archive.org.arc.gz";
var warc="/Users/ianmilligan1/dropbox/wahr/sample-data/arc-warc/ARCHIVEIT-227-QUARTERLY-XUGECV-20091218231727-00039-crawling06.us.
var arcdir="/Users/ianmilligan1/dropbox/warcs-workshop";
```

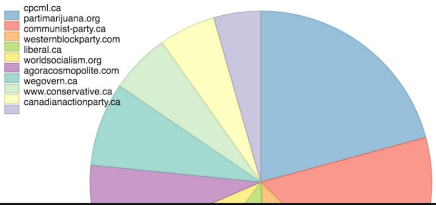
...

In [ ]:
```
val r =
RecordLoader.loadArc(arc,
sc)
.keepValidPages()
.map(r => ExtractTopLevelDomain(r.getUrl))
.countItems()
.take(10)
```

```
r: Array[(String, Int)] = Array((cpcml.ca,271), (partimarijuana.org,215), (communist-party.ca,156), (westernblockpart
y.com,144), (liberal.ca,107), (worldsocialism.org,105), (agoracosmopolite.com,103), (wegovern.ca,74), (www.conservativ
e.ca,70), (canadianactionparty.ca,58))
```

Out[4]:

10 items



Extracting all URLs

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

Results = 186,761,346 URLs, 9.9GB text file

Extracting a Link Graph

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

Results

```
1 ((20090903,http://geocities.com/saganaki2000/ADSLGR/adslgr.htm,
  http://www.adslgr.com),15337)
2 ((20091026,http://geocities.com/saganaki2000/ADSLGR/adslgr.htm,
  http://www.adslgr.com),15337)
3 ((20091027,http://geocities.com/spankbank69hard/,http://pg.photos
  .yahoo.com/ph/spankbank69hard/my_photos/),9807)
4 ((20090903,http://geocities.com/spankbank69hard/index.html,http:/
  /pg.photos.yahoo.com/ph/spankbank69hard/my_photos/),9807)
5 ((20091027,http://geocities.com/CollegePark/Locker/8187/,http://
  www.comercialuruapan.com),8056)
6 ((20090903,http://geocities.com/CollegePark/Locker/8187/,http://
  www.comercialuruapan.com),8056)
```

Creating Entities

403GB of link graph data.

- http://www.geocities.com/EnchantedForest/Grove/1234/index.html
- http://www.geocities.com/EnchantedForest/Grove/1234/pets/cats.html
- http://www.geocities.com/EnchantedForest/Grove/1234/pets/dogs.html
- http://www.geocities.com/EnchantedForest/Grove/1234/pets/rabbits.html

Bash-Fu

Find all four digit numbers:

```
sed 's/[()]*//g; s/^[^,]*,//; s/\([0-9]\{4\}\)[^,]*/\1/g'
enchantedforest-links.txt > enchantedforest-entities-cleaned1.txt
```

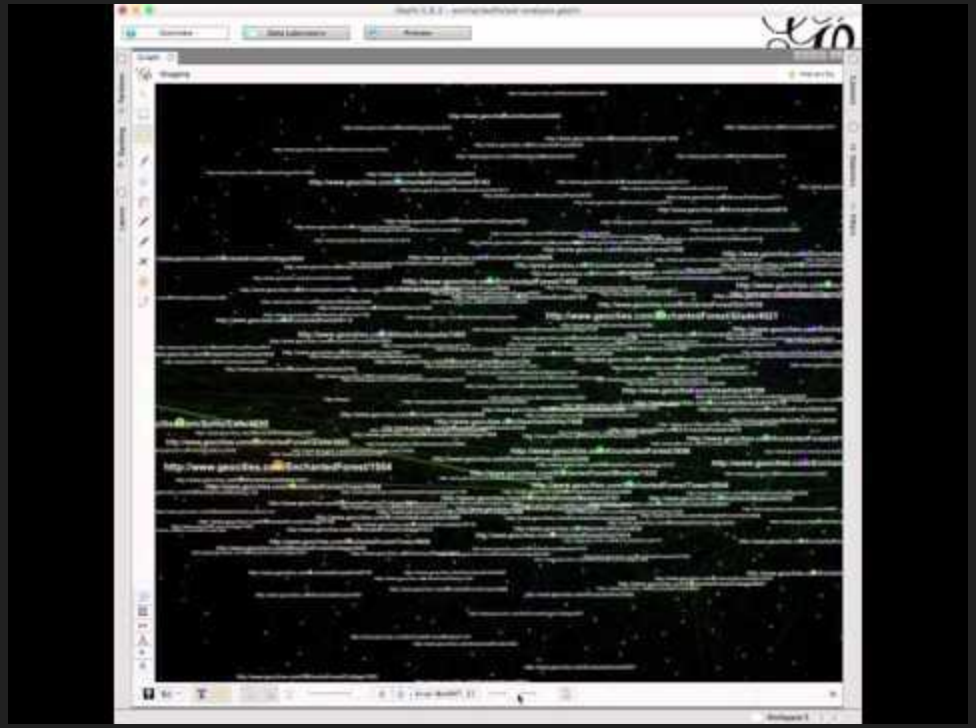
Then find internal:

```
grep -P '(.*/[0-9]{4}){2}' enchantedforest-entities-cleaned1.txt >
enchantedforest-entities-internal.txt
```

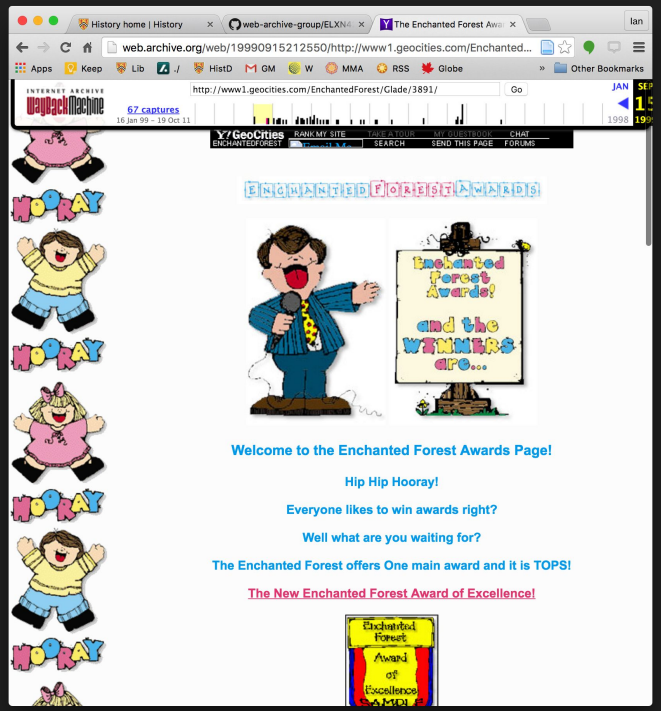
Link Structure

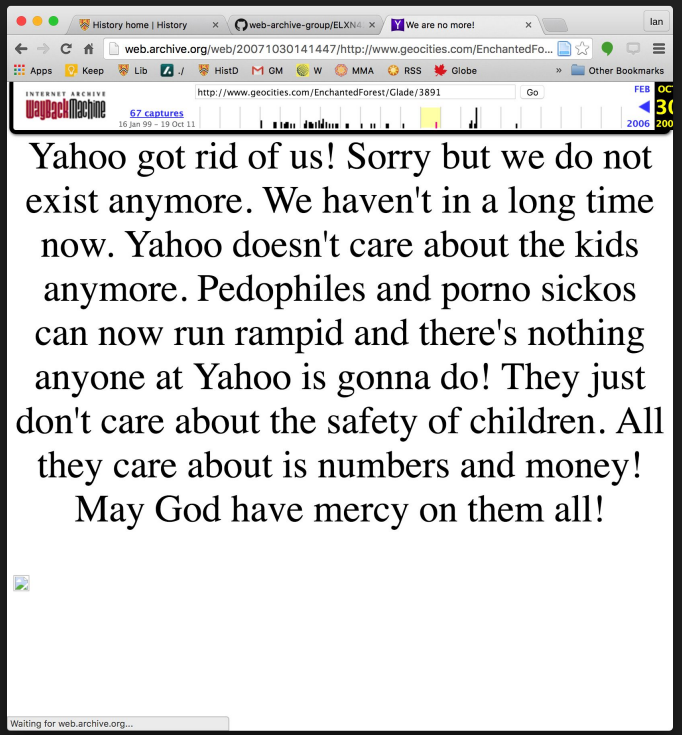
```
1 Source,Target,Weight
2 http://www.geocities.com/EnchantedForest/Meadow/1134,http://www.
  geocities.com/EnchantedForest/1004,83
3 http://www.geocities.com/EnchantedForest/Meadow/1134,http://www.
  geocities.com/EnchantedForest/1004,83
4 http://www.geocities.com/Area51/Stargate/1357,http://www.
  geocities.com/Area51/EnchantedForest/4213,33
5 http://www.geocities.com/Area51/Stargate/1357,http://www.
  geocities.com/Area51/EnchantedForest/4213,33
6 http://www.geocities.com/Eureka/1309,http://www.geocities.com/
  EnchantedForest/Tower/7555,27
7 http://www.geocities.com/Eureka/1309,http://www.geocities.com/
  EnchantedForest/Tower/7555,27
```


http://geocities.com/WestHollywood/9036
http://geocities.com/WestHollywood/2753
http://geocities.com/WestHollywood/Heights/4250
http://geocities.com/WestHollywood/1459
http://geocities.com/WestHollywood/1401
cities.com/WestHollywood/Castro/6902
http://geocities.com/WestHollywood/Village/9414
http://geocities.com/WestHollywood/Park/8021



EnchantedForest/Glade/3891





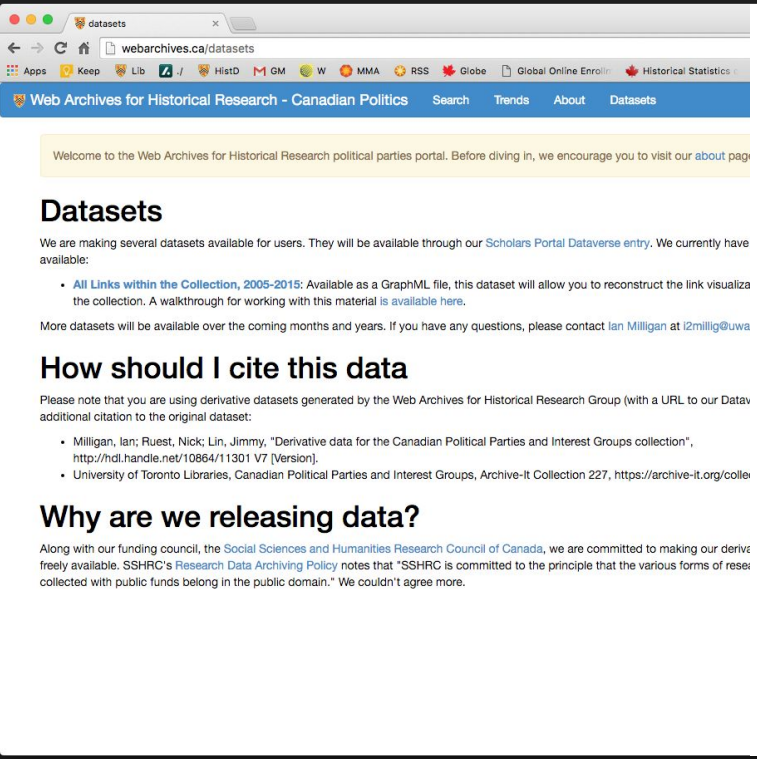
Historical Uses

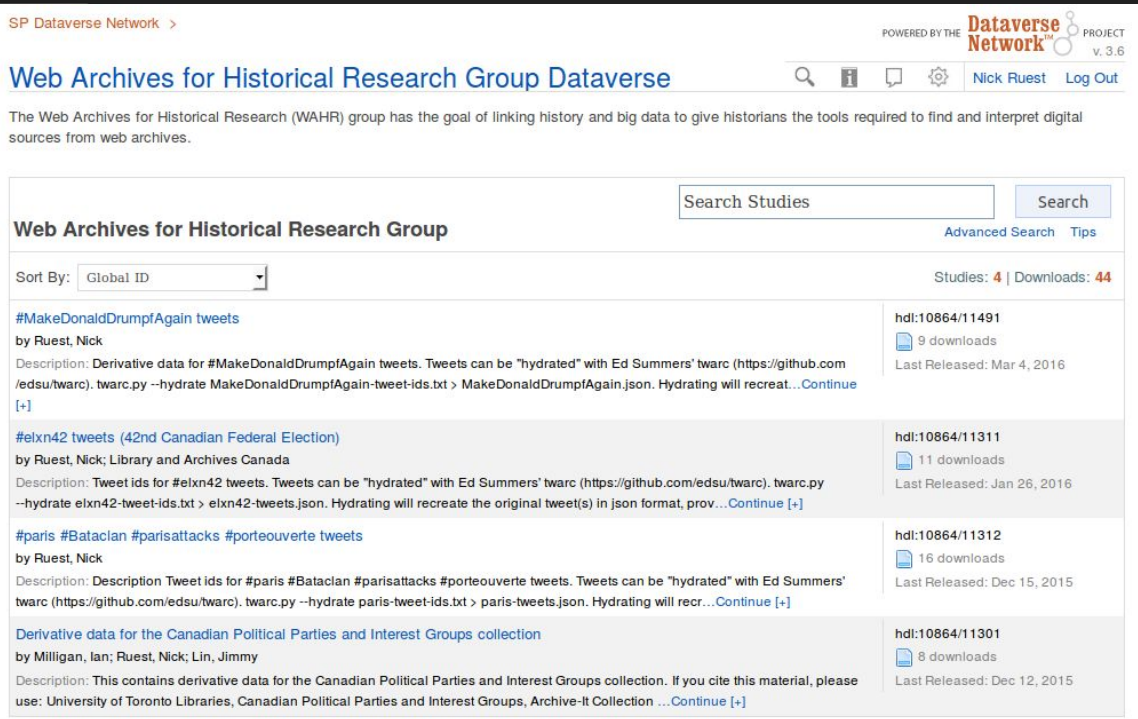
- The prevalence of awards pages and awards hubs within this neighbourhood;
- A protest movement that may have emerged when Yahoo! decided to shut down the neighbourhood;
- We can begin to follow links from this awards page, by highlighting it in Gephi, to find pages that hosted awards in connection with it;

We could do Shine indexing, but metadata might be the best way forward.

Also lets us share datasets!

Datasets


datasets
webarchives.ca/datasets
Web Archives for Historical Research - Canadian Politics
Search
Trends
About
Datasets
Welcome to the Web Archives for Historical Research political parties portal. Before diving in, we encourage you to visit our about pag
Datasets
We are making several datasets available for users. They will be available through our Scholars Portal Dataverse entry. We currently have available:
All Links within the Collection, 2005-2015: Available as a GraphML file, this dataset will allow you to reconstruct the link visualiza the collection. A walkthrough for working with this material is available here.
More datasets will be available over the coming months and years. If you have any questions, please contact Ian Milligan at i2millig@uwa
How should I cite this data
Please note that you are using derivative datasets generated by the Web Archives for Historical Research Group (with a URL to our Datav additional citation to the original dataset:
Milligan, Ian; Ruest, Nick; Lin, Jimmy, "Derivative data for the Canadian Political Parties and Interest Groups collection", http://hdl.handle.net/10864/11301 V7 [Version].
University of Toronto Libraries, Canadian Political Parties and Interest Groups, Archive-It Collection 227, https://archive-it.org/colle
Why are we releasing data?
Along with our funding council, the Social Sciences and Humanities Research Council of Canada, we are committed to making our deriv freely available. SSHRC's Research Data Archiving Policy notes that "SSHRC is committed to the principle that the various forms of rese collected with public funds belong in the public domain." We couldn't agree more.


SP Dataverse Network >
POWERED BY THE Dataverse Network PROJECT v. 3.6
Web Archives for Historical Research Group Dataverse
Nick Ruest
Log Out
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.
Search Studies
Search
Web Archives for Historical Research Group
Advanced Search
Tips
Sort By: Global ID
Studies: 4 | Downloads: 44
#MakeDonaldDrumpfAgain tweets
by Ruest, Nick
Description: Derivative data for #MakeDonaldDrumpfAgain tweets. Tweets can be "hydrated" with Ed Summers' twarc (https://github.com /edsu/twarc). twarc.py --hydrate MakeDonaldDrumpfAgain-tweet-ids.txt > MakeDonaldDrumpfAgain.json. Hydrating will recreat...Continue [+]
hdl:10864/11491
9 downloads
Last Released: Mar 4, 2016
#elxn42 tweets (42nd Canadian Federal Election)
by Ruest, Nick; Library and Archives Canada
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, prov...Continue [+]
hdl:10864/11311
11 downloads
Last Released: Jan 26, 2016
#paris #Bataclan #parisattacks #porteouverte tweets
by Ruest, Nick
Description: 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 recr...Continue [+]
hdl:10864/11312
16 downloads
Last Released: Dec 15, 2015
Derivative data for the Canadian Political Parties and Interest Groups collection
by Milligan, Ian; Ruest, Nick; Lin, Jimmy
Description: 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 ...Continue [+]
hdl:10864/11301
8 downloads
Last Released: Dec 12, 2015

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

By Napalm filled tires (Wu Tang Clan) [CC BY-SA 2.0 (http://creativecommons.org/licenses/by-sa/2.0)], via Wikimedia Commons

Contact

Nick Ruest: @ruebot
ruestn@yorku.ca
Ian Milligan: @ianmilligan1
i2milligan@uwaterloo.ca