Enabling Access to Old Wu-Tang Clan Fan Sites

Facilitating Interdisciplinary Web Archive Collaboration

Nick Ruest (@ruebot)
Ian Milligan (@ianmilligan1)
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/1.0
WARC-Type: resource
WARC-Target-URI: file://var/www/htdocs/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:CHXETFVJD2MUZY6ND6S7ZENMWF7KQ2
WARC-Block-Digest: sha1:CHXETFVJD2MUZY6ND6S7ZENMWF7KQ2
Content-Length: 1662
Farmers Join Occupy Wall Street, Calling for Food Justice

As Wall Street's corrupt influence on the economy has grown, the corporate ownership of our food system has hurt the health and livelihood of some of our most vulnerable communities. This Sunday, December 4th food justice activists and occupiers will be traveling 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, 3:56 p.m. EST by OccupyWallSt

NATIONAL DAY OF ACTION
DEC. 6, 2011

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 the 2012 campaign to occupy homes for the people, not the 1 percent.
GEOCITIES USERS:

OCT. 1995: 10,000 USERS

AUG. 1996: 100,000 USERS

OCT. 1997: 1,000,000 USERS
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!
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.
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

Building a home page for the holidays?
Submit your letters to Santa, favorite holiday recipes and other holiday cheer to our special "North Pole" neighborhood. And share your holiday spirit with GeoCitizens around the world by your virtual holiday tree!
We need interdisciplinary collaboration to tackle this problem!
Team(s)

We form like Voltron
WARCS RULE EVERYTHING AROUND ME (US!)
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
Welcome to the Web Archives for Historical Research - Canadian Politics portal. Before diving in, we encourage you to read our about page.

The graph shows the relative frequency of a term in each year. To see how often a term is used, hover over the graph's line where it intersects a year. You can also use this tool to highlight keywords associated with a particular keyword. To see how often the word "global warming" appeared in the Conservative Party of Canada's domain versus the New Democratic Party of Canada, search with the query: "global warming" domain:conservatives.ca, "global warming" domain:newdemocrats.ca (only for Canada).

For a full list of domains, consult the index page.

Colors are assigned to the graph based on the order they are input. For example, the Conservative Party of Canada will not always be blue.

"Public transit" domain:conservatives.ca, domain:newdemocrats.ca

Some Sample Queries to Get Started With:
- "Public transit": Liberal/NDP Conservatives. We use quotation marks to search on the exact phrase.
- "The Senate": Amongst the three major parties, who has the highest number of seats this year for Canada's 13 Senate?
- "Change Liberal": Similar to the above two queries, but we can see the issue has not yet been raised in public interest.

Feel free to take these examples and adapt them to your own purposes. If you find something interesting, please let us know!

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

awk, sed, grep, parallel, sort, uniq, wc, jq
Geocities
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
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!)
Let’s do a quick walkthrough of how we’ve used it on GeoCities
242 packages can be updated.
130 updates are security updates.


WARM NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
Welcome to

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.

WARM 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> spark
// Entering paste mode (ctrl-D to finish)

import org.apache.spark.sql.SparkSession
import org.apache.spark.SparkContext

val r = spark.sparkContext()

// Exiting paste mode, now interpreting.

INFO SparkSession$InputFormat - Loading file /mnt/voll/data_sets/geocities/warc/GEOCITIES-2000000133634-04399-crawling08.us.archive.org.warc.gz
C4L Hackathon Demo, March 2016

This is a notebook to demo how we're foreseeing 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 []: import org.apache.spark.matchbox._
     import org.apache.spark.rdd.RECORDBDD_

In []: var x = Users/lamilliganl/dropbox/warc-2005100419331-00000-22720-0511129231797-000039-crawling015.archive.org.pf;

In []: var aradir = Users/lamilliganl/dropbox/warc-2005100419331-00000-22720-0511129231797-000039-crawling015.archive.org.pf/

In []: val r = RecordLoader.loadArc(x, aradir)
     .keepValidPages()
     .map{r => ExtractTopLevelDomain{r.getUrl}()
     .countTerms()
     .take(10)

r: Array[(String, Int)] = Array((opmc.ca,271), (partimarijuana.org,215), (communist-party.ca,156), (westernblockparty.com,144), (liberal.ca,147), (worldsocialism.org,105), (agoracommopolite.com,103), (wegovern.ca,74), (www.conservativ.ca,70), (canadianactionparty.ca,59))
```

Out[4]:

10 Items

```
10 Items

<table>
<thead>
<tr>
<th>Domain</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>opmc.ca</td>
<td>271</td>
</tr>
<tr>
<td>partimarijuana.org</td>
<td>215</td>
</tr>
<tr>
<td>communist-party.ca</td>
<td>156</td>
</tr>
<tr>
<td>westernblockparty.com</td>
<td>144</td>
</tr>
<tr>
<td>liberal.ca</td>
<td>147</td>
</tr>
<tr>
<td>worldsocialism.org</td>
<td>105</td>
</tr>
<tr>
<td>agoracommopolite.com</td>
<td>103</td>
</tr>
<tr>
<td>wegovern.ca</td>
<td>74</td>
</tr>
<tr>
<td><a href="http://www.conservativ.ca">www.conservativ.ca</a></td>
<td>70</td>
</tr>
<tr>
<td>canadianactionparty.ca</td>
<td>59</td>
</tr>
</tbody>
</table>
```
Extracting all URLs

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

Results = 186,761,346 URLs, 9.9GB text file
Extracting a Link Graph

```scala
import org.warcbase.spark.matchbox.{ExtractTopLevelDomain, ExtractLinks, RecordLoader}
import org.warcbase.spark.rdd.RecordRDD._

RecordLoader.loadArc("/mnt/vol1/data_sets/geocities/warcs/*", sc)
    .keepValidPages()
    .map(r => (r.getCrawldate, ExtractLinks(r.getUrl, r.getContentType)))
    .flatMap(r => r._2.map(f => (r._1, ExtractTopLevelDomain(f._1).replaceAll("[^\\s*www\\.\"]", ""), ExtractTopLevelDomain(f._2).replaceAll("[^\\s*www\\.\"]", ")
    .filter(r => r._2 != "" && r._3 != "")
    .countItems()
    .filter(r => r._2 > 5)
    .saveAsTextFile("/mnt/vol1/data_sets/geocities/geocities.sitelinks")
```
Creating Entities

403GB of link graph data.

Find all four digit numbers:

```bash
sed 's/[(\)]*/g; s/[^,]*,//; s/\([0-9]{4}\)[^,]*/\1/g'
```

`enchantedforest-links.txt` > `enchantedforest-entities-cleaned1.txt`

Then find internal:

```bash
grep -P '(.*/[0-9]{4}){2}' enchantedforest-entities-cleaned1.txt > enchantedforest-entities-internal.txt
```
<table>
<thead>
<tr>
<th></th>
<th>Source, Target, Weight</th>
</tr>
</thead>
</table>
Yahoo got rid of us! Sorry but we do not exist anymore. We haven't in a long time now. Yahoo doesn't care about the kids anymore. Pedophiles and porno sickos can now run rampant and there's nothing anyone at Yahoo is gonna do! They just don't care about the safety of children. All they care about is numbers and money! May God have mercy on them all!
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

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.

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 visualization 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 smiling_box

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 Data additional citation to the original dataset:


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 data freely available. SSHRC's Research Data Archiving Policy notes that "SSHRC is committed to the principle that the various forms of research collected with public funds belong in the public domain." We couldn't agree more.
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