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.
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 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, 3:54 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 take advantage of foreclosed homes for their use. The action was.
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 neighborhood. And share your holiday spirit with GeoCitizens around the world!
We need interdisciplinary collaboration to tackle this problem!
Team(s)
WARCS RUL一切 WITH 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
CLI tools

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

bash

du -h

4.1T

du -h

4.1T
$ rph
2mwilli@ho.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 7 13:43:20 EST 2016

System load: 0.99
Usage of /: 34.7% of 744.67GB
Memory usage: 16%
Swap usage: 6%
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 7 13:43:21 2016 from 192.168.1.254

$ /spark-1.5.1/bin/spark-shell --jars /warchase/target/warchase-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> import org.apache.spark.sql.Context
import org.apache.spark.sql.Context

scala> val r = RecordLoader.loadWarc("/mnt/voll/data_sets/geocities/warcset/GEOCITIES-20000000133634-04399-crawling08.us.archive.org.warc.gz", sc)
scala> r

scala>
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.

```scala
In []:  scp /Users/ianmlligan1/dropbox/warehouse/target/warehouse-0.1.0-SNAPSHOT-fatjar.jar

In []:  import org.apache.spark.sql.spark

In []:  var archives = "Users/ianmlligan1/dropbox/webcam-data/arc/archive.org.archive.org"

In []:  var arc = "Users/ianmlligan1/dropbox/webcam-data/arc/archive.org.archive.org"

In []:  var arcdir = "Users/ianmlligan1/dropbox/webcam-data/arc/archive.org.archive.org"

In []:  val r =
  RecodeLoader.loadArc(archive,
  keepValidPages()
  .map{r => ExtractTopLevelDomain(r.getUrl)}
  .countItems()
  .take(10)


Out[4]:
```

10 Items

- opcml.ca
- partymarijuana.org
- communist-party.ca
- westernblockparty.com
- liberal.ca
- worldsocialism.org
- agoracommunity.com
- wegovern.ca
- www.conservative.ca
- canadianactionparty.ca

![Pie chart image]
Extracting all URLs

```scala
import org.warctools.spark.matchbox._
import org.warctools.spark.rdd.RecordRDD._

val r = RecordLoader.loadWarc("/mnt/vol1/data_sets/geocities/warcs", sc)
  .keepValidPages()
  .map(r => r.getUrl)
  .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.
        getContentString)))
    .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")
```
Results

Creating Entities

403GB of link graph data.

Bash-Fu

Find all four digit numbers:

```
sed 's/([()]*//g; s/^[^,]*,;//; s/\([0-9]{4}\){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
```
<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Target</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td><a href="http://www.geocities.com/Area51/Stargate/1357">http://www.geocities.com/Area51/Stargate/1357</a></td>
<td><a href="http://www.geocities.com/Area51/EnchantedForest/4213,33">http://www.geocities.com/Area51/EnchantedForest/4213,33</a></td>
<td></td>
</tr>
</tbody>
</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 rampid 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

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, 2009-2015. Available as a GraphML file, this dataset will allow you to reconstruct the link visualization 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 Miligan at SmilingBox.

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:

- Milligan, Ian; Ruest, Nick; Lin, Jimmy. "Derivative data for the Canadian Political Parties and Interest Groups collection," http://hdl.handle.net/10664/12027 (version).

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://uwaterloo.ca/web-archive-group/)
- [https://github.com/web-archive-group/](https://github.com/web-archive-group/)
- [https://github.com/ianmilligan1/](https://github.com/ianmilligan1/)
- [https://github.com/ruebot](https://github.com/ruebot)
Contact

Nick Ruest: @ruebot
ruestn@yorku.ca

Ian Milligan: @ianmilligan1
i2milligan@uwaterloo.ca