

# #SeedsOfAnarchy

...

Creatively thinking about automating quality in web archiving using visual analysis...

# Brenda's PhD in a nutshell

- Created a multi-dimensional model of Informational Quality for a web archives
- Proposed and explored which could be automated
- Three main:
  - Correspondence
  - Relevance
  - Archivability

# Subdimensions

- Correspondence
  - Visual correspondence
  - Interactional correspondence
  - Completeness
- Relevance
  - Size relevance
  - Topic relevance

**Much discussion messing  
about with Gephi & various  
web collections...before a  
plan formed...**

Working with Anarchist Archives incited  
**ANARCHY** within the group!

## Two projects emerged!

1. Comparing the images from the seed websites to images from URLs as they get farther from seed.

Theory being that they will get less-relevant the farther away from original.

2. Compare screenshots of the current live websites to screenshots of seeds crawled.

Theory being that variance from the current original will create useful flags for whoever is doing Quality.

# Project 1

Chose some seeds from  
University of Victoria's:  
**Anarchist Archives**

Title: 325

URL: <http://325.nostate.net/>

Captured 4 times between Nov 3, 2016 and Nov 3, 2016

Title: 38 Blood Alley Square | the anarchist space menacing blood alley

URL: <http://38bloodalley.wordpress.com/>

Captured 5 times between Jul 8, 2014 and Nov 3, 2016

Title: Halifax Anarchist Black Cross

URL: <http://abc.h-a-z.org/>

Captured 4 times between Jul 8, 2014 and Nov 3, 2016

Title: Calgary Anarchist Bookfair

URL: <http://calgaryanarchistbookfair.blogspot.ca/>

Captured 4 times between Jul 8, 2014 and Nov 3, 2016

Title: Camas Books & Infoshop

URL: <http://camas.ca/>

Captured 5 times between Jul 8, 2014 and Nov 3, 2016

Title: Edmonton Anarchist Bookfair

URL: <http://eabf.ca/>

Captured once on Jul 8, 2014

# Ryan

Wrote some code. The plan was to compare images - **technical difficulties** - it turns out we learned useful things from the concentric lists of URLs as they get further from the seed(s).



# This code worked! (contact #seedsofanarchy if you want it)

7523

```
h [224]: one <- neighbors(graph, vid, "out")$label
```

```
h [225]: FUN <- function(url) {  
  return (neighbors(graph, which(V(graph)$label == url), "out")$label)  
}
```

```
FILTERFUN <- function (url) {  
  return (!(url %in% one))  
}
```

```
two <- lapply(one, FUN)  
two <- lapply(two, rbind)  
two <- unlist(two, recursive=FALSE)  
two <- two[!(two %in% one)]  
two <- unique(two)
```

```
h [226]: three <- lapply(two, FUN)  
three <- lapply(three, rbind)  
three <- unlist(three, recursive=FALSE)
```

```
import io.archivesunleashed._  
import io.archivesunleashed.app._  
import io.archivesunleashed.matchbox._  
sc.setLogLevel("INFO")
```

```
val links = RecordLoader.loadArchives("/mnt/data/uvic-anarchist-archives/warcs/*.gz", sc)  
  .keepValidPages()  
  .map(r => (r.getCrawlDate, ExtractLinks(r.getUrl, r.getContentString)))  
  .flatMap(r => r._2.map(f => (r._1, ExtractDomain(f._1).replaceAll("^\\s*www\\.\"", ""))  
  .countItems()  
WriteGraphML(links, "/home/ubuntu/wildfires3.graphml")
```

```
import io.archivesunleashed._  
import io.archivesunleashed.app._  
import io.archivesunleashed.matchbox._
```

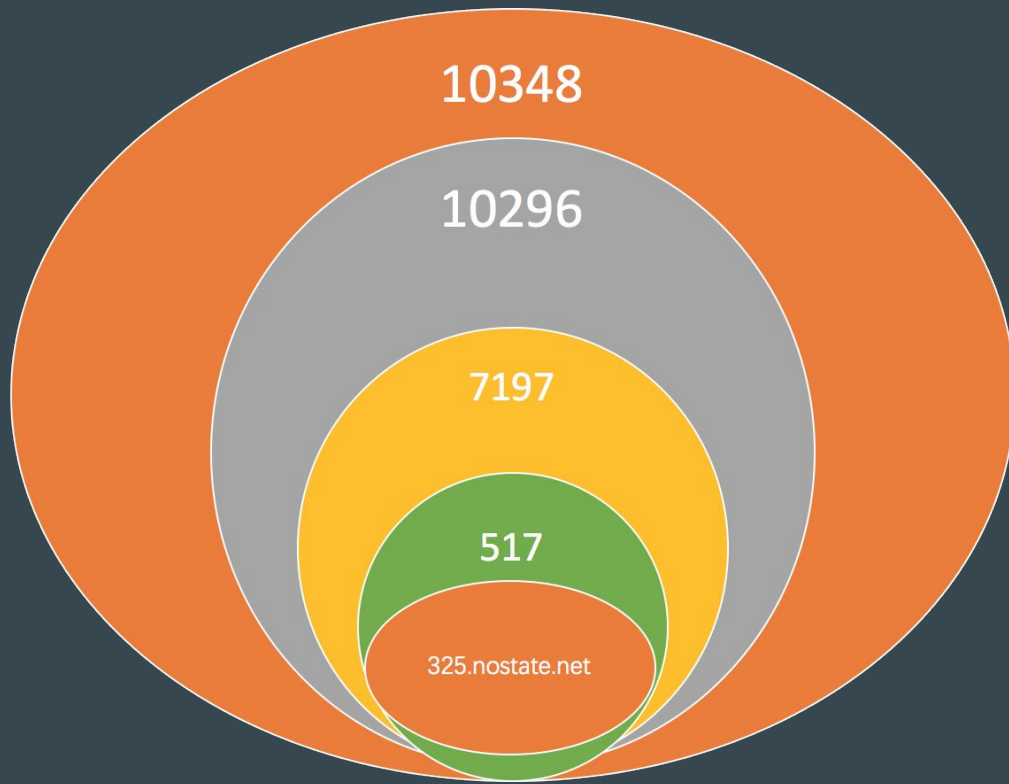
```
val urls = spark.read.csv("/home/ubuntu/QUALITY/items.csv").rdd  
val partitions = urls.getNumPartitions  
val A: org.apache.spark.rdd.RDD[scala.util.matching.Regex] = urls  
  .map(r => "\\b" + r(1).toString.replaceAll(raw"""\"?*\(\)\<\>\"?\"""", "")) +  
  .map(s => s.r)  
  .mapPartitionsWithIndex { (idx, iter) =>  
    if (idx == 0) iter.drop(1)  
    else if (idx == partitions - 1) iter.sliding(2).map(_.head)  
    else iter  
  }
```

```
org.apache.spark.rdd.RDD[scala.util.matching.Regex] = urls.map(r => "\\b"  
  .mapPartitionsWithIndex { (idx, iter) =>  
    if (idx == 0) iter.drop(1)  
    else if (idx == partitions - 1) iter.sliding(2).map(_.head)
```

```
org.apache.spark.rdd.RDD[scala.util.matching.Regex] = urls.map(r => "\\b"  
  .mapPartitionsWithIndex { (idx, iter) =>  
    if (idx == 0) iter.drop(1)  
    else if (idx == partitions - 1) iter.sliding(2).map(_.head)
```

# Concentric links

out from the seed URL:  
325.nostate.net; each  
number refers to the  
number of urls in the  
stage.



A	B	C	D	E
demotix.com	prisonactivist.org	malmo.e.org	dmpibooks.com	sfu.ca
ines.wordpress.com	justiceformarissa.blogspot.com	monk-e.bandcamp.com	pagesfromtheoaktree.bandcamp.com	american.edu
325.nostate.net	publiceyeonline.com	hawaiinewsdaily.com	blacklistednews.com	poisonedfiction.blogspot.com
pastebin.com	mathaba.net	zaf.anarhija.org	buty.a349.info	ceaa-acee.gc.ca
ocial.wordpress.com	nevertrustacop.org	https	miss222.com	cispes.org
youtu.be	tech.rebellyon.info	chaoticinsurrectionensemble.org	myav.good-tw.info	portland.indymedia.org
files.wordpress.com	maisonmelleurprix.com	multi.lectual.net	canlil.ca	aas383.e465.info
autistici.org	dryriver.org	rhizomecafe.ca	adriennemareebrown.net	7659.info
offire.espiwblogs.net	politicalfailblog.com	bevreece.co.za	southwalesanarchists.wordpress.com	juarezdialoga.org



Contact

Direct

Action

Distro

## Indonesia: 'No (ITS), iNo Seguirás Adelante!' por Eat

November 1st, 2018

EN: 'No (ITS), You Will Not Go On Forward!' by Eat (Indonesia)

Han pasado tanta mierda de parte de la llamada pandilla Eco-Extremista y de todas sus porristas misántropas, nihilista y aspirantes. Encontré su texto del año pasado, "El mito anarquista", lleno de cristianismo vulgar que esconden como eco-extremismo. Vamos, ¿es esto lo mejor que pueden hacer? ¿Afirmando haber matado a personas sin esperanza en alguna ocasión? Y leyendo sus amenazas de muerte mientras publicas fotos de alguien que asesinaste (según tu reivindicación, por supuesto), es realmente divertido, no, realmente lo es (realmente me río, literalmente cuando leo los comunicados y las fotografías, ¿qué mierda?).

(MÉXICO) 60 COMUNICADO DE ITS: SOBRE LA DELACIÓN DE LA ONG 325

(MÉXICO) 60 COMUNICADO DE ITS: SOBRE LA DELACIÓN DE LA ONG 325



### Anti-Info

- ABC Belarus
- ABC Brighton
- ABC Dresden
- ABC Indonesia
- ABC Wien
- Act for Freedom Now
- Agitasi
- Anarchist Defence Fund
- Anarchy Today
- Anarhija
- Anarkism
- Anarchist Libraries
- Anti-Fascist Network

# 325.nostate.net

Screenshot of items in each list of URLs. By 3 links away it's mostly local news sites and by 5 away it's all spam & gov't sites.

# camas.ca

Screenshot of items  
in each list of URLs.  
By 3 links away it's  
NGO sites not as  
dramatic of a  
change. Possibly  
indicating a closer  
community.

AA	BB	CC	DD	EE
warriorpublications.wordpress.com	thekebabonakeepers.com	latinwavesmedia.com	channel-chat.info	eastbayexpress.com
indiegogo.com	theturtleislandnews.com	luxdev.org	nybooks.com	thebolditalic.com
tinyurl.com	maps.ubc.ca	buynothingxmas.org	satyamag.com	blackoutprint.tumblr.com
facebook.com	elfpressoffice.org	doctorsforrefugeecare.ca	labourspeaksout.com	channel-tube.info
judibari.org	stmarksbookshop.com	varjokirjamesut.wordpress.com	mobi.good-tw.info	thebarrieexaminer.com
adastracomix.com	raisethereates.org	ecovillagenews.org	riot2011.vpd.ca	aki-kawamura.7359.info
wildcoast.ca	RisingTide804.net	electricgreen.org	greens.org	pics.lockerz.com
unistotencamp.com	xueronghua.org	blacklawrence.com	outerspace.good-tw.info	buzz.blogger.com
shac7.com	photos.google.com	greenisthenewred.com	anti-racistcanada.blogspot.com	sexdiy.b162.info
cloggedarteries.bandcamp.com	pacifictfutureenergy.com	chiapas.indymedia.org	dolove.a349.info	earthwarriorsrising.wordpress.com
shawnsunky.com	stoptheflows.com	geogroup.com	joannelehrer.wordpress.com	occuworld.org
greenisthenewred.com	goo.gl	roguemedia.org	gorillaradioblog.blogspot.ca	live.b162.info
beehivecollective.org	intelligence.ca	accessinuruguay.wordpress.com	shiningsoul-music.blogspot.com	tanfonline.com
afreeskool.wordpress.com	cyresshill.com	broadwayworld.com	journals.plos.org	urban75.org
upsideownworld.org	mqup.ca	aequanimio.com	proquest.umt.com.ezproxy.library.yorku.ca	gendertrender.wordpress.com
jevents.net	dontgetlazy.wordpress.com	submedia.tv	dudusex.176go.info	gilderlehman.org
friesenpress.com	bchousing.org	videofox.com	bibliothekderfreien.de	oilsandsrealitycheck.org
madilli.com	anarkismo.net	pnwer.org	communities.canada.com	080a.g759.com
google.ca	streams4justice.org	pedalandplow.com	occupygeepsics.tumblr.com	dspace.library.uvic.ca
victoriaanarchistbookfair.ca	immigrantforsale.org	freealabamamovement.wordpress.com	committedtoendhomelessnessvictoria.wordpress.com	encyclopedia.wikia.com
mail.uvic.ca	dreamwalkerdieries.blogspot.com	pdxcabc.org	thehill.com	discoqs.com
woodlanddeer.blogspot.ca	maximumtolerateddose.org	treehugger.com	419bird110.wordpress.com	indymedia.org.uk
camas.ca	seattletimes.nwsource.com	edgeofsports.com	killcap.org	kk121.g063.info
shac.net	Bump724.com	rainbow.coop	politicsinotabanana.com	sundog.usask.ca
wildmembers.com	network.nationalpost.com	jfofeman.com	davidasuzuki.org	shortbusbook.blogspot.com
facebook.com	gatewayspanel.review-examen.gc.ca	chiv.ca	readingforsanity.blogspot.com	untorelpress.noblogs.org
twitter.com	newmatilda.com	october2011.org	radicalpolytics.org	stopkevinnanett.wordpress.com
en.wikipedia.org	truthinanutshell.wordpress.com	antifascistnetwork.wordpress.com	ia601507.us.archive.org	syracuse.com
cms.paypal.com	solone.org	thepetalia.org	my.good-tw.info	tpccanada.org
youtube.com	antibacconomads.wordpress.com		antipetalia.org	vsw.ca
				en.alexexperts.com
				blackamericaweb.com

### Camas Books & Infoshop

2620 Quadra Street - Victoria - BC  
Traditional Lekwungen Territory

WELCOME CALENDAR ABOUT US GET INVOLVED CONTACTS

We are open every day: [Check the calendar for events.](#)

PLEASE DONATE

Single Donation

Donate

Recurring Donations

Supporter: \$10.00 CAD - mont

Subscribe

#### Take the Rap

Ann Hansen on the Prison Industrial Complex

**Take the Rap**  
Women Doing Time for Society's Crimes  
Dr. Hansen is author of *Take the Rap*, a collection of essays on the prison industrial complex and the role of women in the justice system.

**Saturday, September 22**  
**Book Launch**  
6:00 pm  
Come! Book: 2620 Quadra  
Unceded Lekwungen Territory

#### Camas 10th Anniversary Celebration Fundraiser

NOVEMBER 2ND, 7-10PM  
HOSTED AT CAMAS INSTITUTE ON KINGS

#### 50% Off All "Critical Thinker" Books at Camas!

Four more boxes of books have been donated to us from the PACTAC and C-Theory archives of Dr. Arthur Kroker.

All used books in the "Critical Thinker" category will now be

November 2018

S	M	T	W	T	F	S
28	29	30	31	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	1

#### OTHER PLACES TO VISIT

- Warrior Publications
- PM Press
- VIC Forest Action Network
- Anarchy Radio
- The Talon Conspiracy
- Unist'ot'en Camp
- Recyclistas
- Citizens' Counselling Centre
- Village Muse Books
- Victoria Makerspace
- People's Apothecary
- SubMedia.TV
- Black Coffee Co-Op

# Project 2

Chose some seeds from  
University of Victoria's:  
**Anarchist Archives**

Title: 325

URL: <http://325.nostate.net/>

Captured 4 times between Nov 3, 2016 and Nov 3, 2016

Title: 38 Blood Alley Square | the anarchist space menacing blood alley

URL: <http://38bloodalley.wordpress.com/>

Captured 5 times between Jul 8, 2014 and Nov 3, 2016

Title: Halifax Anarchist Black Cross

URL: <http://abc.h-a-z.org/>

Captured 4 times between Jul 8, 2014 and Nov 3, 2016

Title: Calgary Anarchist Bookfair

URL: <http://calgaryanarchistbookfair.blogspot.ca/>

Captured 4 times between Jul 8, 2014 and Nov 3, 2016

Title: Camas Books & Infoshop

URL: <http://camas.ca/>

Captured 5 times between Jul 8, 2014 and Nov 3, 2016

Title: Edmonton Anarchist Bookfair

URL: <http://eabf.ca/>

Captured once on Jul 8, 2014

# Brenda

Wrote some code. We will show you that and then the interesting results we got! (if you want the code let us know)

# Activates a headless Chrome browser to take screenshots of a list of seeds

```
import os
import sys
from itertools import izip
from PIL import Image

url_list = sys.argv[1]

with open(url_list) as f:
    urls = f.readlines()
# you may also want to remove whitespace characters like
# '\n' at the end of each line
urls = [x.strip() for x in urls]

#command to use cutycapt to take screenshot
#cutycapt --url=google.com --out=google.png
```

```
#command for xserver
#xvfb-run --server-args="-screen 0, 1024x768x24" cutycapt
--url=... --out=...

print "Taking screenshots"
cnt=1
for url in urls:
    #command = "xvfb-run --server-args="-screen 0,
    1024x768x24\" cutycapt --url="+url+" --out="+url+".jpg"
    clean_url = url.strip("/")
    command = "/usr/bin/google-chrome --headless
    --hide-scrollbars --disable-gpu
    --screenshot=./anarchy_collection_screenshots_no_banner/"
    +str(cnt)+".png --window-size=800,600 "+url
    print command
    os.system(command)
    cnt = cnt+1
```



# Compares pairs of screenshots: the original website and then its archived version

#adapted from here:

[https://rosettacode.org/wiki/Percentage\\_difference\\_between\\_images#Python](https://rosettacode.org/wiki/Percentage_difference_between_images#Python)

thon

import os

import sys

from itertools import izip

from PIL import Image

from os import listdir

image\_locations = sys.argv[1]

from os.path import isfile, join

onlyfiles = [f for f in listdir(image\_locations) if

isfile(join(image\_locations, f))]

print onlyfiles

i=1

while i<len(onlyfiles):

    i1 = Image.open(image\_locations+str(i)+".png")

    i2 = Image.open(image\_locations+str(i+1)+".png")

    size\_i1 = i1.size

    size\_i2 = i2.size

    print "Reading images: ",str(i)+".png", str(i+1)+".png"

    #i2.resize(i1.size, Image.ANTIALIAS)

    assert i1.mode == i2.mode, "Different kinds of images."

    assert i1.size == i2.size, "Different sizes."

    pairs = izip(i1.getdata(), i2.getdata())

    if len(i1.getbands()) == 1:

        # for gray-scale jpegs

        dif = sum(abs(p1-p2) for p1,p2 in pairs)

    else:

        dif = sum(abs(c1-c2) for p1,p2 in pairs for c1,c2 in zip(p1,p2))

    ncomponents = i1.size[0] \* i1.size[1] \* 3

    print "Difference (percentage).:", (dif / 255.0 \* 100) / ncomponents

    i=i+2

Collapse

If you want the code talk to  
#seedsofanarchy



# Small degrees of difference: 1-10% page has been updated

Current website



## 325

Contact  
Direct  
Action  
Distro

### Indonesia: 'No (ITS), ¡No Seguirás Adelante!' por Eat

November 1st, 2018

EN: 'No (ITS), You Will Not Go On Forward!' by Eat (Indonesia)

Han pasado tanta mierda de parte de la llamada pandilla Eco-Extremista y de todas sus porristas misántropas, nihilista y aspirantes. Encontré su texto del año pasado, "El mito anarquista", lleno de cristianismo vulgar que esconden como eco-extremismo. Vamos, ¿es esto lo mejor que pueden hacer? ¿Afirmando haber matado a personas sin esperanza en alguna ocasión? Y leyendo sus amenazas de muerte mientras publicas fotos de alguien que asesinaste (según tu reivindicación, por supuesto), es realmente divertido, no, realmente lo es (realmente me río, literalmente cuando leo los comunicados y las fotografías, ¿qué mierda?).

(MÉXICO) 60 COMUNICADO DE ITS: SOBRE LA DELACIÓN DE LA ONG 325

Search

Anti-I

ABC Belar  
ABC Brigh  
ABC Dresc  
ABC Indor  
ABC Wien  
Act for Fre  
Agitasi  
Anarchist  
Anarchy T

Crawled site



## 325

Contact  
Direct  
Action  
Distro

### Anarchist call against the G20 summit In Hamburg (Germany)

November 2nd, 2016

**On the 7th and 8th of July 2017, when the most successful war criminals of the present, the most unscrupulous sweaters of human and nature, the self-titled leaders of this planet, meet in Hamburg, they will not be confronted and thus be revaluated with some demands for better governance or social enslavement.**

**They will feel the rage of the street, when they are rushing with their convoys through deserted districts and talk about the nightly attacks of the last few weeks.**

**This call, like so many others at similar meetings, does not want to lose itself in an analysis of the importance of the G-20 summit or the policy of their participants. The injustice of the world has been declared a thousand times, anyone who now feels no urge**

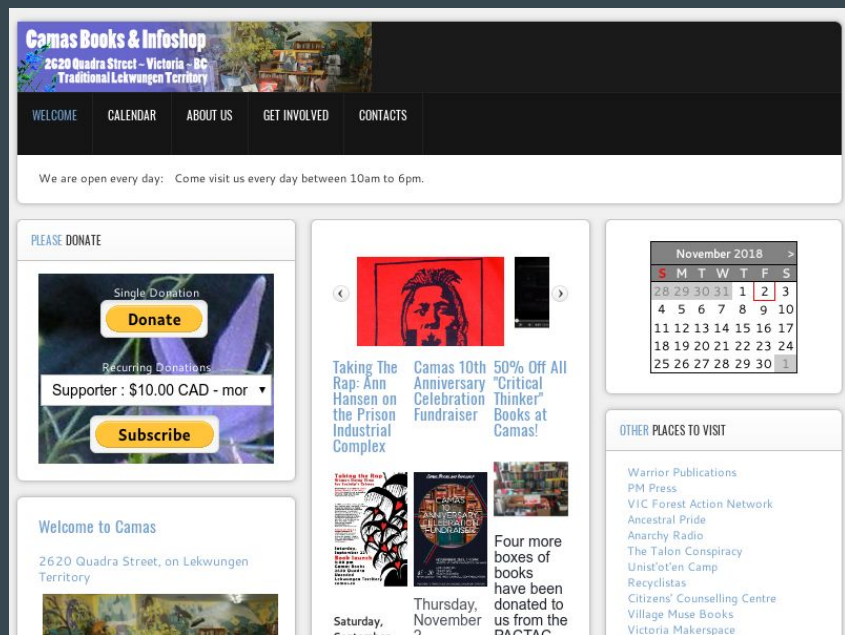
Search

Radio

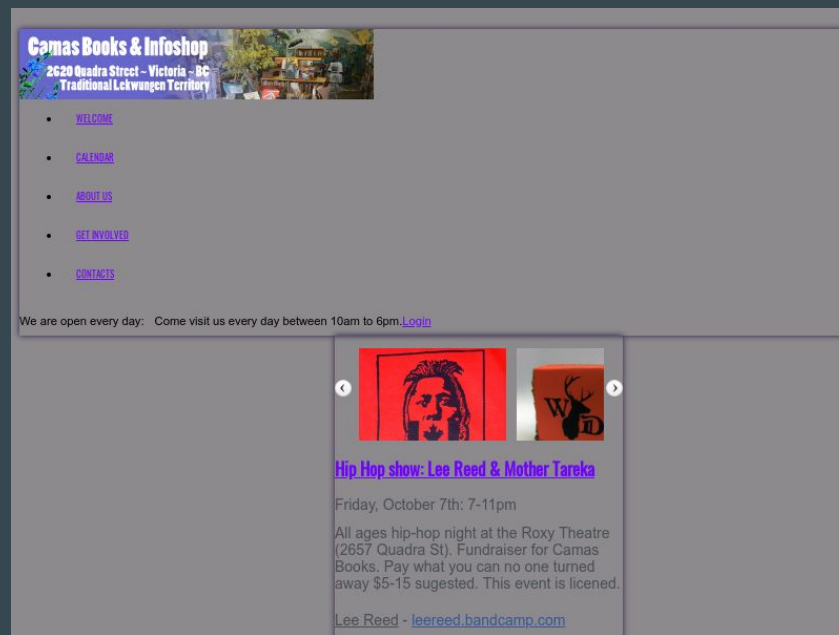
Act for Fre  
Attaque  
Avtonom  
Bite Back  
Cette Sem  
Contra Inf  
Contra Inf  
Contra Ma  
CNA Italia

# Larger degrees of difference: ~40% Quality Problem!

Current website



Crawled website



# Cool thing that Brenda discovered...



[http://wayback.archive-it.org/4594/20161103041607id\\_/https://38bloodalley.wordpress.com/](http://wayback.archive-it.org/4594/20161103041607id_/https://38bloodalley.wordpress.com/)

Add 'id\_' before the 'http' to the Archive-It URL to display without the header (turns out that the header leaves an ~18% difference).

# Results and thoughts....

- Could do more with computational comparison of recent crawl and current website...
- Our results for comparison of live v. crawled screenshots:
  - ~1-10% page has been updated
  - ~20% images on page changed
  - ~40%+ =flag for whoever is in charge of the collection: capture quality issue, page no longer exists, etc.

# Questions

And comments please! How successful were we?

# Nonrecognition of authority!!!

## Anarchy/web archives haiku(s):

Run Archive-It crawls/Digital humanities/Seeds of anarchy

#HackArchives @SFU/data brought us together/curry fries with cheese

collect web archives/next step? there's a toolkit for that/thanks ArchivesUnleashed

# More Haikus...

counting syllables/much like parsing datasets/archivist's happy place

data quality / import aut matchbox / file exists error

seeds of anarchy / history without the state / keepValidPages()

entering paste mode/nothing pasted nothing gained/Spark scala> humour