CS 751: Introduction to Digital Libraries - Assignment 2

Jessica McConnell

March 5, 2015

1 Q1

We were required to pick 100 URIs and get WARC files for them using 4 different methods. The methods were WARCreate, WAIL, webrecorder and wget. We were then supposed to replay some WARC files using webrecorder and the wayback machine.

1.1 Picking the URIs

To get my 100 URIs I wrote a python script 'pickuris.py' that picked a random number. It then checked the size of the previously downloaded body from assignment 1. If the size of the body was more than 20,000 we used it as one of our URIs. I picked this number by picking several different sizes and browsing to them in my web page. At this size, the web pages started looking more intricate.

For each of the URIs I created a directory with 4 subdirectories associated with each of different methods we were supposed to use.

These URIs and associated folder number were placed in the 'urissort' file.

1.2 WGET

I wrote a shell script to perform my wgets on all the URIs. This script is called 'wget-command'

WGET was the fastest method of getting the WARC files.

WGET appeared to be consistently the smallest file type.

SUM: 57950068
LENGTH: 100
MEAN: 579500.68
MEDIAN: 41694.5

1.3 webrecorder

Webrecorder was the second tool I used to get the WARC files. At first I maded the mistake of not erasing the previous file and kept getting WARC files for a playlist of URIs. After figuring that out this tool was very simple.

I was not able to get a WARC file for every URI I attempted to use with this tool. I skipped these URIs for this method.

`http://www.louisianamusicfactory.com/shop/compact-disc/various-artists-modern-brass-b` would not load in webrecorder.

`http://loveofmusicblog.blogspot.com/p/clothing.html` would cause Chrome to crash saying the site was unresponsive although it would load most of the page before crashing.

`https://bhamqmm.files.wordpress.com/2015/01/qmm-winter-2014.pdf` would cause a 'Temporary Warc Error' whenever I tried to download the WARC.

The webrecorder files were about twice the size of the files from wget.

SUM: 176759581
LENGTH: 97
MEAN: 1822263.7216494845
MEDIAN: 1005967

1.4 WARCreate

WARCreate was very difficult to use. It would lock up very often. It would also, seemingly at random times, start failing downloads saying 'Failed - No File'. The only way to resolve either of these issues was to close Chrome, open task manager and kill all processes associated with Chrome. After that I could start downloading WARCs again.

I was also not able to get a WARC for every URI I tried with this method. In these instances of failure I would click ‘Generate WARC’ wait for several minutes and nothing ever downloaded. These URIs were:

https://bhamqmm.files.wordpress.com/2015/01/qmm-winter-2014.pdf
https://www.etsy.com/shop/stationeryCiaffi?utm_source=Twitter&utm_medium=PageTools&utm_campaign=Share
http://www.careersinconstruction.com/job/842929/quantity-surveyor-/?TrackID=17
http://solarsystem.nasa.gov/missions/profile.cfm?Sort=Alpha&Alias=Surveyor%2003&Letter=S&Display=ReadMore
http://dg-di.us/OM/FR/613029.htm
http://www.louisianamusicfactory.com/shop/compact-disc/various-artists-modern-brass-banc
http://www.iol.co.za/motoring/f1-grand-prix/ferrari-reveals-new-f1-race-car-1.1811434?utm_medium=twitter&utm_source=twitterfeed

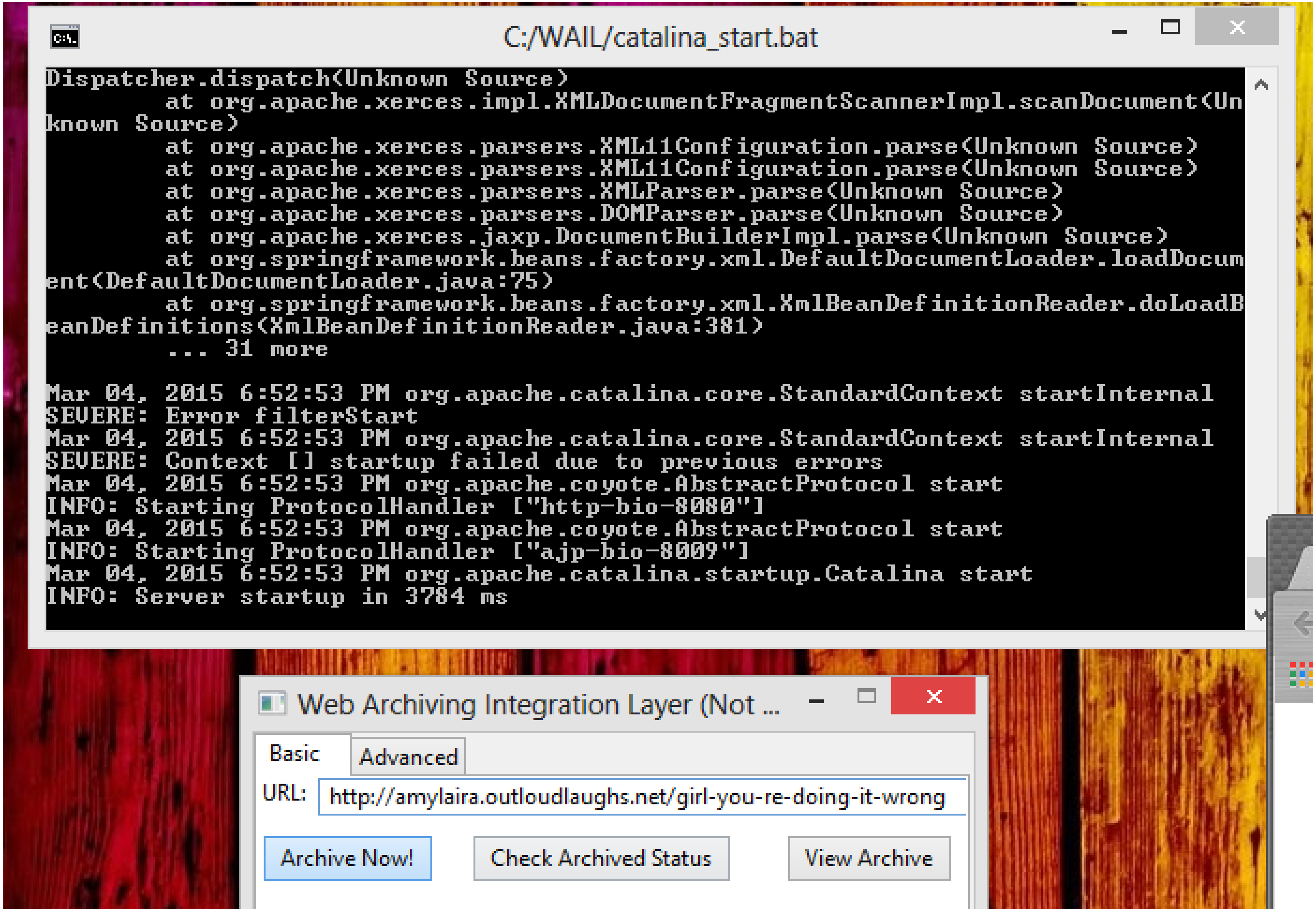
The WARCreate files were by far the largest files. They were about 4 times as large as the webrecorder files and 8 times as large as the wget files.

SUM: 976675632
LENGTH: 93
MEAN: 10501888.516129032
MEDIAN: 9785880

1.5 WAIL

I was unable to get WARC files using WAIL. Everytime I tried to use it the application would lock up and get in a ‘not responding’ state.

Figure 1: Failed WAIL



```
Dispatcher.dispatch(Unknown Source)
        at org.apache.xerces.impl.XMLDocumentFragmentScannerImpl.scanDocument(Un
known Source)
        at org.apache.xerces.parsers.XML11Configuration.parse(Unknown Source)
        at org.apache.xerces.parsers.XML11Configuration.parse(Unknown Source)
        at org.apache.xerces.parsers.XMLParser.parse(Unknown Source)
        at org.apache.xerces.parsers.DOMParser.parse(Unknown Source)
        at org.apache.xerces.jaxp.DocumentBuilderImpl.parse(Unknown Source)
        at org.springframework.beans.factory.xml.DefaultDocumentLoader.loadDocum
ent(DefaultDocumentLoader.java:75)
        at org.springframework.beans.factory.xml.XmlBeanDefinitionReader.doLoadB
eanDefinitions(XmlBeanDefinitionReader.java:381)
        ... 31 more

Mar 04, 2015 6:52:53 PM org.apache.catalina.core.StandardContext startInternal
SEVERE: Error filterStart
Mar 04, 2015 6:52:53 PM org.apache.catalina.core.StandardContext startInternal
SEVERE: Context [] startup failed due to previous errors
Mar 04, 2015 6:52:53 PM org.apache.coyote.AbstractProtocol start
INFO: Starting ProtocolHandler ["http-bio-8080"]
Mar 04, 2015 6:52:53 PM org.apache.coyote.AbstractProtocol start
INFO: Starting ProtocolHandler ["ajp-bio-8009"]
Mar 04, 2015 6:52:53 PM org.apache.catalina.startup.Catalina start
INFO: Server startup in 3784 ms
```

URL: http://amylaira.outloudlaughs.net/girl-you-re-doing-it-wrong

1.6 Statistics

Figure 2: Mean and Median of the different methods

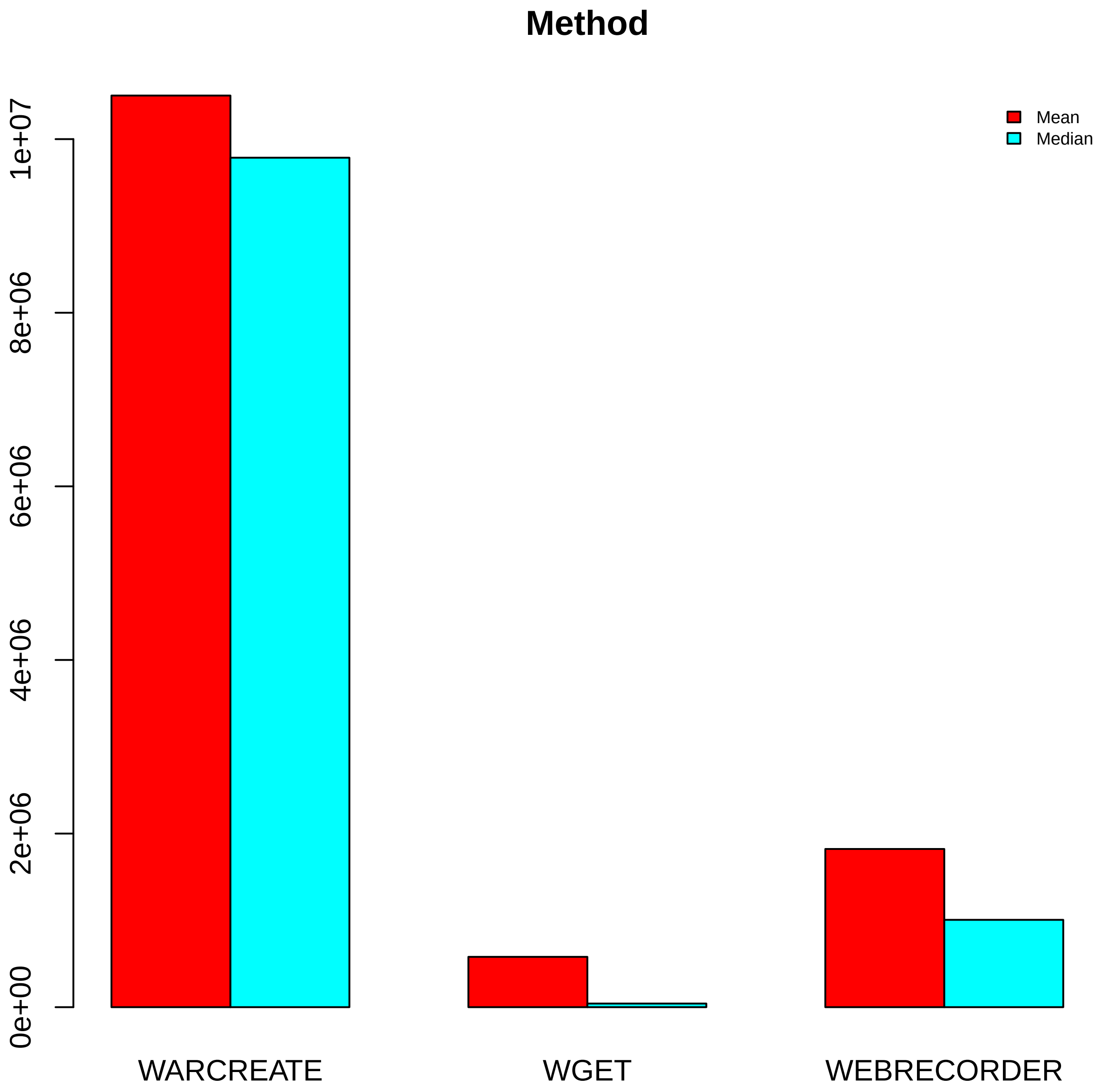
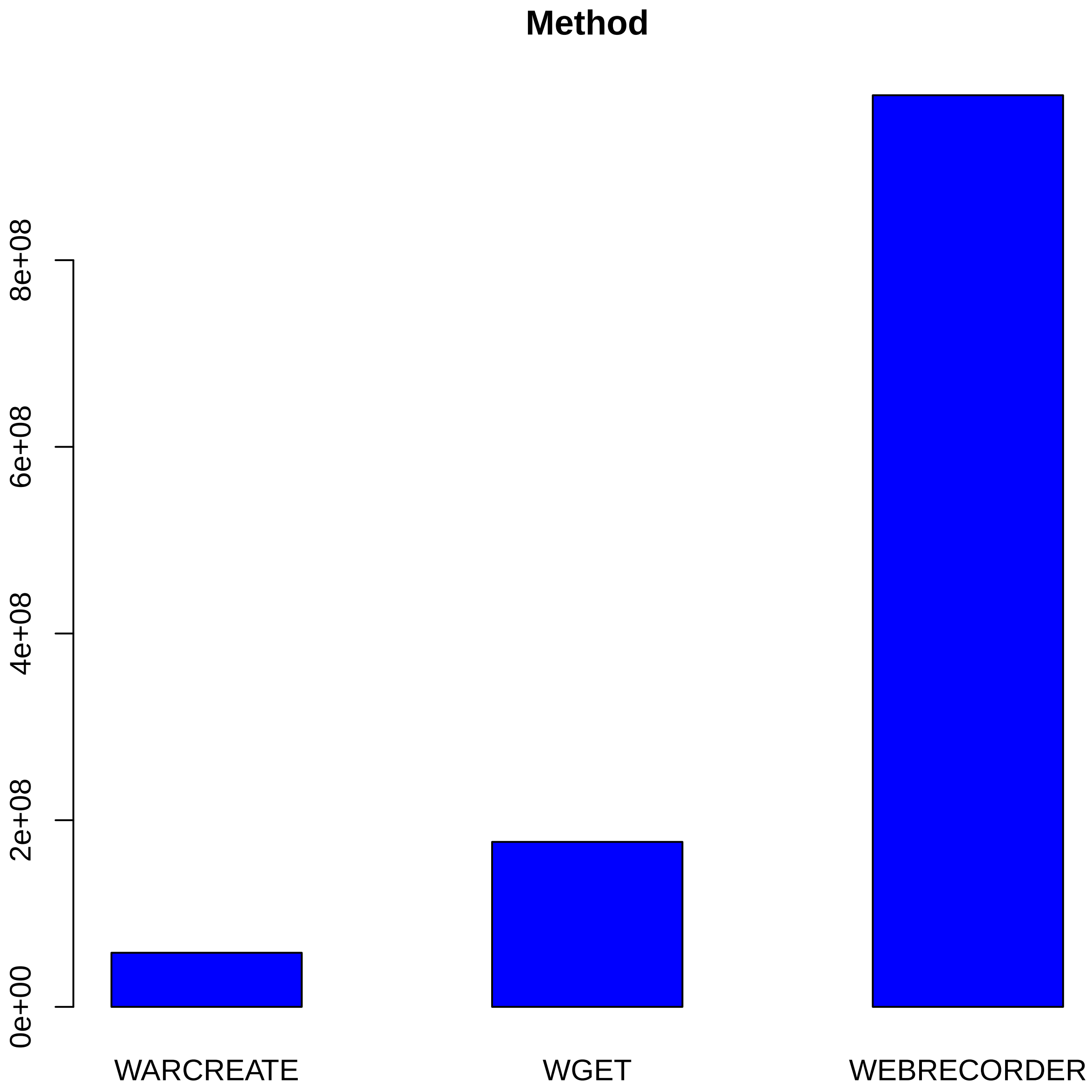


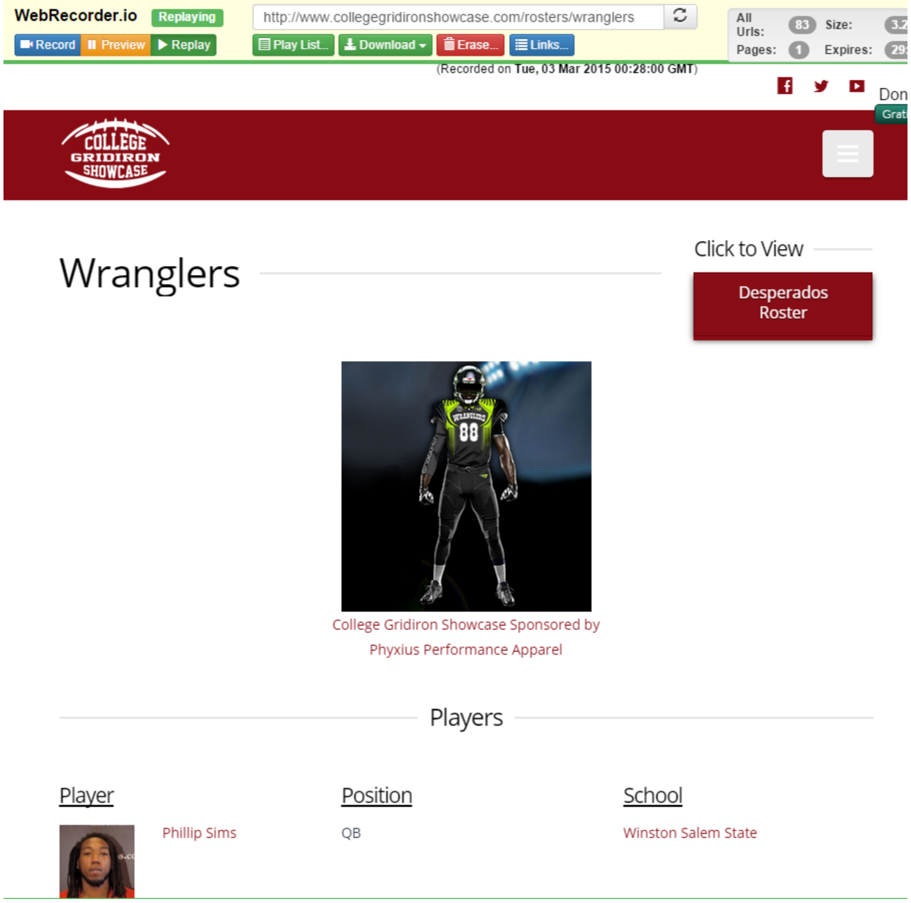
Figure 3: Number of successful WARCs



1.7 Playback

I chose to playback 3 WARC files. The WARCs I chose can be found in the report directory of this folder. In my playback I downloaded a new WARC of the replay. During the replay some images appeared to be smaller, but otherwise everything seemed the same as when I first loaded the page.

Figure 4: URI number 6826



http://www.collegegridironshowcase.com/rosters/wranglers

(Recorded on Tue, 03 Mar 2015 00:28:00 GMT)

Wranglers

Click to View

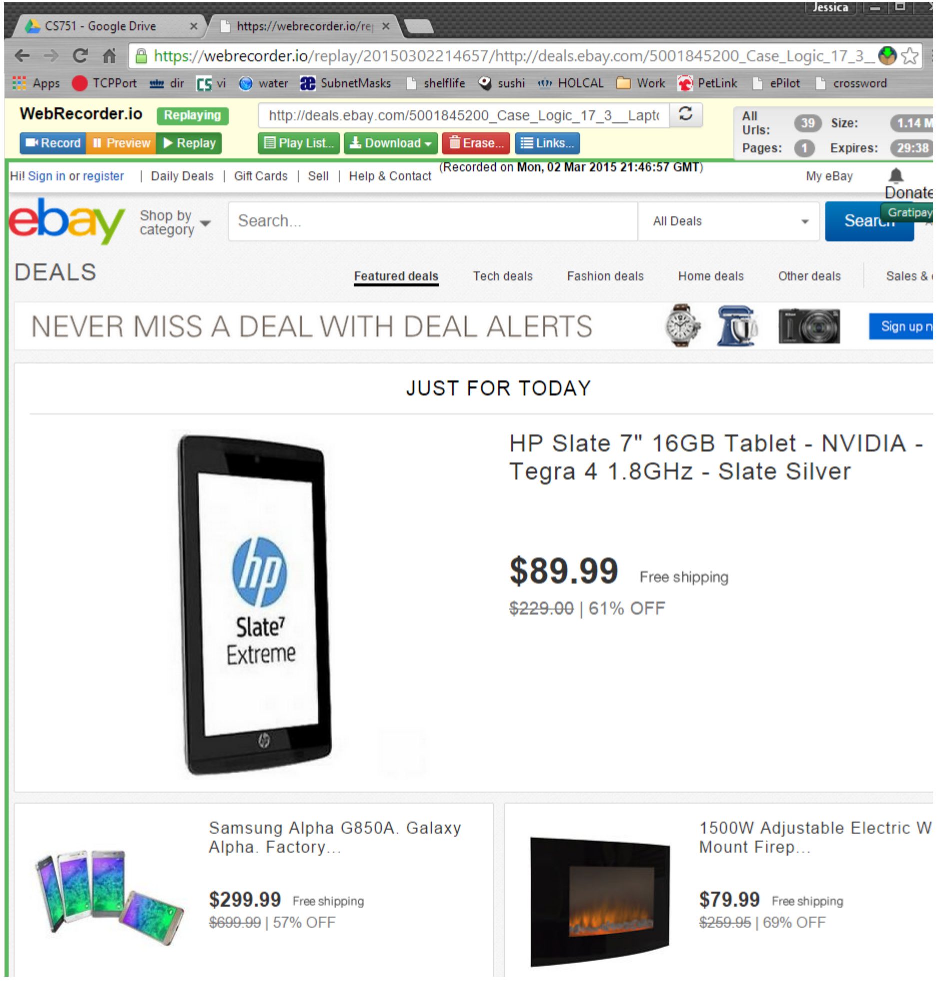
Desperados Roster

College Gridiron Showcase Sponsored by Phyxius Performance Apparel

Players

Player	Position	School
Phillip Sims	QB	Winston Salem State

Figure 5: URI number 1920

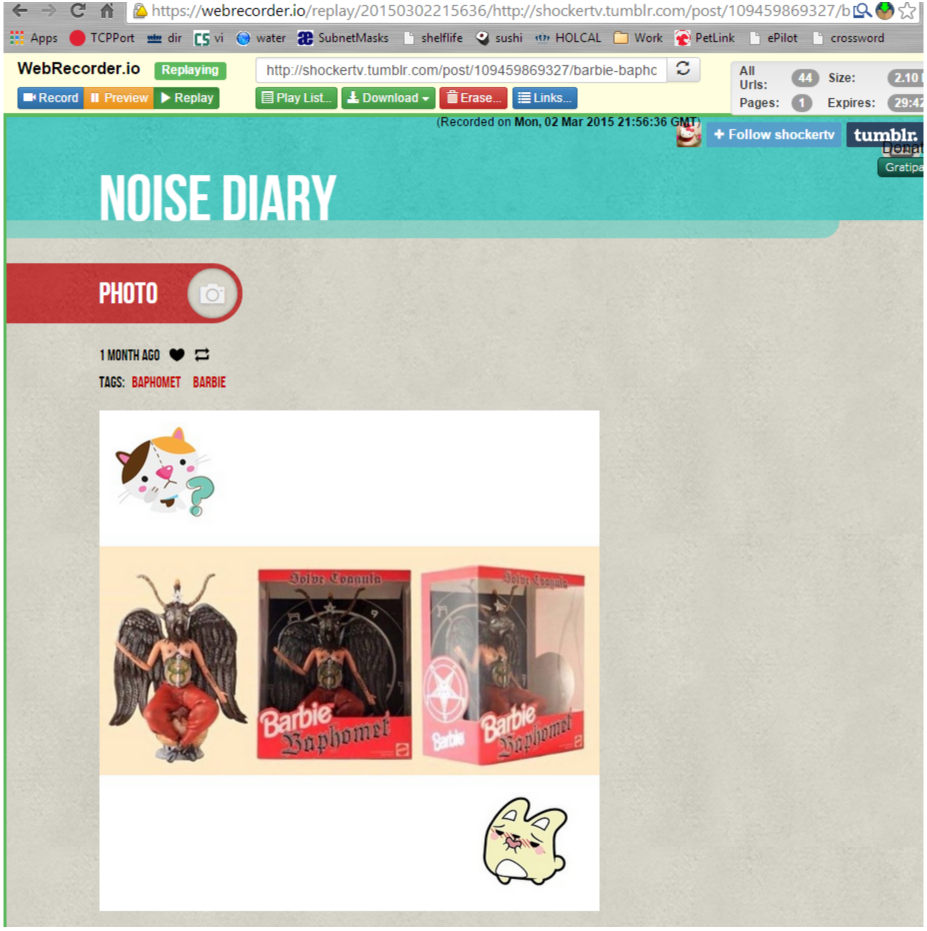


https://webrecorder.io/replay/20150302214657/http://deals.ebay.com/5001845200_Case_Logic_17_3_

http://deals.ebay.com/5001845200_Case_Logic_17_3__Lapt

(Recorded on Mon, 02 Mar 2015 21:46:57 GMT)

Figure 6: URI number 3444



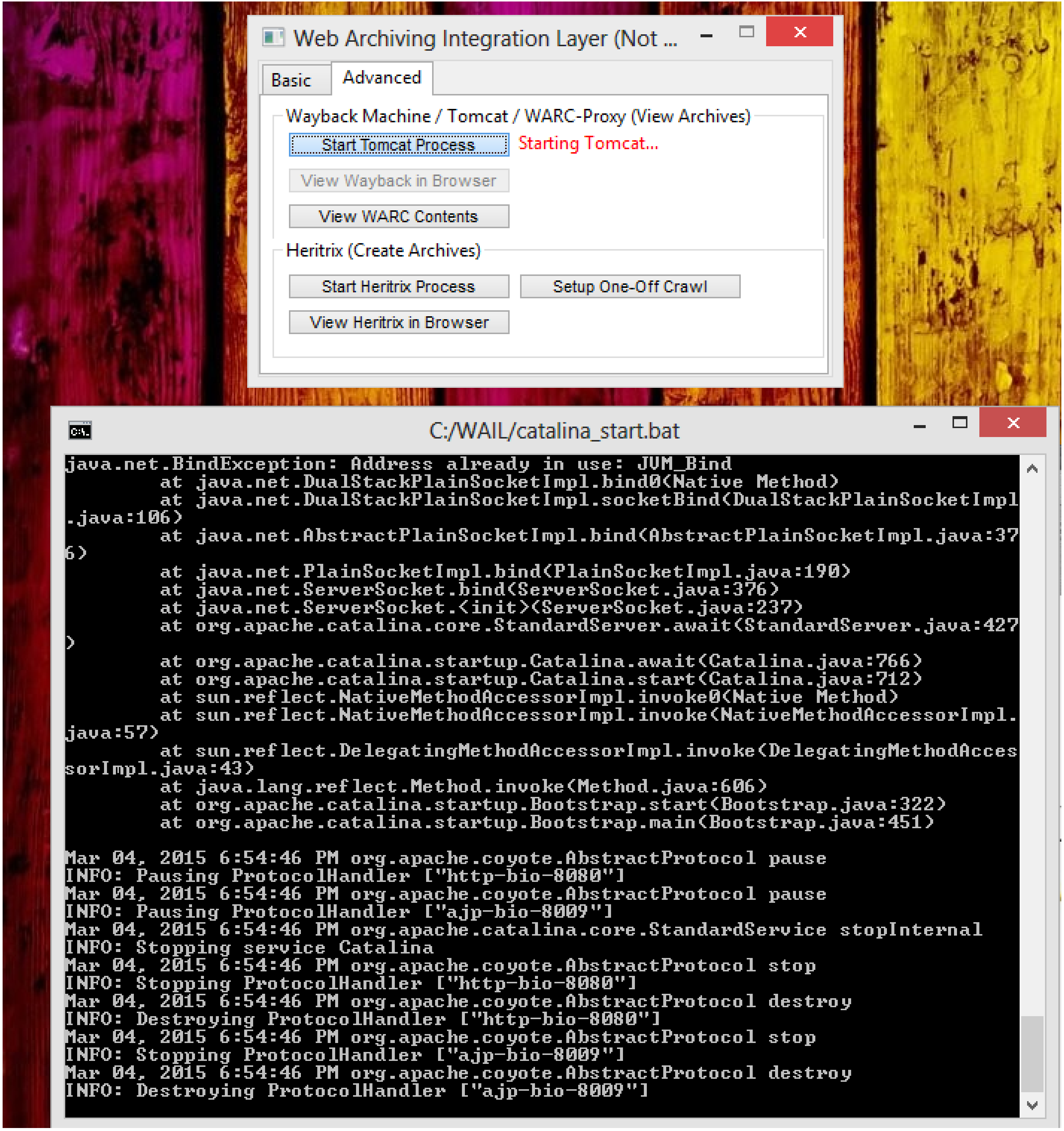
https://webrecorder.io/replay/20150302215636/http://shockertv.tumblr.com/post/109459869327/b

http://shockertv.tumblr.com/post/109459869327/barbie-baphc

(Recorded on Mon, 02 Mar 2015 21:56:36 GMT)

I was unable to replay WARC files using WAIL. Everytime I tried to use it the application would lock up and get in a 'not responding' state.

Figure 7: Failed WAIL Wayback


Web Archiving Integration Layer (Not ...
Basic
Advanced
Wayback Machine / Tomcat / WARC-Proxy (View Archives)
Start Tomcat Process
View Wayback in Browser
View WARC Contents
Heritrix (Create Archives)
Start Heritrix Process
Setup One-Off Crawl
View Heritrix in Browser
C:/WAIL/catalina_start.bat

Starting Tomcat...

```
java.net.BindException: Address already in use: JVM_Bind
        at java.net.DualStackPlainSocketImpl.bind0(Native Method)
        at java.net.DualStackPlainSocketImpl.socketBind(DualStackPlainSocketImpl
.java:106)
        at java.net.AbstractPlainSocketImpl.bind(AbstractPlainSocketImpl.java:37
6)
        at java.net.PlainSocketImpl.bind(PlainSocketImpl.java:190)
        at java.net.ServerSocket.bind(ServerSocket.java:376)
        at java.net.ServerSocket.<init>(ServerSocket.java:237)
        at org.apache.catalina.core.StandardServer.await(StandardServer.java:427
)
        at org.apache.catalina.startup.Catalina.await(Catalina.java:766)
        at org.apache.catalina.startup.Catalina.start(Catalina.java:712)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.
java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAcces
sorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:606)
        at org.apache.catalina.startup.Bootstrap.start(Bootstrap.java:322)
        at org.apache.catalina.startup.Bootstrap.main(Bootstrap.java:451)

Mar 04, 2015 6:54:46 PM org.apache.coyote.AbstractProtocol pause
INFO: Pausing ProtocolHandler ["http-bio-8080"]
Mar 04, 2015 6:54:46 PM org.apache.coyote.AbstractProtocol pause
INFO: Pausing ProtocolHandler ["ajp-bio-8009"]
Mar 04, 2015 6:54:46 PM org.apache.catalina.core.StandardService stopInternal
INFO: Stopping service Catalina
Mar 04, 2015 6:54:46 PM org.apache.coyote.AbstractProtocol stop
INFO: Stopping ProtocolHandler ["http-bio-8080"]
Mar 04, 2015 6:54:46 PM org.apache.coyote.AbstractProtocol destroy
INFO: Destroying ProtocolHandler ["http-bio-8080"]
Mar 04, 2015 6:54:46 PM org.apache.coyote.AbstractProtocol stop
INFO: Stopping ProtocolHandler ["ajp-bio-8009"]
Mar 04, 2015 6:54:46 PM org.apache.coyote.AbstractProtocol destroy
INFO: Destroying ProtocolHandler ["ajp-bio-8009"]
```

2 Q2

For this question we were required to build a SOLR instance, ingest 100 URIs and search.

2.1 Setting up SOLR

I downloaded SOLR from `https://github.com/ukwa/webarchive-discovery` and followed the instructions at its wiki. I was required to also install Maven. I got the binary version from `http://maven.apache.org/download.cgi`.

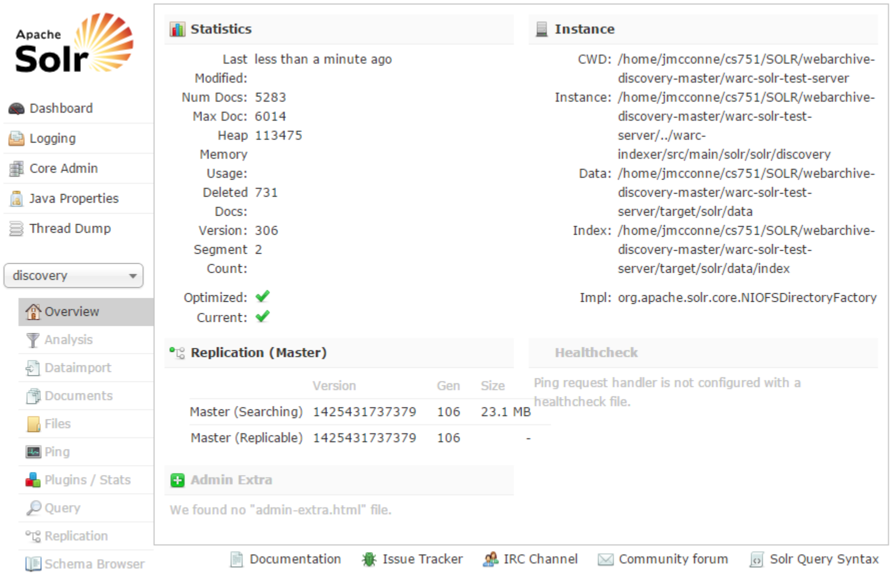
I did not do any special configurations for my SOLR instance. Running the server from linux.cs.odu.edu did require me to use the -Djetty.port option to change the port from 8080 to something between 10000 and 11000.

2.2 Ingesting WARCs

I noticed when ingesting my warc files a few of them hit the word limit and quit ingesting more words from that document.

I chose to use the WARCs received when using WARCreate due to their size. I thought it would give me more words to ingest and search on. Here is my index stats after importing all the WARCs.

Figure 8: Indexes



Statistics

Last Modified: less than a minute ago
Num Docs: 5283
Max Doc: 6014
Heap Memory Usage: 113475
Deleted Docs: 731
Version: 306
Segment Count: 2

Optimized: ✓
Current: ✓

Instance

CWD: /home/jmcconne/cs751/SOLR/webarchive-discovery-master/warc-solr-test-server
Instance: /home/jmcconne/cs751/SOLR/webarchive-discovery-master/warc-solr-test-server/../warc-indexer/src/main/solr/solr/discovery
Data: /home/jmcconne/cs751/SOLR/webarchive-discovery-master/warc-solr-test-server/target/solr/data
Index: /home/jmcconne/cs751/SOLR/webarchive-discovery-master/warc-solr-test-server/target/solr/data/index

Impl: org.apache.solr.core.NIOFSDirectoryFactory

Replication (Master)

	Version	Gen	Size
Master (Searching)	1425431737379	106	23.1 MB
Master (Replicable)	1425431737379	106	-

Healthcheck

Ping request handler is not configured with a healthcheck file.

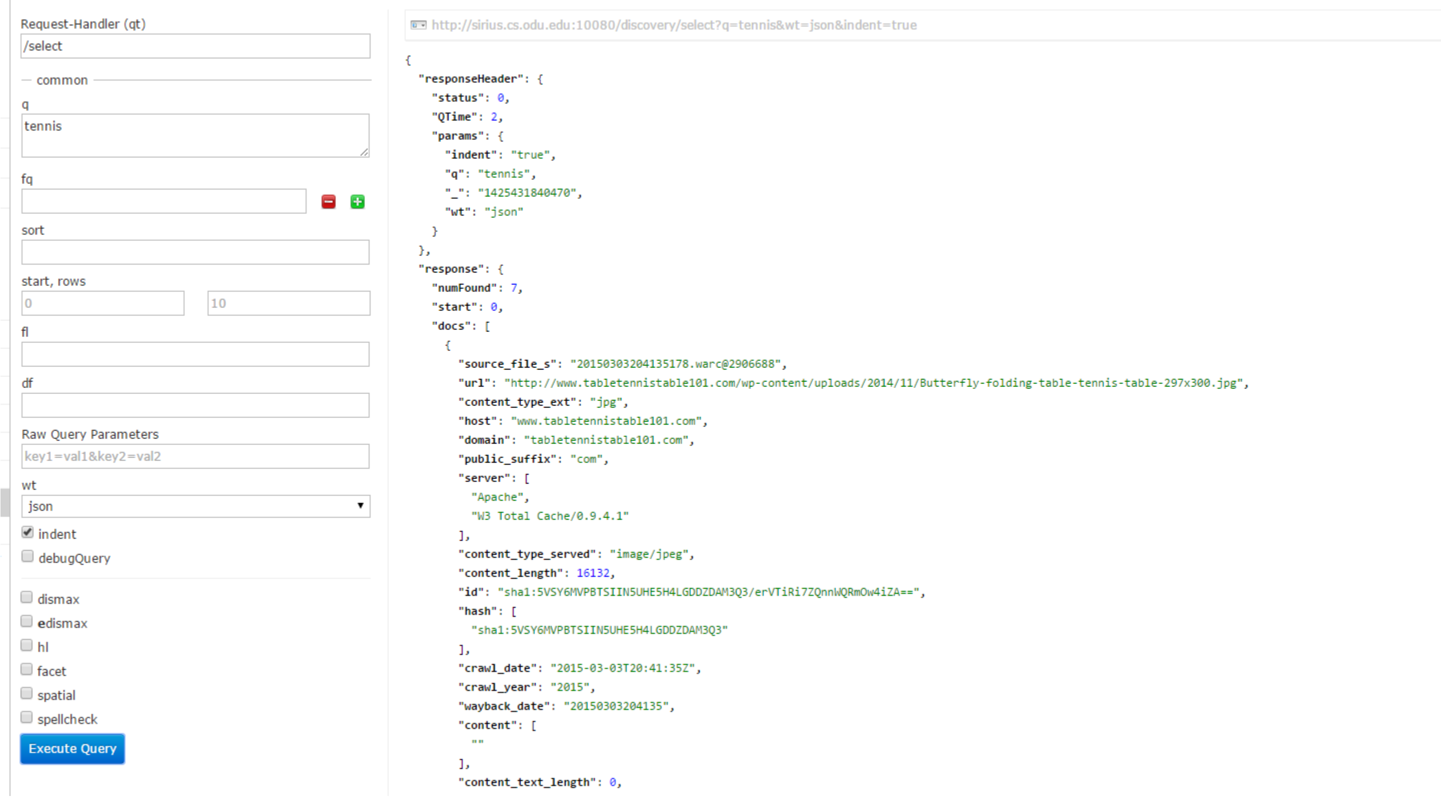
Admin Extra

We found no "admin-extra.html" file.

2.3 Querying

I performed five queries. I was surprised when it worked so well! I knew that I had a WARC file that included a web page that used the word ‘tennis’. To test out querying while still ingesting the WARCs, I searched for ‘tennis’ knowing it was in at least one document.

Figure 9: TENNIS



Request-Handler (qt)
/select

common

q
tennis

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

wt
json

☑ indent
☐ debugQuery

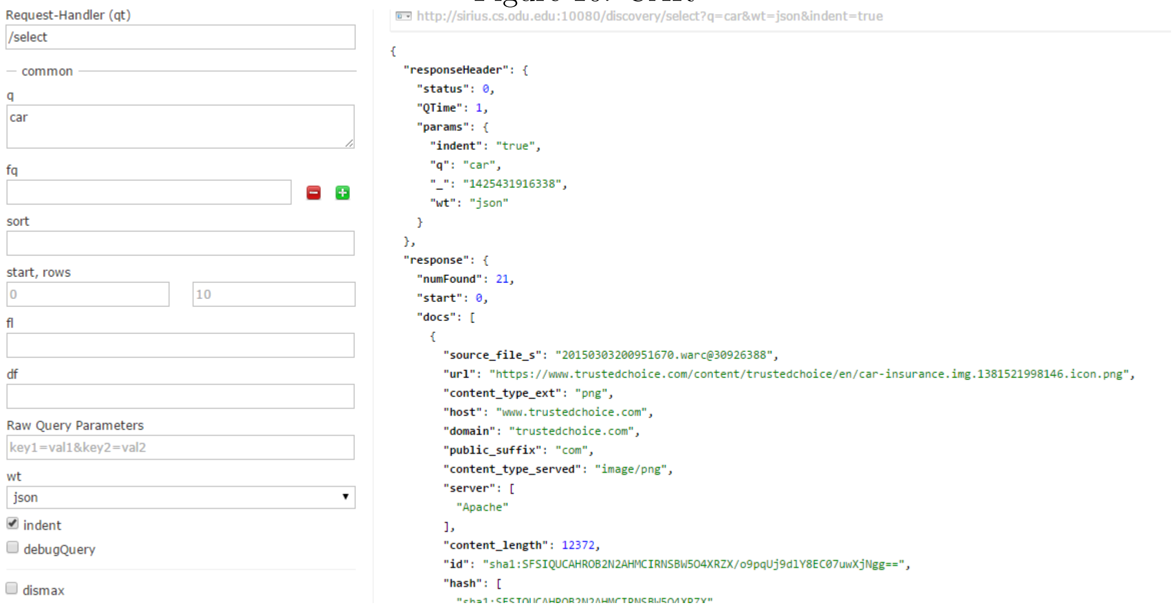
☐ dismax
☐ edismax
☐ hl
☐ facet
☐ spatial
☐ spellcheck

http://sirius.cs.odu.edu:10080/discovery/select?q=tennis&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 2,
    "params": {
      "indent": "true",
      "q": "tennis",
      "_": "1425431840470",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 7,
    "start": 0,
    "docs": [
      {
        "source_file_s": "20150303204135178.warc@2906688",
        "url": "http://www.tabletennistable101.com/wp-content/uploads/2014/11/Butterfly-folding-table-tennis-table-297x300.jpg",
        "content_type_ext": "jpg",
        "host": "www.tabletennistable101.com",
        "domain": "tabletennistable101.com",
        "public_suffix": "com",
        "server": [
          "Apache",
          "W3 Total Cache/0.9.4.1"
        ],
        "content_type_served": "image/jpeg",
        "content_length": 16132,
        "id": "sha1:5VSY6MVPBTSIIN5UHE5H4LGDDZDAM3Q3/erVTiRi7ZQnnWQRmOw4iZA==",
        "hash": [
          "sha1:5VSY6MVPBTSIIN5UHE5H4LGDDZDAM3Q3"
        ],
        "crawl_date": "2015-03-03T20:41:35Z",
        "crawl_year": "2015",
        "wayback_date": "20150303204135",
        "content": [
          ""
        ],
        "content_text_length": 0,
```

After ingesting the rest of my WARC files I began searching for random words.

Figure 10: CAR



Request-Handler (qt)
/select

common

q
car

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

wt
json

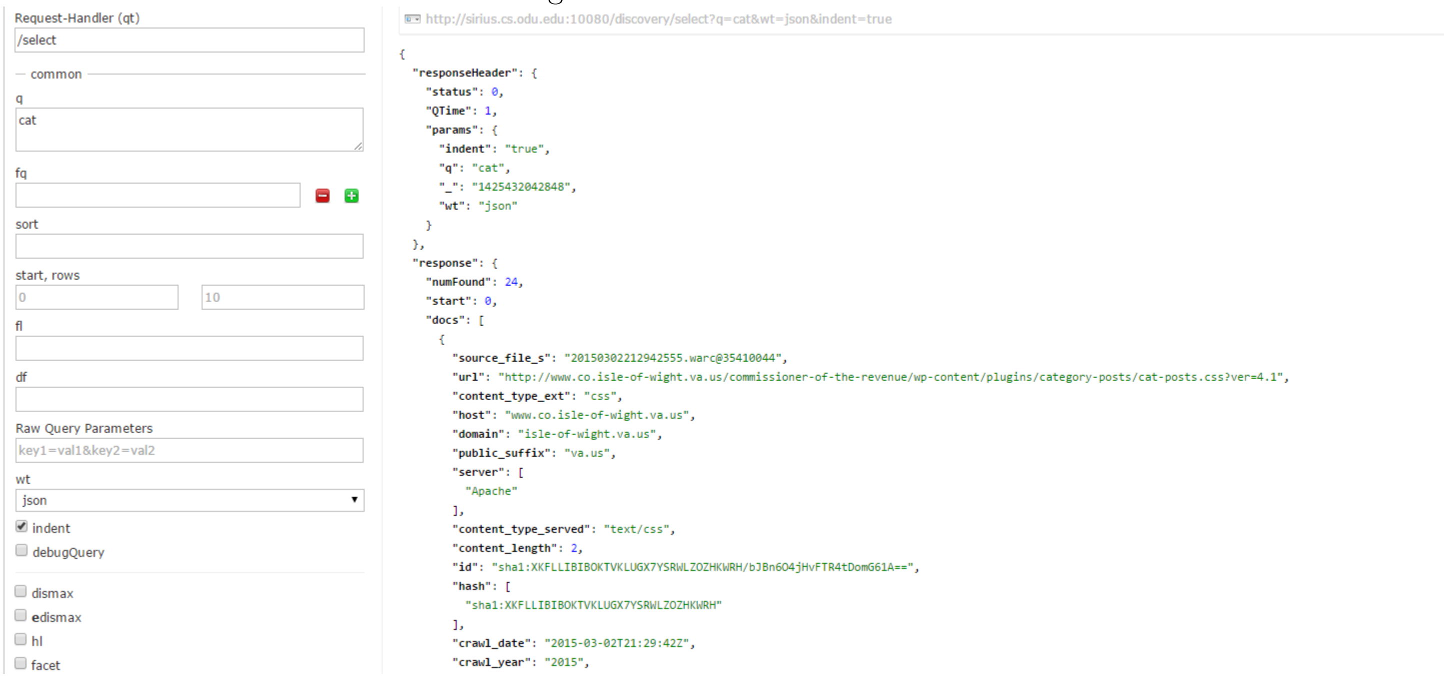
☑ indent
☐ debugQuery

☐ dismax

http://sirius.cs.odu.edu:10080/discovery/select?q=car&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "indent": "true",
      "q": "car",
      "_": "1425431916338",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 21,
    "start": 0,
    "docs": [
      {
        "source_file_s": "20150303200951670.warc@30926388",
        "url": "https://www.trustedchoice.com/content/trustedchoice/en/car-insurance.img.1381521998146.icon.png",
        "content_type_ext": "png",
        "host": "www.trustedchoice.com",
        "domain": "trustedchoice.com",
        "public_suffix": "com",
        "content_type_served": "image/png",
        "server": [
          "Apache"
        ],
        "content_length": 12372,
        "id": "sha1:SFSIQUCAHROB2N2AHMCIRNSBW5O4XRZX/o9pqUj9dlY8EC07uwXjNgg==",
        "hash": [
          "sha1:SFSIQUCAHROB2N2AHMCIRNSBW5O4XRZX"
```

Figure 11: CAT



Request-Handler (qt)
/select

common

q
cat

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

wt
json

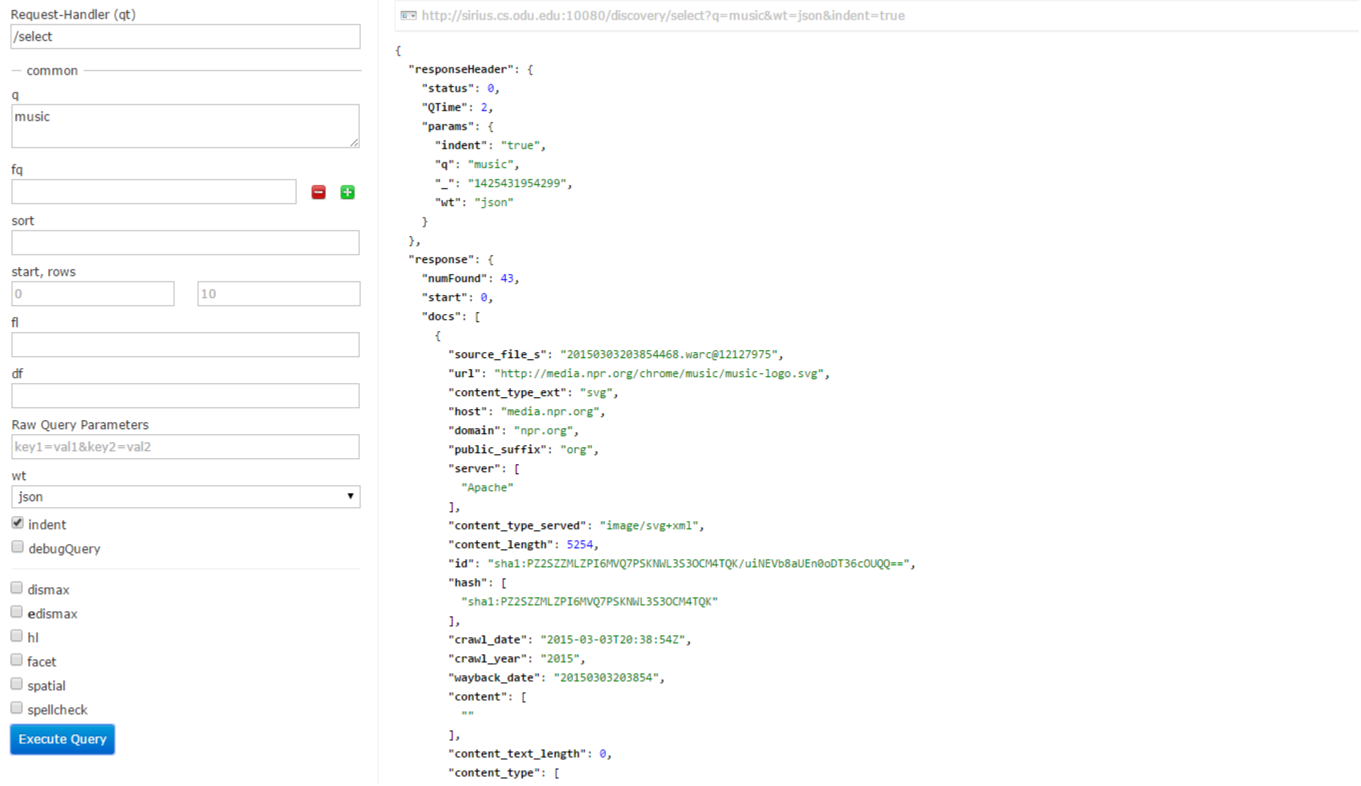
☑ indent
□ debugQuery

□ dismax
□ edismax
□ hl
□ facet

http://sirius.cs.odu.edu:10080/discovery/select?q=cat&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "indent": "true",
      "q": "cat",
      "_": "1425432042848",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 24,
    "start": 0,
    "docs": [
      {
        "source_file_s": "20150302212942555.warc@35410044",
        "url": "http://www.co.isle-of-wight.va.us/commissioner-of-the-revenue/wp-content/plugins/category-posts/cat-posts.css?ver=4.1",
        "content_type_ext": "css",
        "host": "www.co.isle-of-wight.va.us",
        "domain": "isle-of-wight.va.us",
        "public_suffix": "va.us",
        "server": [
          "Apache"
        ],
        "content_type_served": "text/css",
        "content_length": 2,
        "id": "sha1:XKFLLIBIBOKTVKLUGX7YSRWLZOZHKWRH/bJBn6O4jHvFTR4tDomG61A==",
        "hash": [
          "sha1:XKFLLIBIBOKTVKLUGX7YSRWLZOZHKWRH"
        ],
        "crawl_date": "2015-03-02T21:29:42Z",
        "crawl_year": "2015",
```

Figure 12: MUSIC



Request-Handler (qt)
/select

common

q
music

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

wt
json

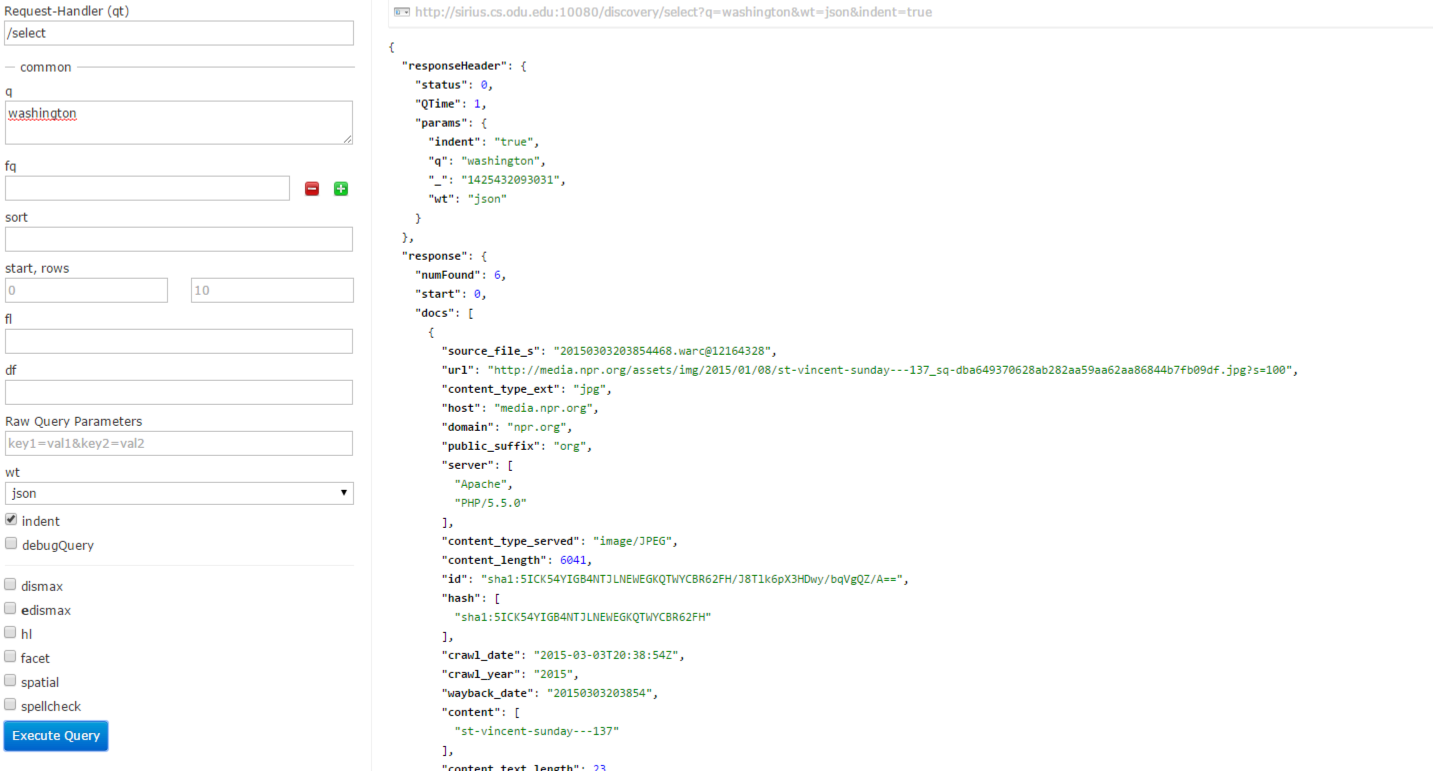
☑ indent
□ debugQuery

□ dismax
□ edismax
□ hl
□ facet
□ spatial
□ spellcheck

http://sirius.cs.odu.edu:10080/discovery/select?q=music&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 2,
    "params": {
      "indent": "true",
      "q": "music",
      "_": "1425431954299",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 43,
    "start": 0,
    "docs": [
      {
        "source_file_s": "20150303203854468.warc@12127975",
        "url": "http://media.npr.org/chrome/music/music-logo.svg",
        "content_type_ext": "svg",
        "host": "media.npr.org",
        "domain": "npr.org",
        "public_suffix": "org",
        "server": [
          "Apache"
        ],
        "content_type_served": "image/svg+xml",
        "content_length": 5254,
        "id": "sha1:PZ2SZZMLZPI6MVQ7PSKNWL3S3OCM4TQK/uiNEVb8aUEn0oDT36cOUQQ==",
        "hash": [
          "sha1:PZ2SZZMLZPI6MVQ7PSKNWL3S3OCM4TQK"
        ],
        "crawl_date": "2015-03-03T20:38:54Z",
        "crawl_year": "2015",
        "wayback_date": "20150303203854",
        "content": [
          ""
        ],
        "content_text_length": 0,
        "content_type": [
```

Figure 13: WASHINGTON



Request-Handler (qt)
/select

common

q
washington

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

wt
json

☑ indent
□ debugQuery

□ dismax
□ edismax
□ hl
□ facet
□ spatial
□ spellcheck

http://sirius.cs.odu.edu:10080/discovery/select?q=washington&wt=json&indent=true

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "indent": "true",
      "q": "washington",
      "_": "1425432093031",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 6,
    "start": 0,
    "docs": [
      {
        "source_file_s": "20150303203854468.warc@12164328",
        "url": "http://media.npr.org/assets/img/2015/01/08/st-vincent-sunday---137_sq-dba649370628ab282aa59aa62aa86844b7fb09df.jpg?s=100",
        "content_type_ext": "jpg",
        "host": "media.npr.org",
        "domain": "npr.org",
        "public_suffix": "org",
        "server": [
          "Apache",
          "PHP/5.5.0"
        ],
        "content_type_served": "image/JPEG",
        "content_length": 6041,
        "id": "sha1:5ICK54YIGB4NTJLNEWEGKQTWYCBR62FH/J8Tlk6pX3HDwy/bqVgQZ/A==",
        "hash": [
          "sha1:5ICK54YIGB4NTJLNEWEGKQTWYCBR62FH"
        ],
        "crawl_date": "2015-03-03T20:38:54Z",
        "crawl_year": "2015",
        "wayback_date": "20150303203854",
        "content": [
          "st-vincent-sunday---137"
        ],
        "content_text_length": 23
```

All of the JSON responses for these queries can be found in the report folder in files named *word*query.txt.