

## Assignment 2

---

Victor Dasari

March 5, 2015

### 1 GENERATE WARC FILES

The first half of the assignment is to generate WARC files using webrecorder.io, WARCreate, WAIL and wget for 100 URIs used in assignment 1.

I've written a shell script which takes a file as an input. This file has 100 URIs in it. Each URI is read and the wget command is used to extract WARC files. The WARC files generated are in .warc format. Webrecorder.io is a website that is used to generate .warc files. Any URL can be recorded using the record button and the replay option can be used to upload WARC files to the webrecorder.io and the archived URL can be viewed.

WARCreate is another plug-in that is used to generate the WARC files. Once this extension is downloaded any website can be archived by hitting the WARCreate extension and WARC files are generated .gz format.

I've downloaded WAIL from Github. Tried to run it by following the instructions on github I couldn't get it running for 2-3 times. I had to download maven and a python extension to run a maven command. I also faced issues when installing maven on windows. Finally, I installed maven and followed the guidelines which popped up a solar instance.

Heritrix has been used via the WAIL interface. This has been downloaded from Mat Kelly's website and has been extracted to C Drive. The WAIL.py is executed to run WAIL. From WAIL, Heritrix has been used. To install WAIL, the instructions posted in the class group have been followed.

Format:

In wget, WARC files can be generated in either compressed or uncompressed format. When webrecorder.io is used WARC files are generated in compressed formats(.gz) When WARCreate is used the format of WARC files is also in compressed format.

Playback of a WARC file using webrecorder:

Create high-quality, verifiable archival recording of the content you browse.  
Download and preserve the content for future use. Upload later to view your archive.  
All free. No sign-up or browser extensions required.


Enter url below to begin recording:

Record

Upload and replay previously recorded web archive file or url:

(WebRecorder.io supports replay of any **WARC** or **ARC** file already hosted online, or uploads of up to **500MB**)


Replay

Choose File... or  Choose from Dropbox

(You may also download the new desktop [Web Archive Player](#) to browse your archives offline)

## About WebRecorder

For any questions, comments, inquiries or feature requests,  
contact: [info@webrecorder.io](mailto:info@webrecorder.io)

Donations graciously accepted: 

https://webrecorder.io/replay/

WebRecorder.io REPLAY Expires In: 29:12

Note: Some or all of this archived data was not created in WebRecorder.io and its authenticity can not be verified by WebRecorder.io

Recorded Pages [Url Search](#) Total Archive Size: 240.77 KB

Below is WebRecorder.io best-guess on which urls are actual pages (up to 500 pages per archive). Some pages may not have been detected. Use the *Url Search* tab to lookup specific urls that may not be listed here.

Search:

Showing 1 to 1 of 1 entries

Page	Recorded At
<a href="http://www.weather.com/weather/today//06102%0D">http://www.weather.com/weather/today//06102%0D</a>	3/2/2015, 7:17:38 PM

https://webrecorder.io/replay/20150303001738/http://www.weather.com/weather/today//06102%0D

WebRecorder.io Replaying http://www.weather.com/weather/today//06102

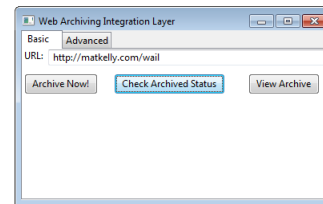
Record Preview Replay Play List... Download... Erase... Links...

All Urls: 2 Size: 240.77 KB Pages: 1 Expires: 29:48

Skip to main content

- [Home](#)
- [Forecasts](#)
  - [Today's Forecast](#)
  - [National Forecast](#)
  - [Alerts](#)
  - [Severe Weather](#)
  - [Safety & Preparedness](#)
  - [Winter Storm Central](#)
  - [Commuter Forecast](#)
  - [NEW Allergy Tracker](#)
  - [NEW Pollen Forecast](#)
- [Winter Storm Thor: Widespread Wintry Mess](#)
- [Taste of Spring in the South](#)
- [More Record Cold Ahead](#)
- [Nearing a Record Snow Season](#)
- [Maps](#)
  - [Weather In Motion®](#)
  - [Severe Alerts](#)
  - [Current Temps](#)
  - [Regional Satellite](#)
  - [World Satellite](#)
  - [Commuter Forecast](#)
  - [Forecast Maps](#)
  - [Classic Weather Maps](#)
- [Current Weather](#)
- [Health Maps](#)

Playback using wayback machine via WAIL:



## WAIL Web Archiving Integration Layer One-Click User Instigated Preservation

### What Is It?

Web Archiving Integration Layer (WAIL) is a graphical user interface (GUI) atop multiple web archiving tools intended to be used as an easy way for anyone to preserve and replay web pages.

Tools included and accessible through the GUI are [Heritrix 3.1.2](#), [Wayback 1.7](#), and [marg-proxy](#). Support packages include [Apache Tomcat](#), [phantomjs](#) and [pyinstaller](#).

WAIL is written mostly in Python and a small amount of JavaScript.

### creenshots



## 2 SOLR INSTANCE

To fire up a solr instance, in one terminal a solr server is started and in another terminal a java command is run along with location of .warc file this in turn will index the URL provided. The Solr UI is at <http://localhost:8080/>. To know if a given warc file has been indexed in the overview section of Solr view "statistics". To query on an indexed URL the query section should be selected and a query should be written and execute query should be pressed

to view the result. Warcmerge is used to consolidate all the 100 warc files into one single warc file. All the dependencies for warcmerge are installed and warcmerge.py is run to consolidate all the 100 warc files. This single warc file is fed to solr and it gets indexed. In the below diagram the value of the field wt is set to xml and the results are in xml format and also the value of the field rows is set to 10. In the second Solar UI image the value of wt is set to json and the rows value to 1 and the results are in json format. In the third image the value of facet is set to "true".

The screenshot displays the Apache Solr Admin UI. On the left is a sidebar with navigation links: Dashboard, Logging, Core Admin, Java Properties, Thread Dump, and a 'discovery' dropdown menu. The 'discovery' menu is expanded, showing options like Overview, Analysis, Dataimport, Documents, Files, Ping, Plugins / Stats, Query (which is selected), Replication, and Schema Browser.

The main interface is titled 'Request-Handler (qt)' and shows a query configuration for the '/select' handler. The configuration includes:
 

- q: /select
- fq: (empty)
- sort: (empty)
- start, rows: 0 to 10
- fl: (empty)
- df: (empty)
- Raw Query Parameters: key1=val1&key2=val2
- wt: xml (selected from a dropdown)
- indent: ☒ (checked)
- debugQuery: ☐ (unchecked)
- Other options: dismax, edismax, hl, facet, spatial, spellcheck (all unchecked)

 An 'Execute Query' button is at the bottom left of the configuration panel.

On the right, the URL bar shows the request: `http://localhost:8080/discovery/select?q=%3A%6rows=10&wt=xml&indent=true`. Below the URL, the XML response is displayed, showing a single document with various metadata fields like `source_file_s`, `url`, `content_type_ext`, `id`, `hash`, `crawl_date`, `crawl_year`, `wayback_date`, and `content_type`.



- Dashboard
- Logging
- Core Admin
- Java Properties
- Thread Dump
- discovery
- Overview
- Analysis
- Dataimport
- Documents
- Files
- Ping
- Plugins / Stats
- Query
- Replication
- Schema Browser

Request-Handler (qt)

/select

— common —

q  
\*,\*

fq

sort

start, rows  
0 1

fl

df

Raw Query Parameters  
key1=val1&key2=val2

wt  
json

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

☐ hl

☐ facet

☐ spatial

☐ spellcheck

```
http://localhost:8080/discovery/select?q=%3A*&rows=1&wt=json&indent=true

{
  "responseHeader": {
    "status": 0,
    "QTime": 1,
    "params": {
      "indent": "true",
      "q": "*,*",
      "_": "1425611452344",
      "wt": "json",
      "rows": "1"
    }
  },
  "response": {
    "numFound": 640,
    "start": 0,
    "docs": [
      {
        "source_file_s": "weeder.org-zore.blue-2015030804952.warc.gz@96437329",
        "url": "http://knzmuslim.com/index.php",
        "content_type_ext": "php",
        "url_type": "slashpage",
        "host": "knzmuslim.com",
        "domain": "knzmuslim.com",
        "public_suffix": "com",
        "server": [
          "Apache",
          "PHP/5.4.34"
        ],
        "content_type_served": "text/html",
        "content_length": 2162,
        "id": "sha1:5FK6YPINQZUPZQP7OUVAO3PEFUTGGK7/B011bmI+xtH4EVDS839VGw==",
        "hash": [
          "sha1:5FK6YPINQZUPZQP7OUVAO3PEFUTGGK7"
        ],
        "crawl_date": "2015-03-02T23:59:55Z",

```



- Dashboard
- Logging
- Core Admin
- Java Properties
- Thread Dump
- discovery
- Overview
- Analysis
- Dataimport
- Documents
- Files
- Ping
- Plugins / Stats
- Query
- Replication
- Schema Browser

Request-Handler (qt)

/select

— common —

q  
\*,\*

fq

sort

start, rows  
0 10

fl

df

Raw Query Parameters  
key1=val1&key2=val2

wt  
json

☒ indent

☐ debugQuery

☐ dismax

☐ edismax

☐ hl

☒ facet

facet.query  
true

facet.field

```
http://localhost:8080/discovery/select?q=%3A*&wt=json&indent=true&facet=true&facet.query=true

{
  "responseHeader": {
    "status": 0,
    "QTime": 57,
    "params": {
      "facet": "true",
      "indent": "true",
      "facet.query": "true",
      "q": "*,*",
      "_": "1425616771601",
      "wt": "json"
    }
  },
  "response": {
    "numFound": 640,
    "start": 0,
    "docs": [
      {
        "source_file_s": "weeder.org-zore.blue-2015030804952.warc.gz@96437329",
        "url": "http://knzmuslim.com/index.php",
        "content_type_ext": "php",
        "url_type": "slashpage",
        "host": "knzmuslim.com",
        "domain": "knzmuslim.com",
        "public_suffix": "com",
        "server": [
          "Apache",
          "PHP/5.4.34"
        ],
        "content_type_served": "text/html",
        "content_length": 2162,
        "id": "sha1:5FK6YPINQZUPZQP7OUVAO3PEFUTGGK7/B011bmI+xtH4EVDS839VGw==",
        "hash": [
          "sha1:5FK6YPINQZUPZQP7OUVAO3PEFUTGGK7"
        ],
        "crawl_date": "2015-03-02T23:59:55Z",
        "crawl_year": "2015",
        "wayback_date": "20150302235955",
        "content_type": {

```