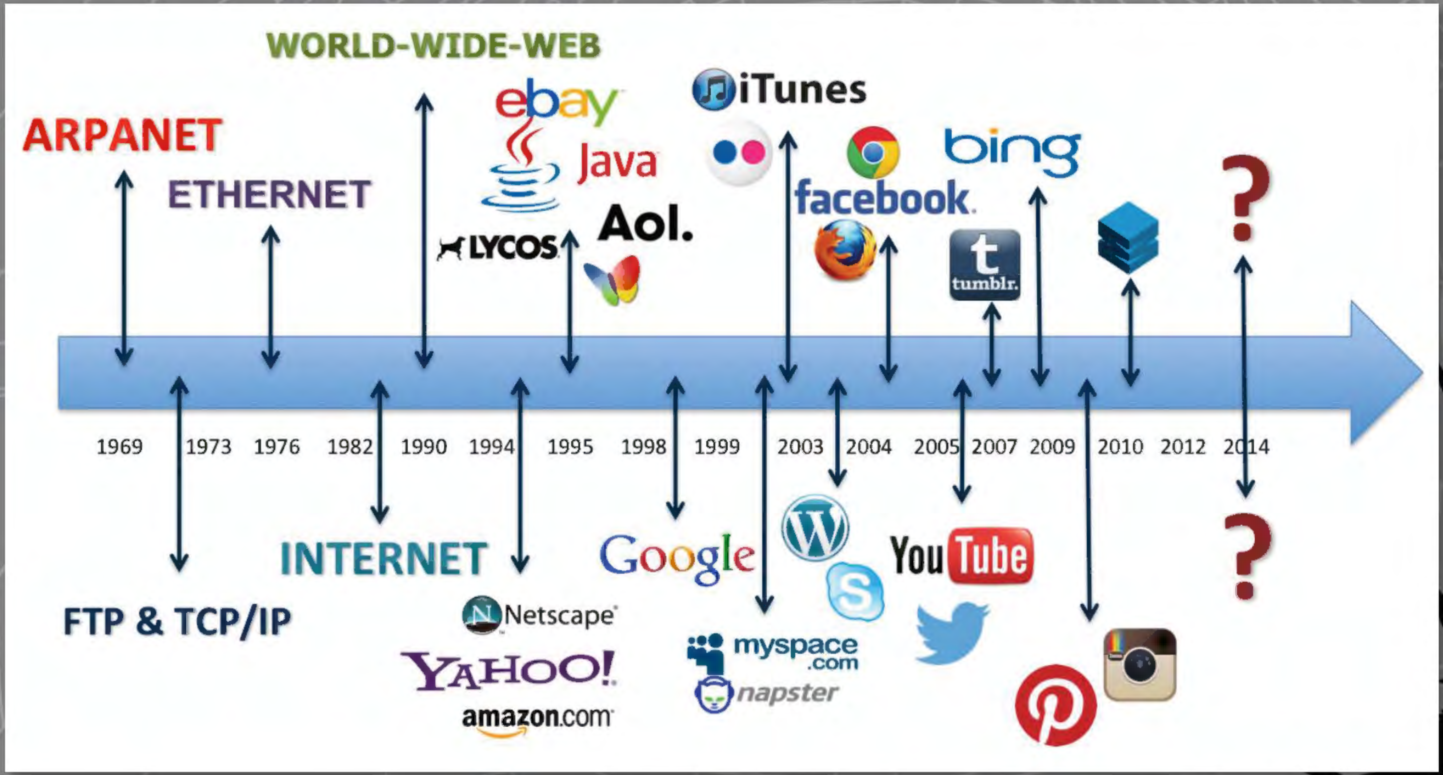
Web Crawlers and Web Archiving

Caroline Z. Oliveira
Digital Preservation
Fall 2017

The Web



ARPANET

ETHERNET

WORLD-WIDE-WEB

LYCOS

ebay

Java

Aol.

iTunes

facebook

bing

tumblr.

?

1969 1973 1976 1982 1990 1994 1995 1998 1999 2003 2004 2005 2007 2009 2010 2012 2014

FTP & TCP/IP

INTERNET

Netscape

YAHOO!

amazon.com

Google

myspace.com

napster

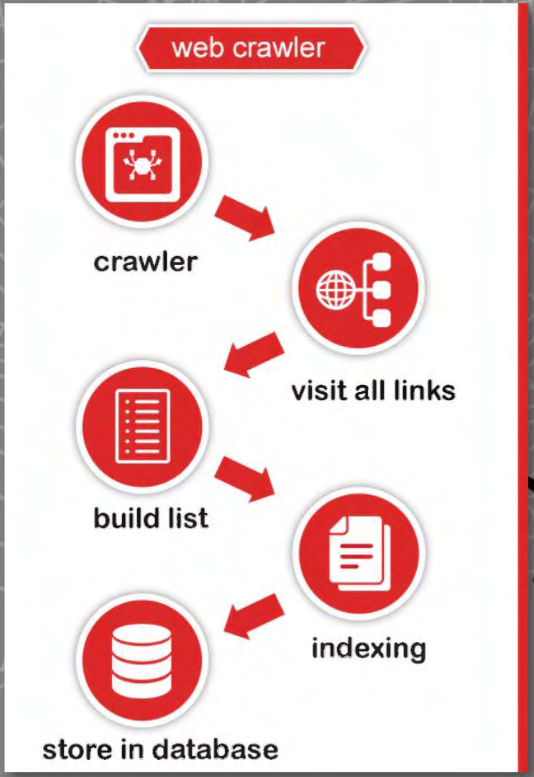
You Tube

?

Crawling or Spidering

- How do crawlers work?

- What are *Seeds*?



web crawler

crawler

visit all links

build list

indexing

store in database

The Internet Archive

First, some statistics:

- 80 % of web pages are not available in their original form after 1 year

- 13 % of web references in scholarly articles disappear after 27 months

- 11 % of social media resources, such as the ones posted on Twitter, are lost after 1 year

What is the Internet Archive?

- Founded by Brewster Kahle, Bruce Gilliat, and other technologists
- Aims to archive content from the Web
- Access
- Alexa Internet and the Wayback Machine

Wayback Machine’s Flaws

- Robots.txt
- Javascript
- Server-side image maps
- Orphan pages

```
User-Agent: *

Disallow: /wp-admin/
Disallow: /wp-includes
Disallow: /wp-content
Disallow: /wp-
Disallow: /*.css$

User-agent: Googlebot-I
Disallow:

User-agent: Mediapartner
Disallow:
```

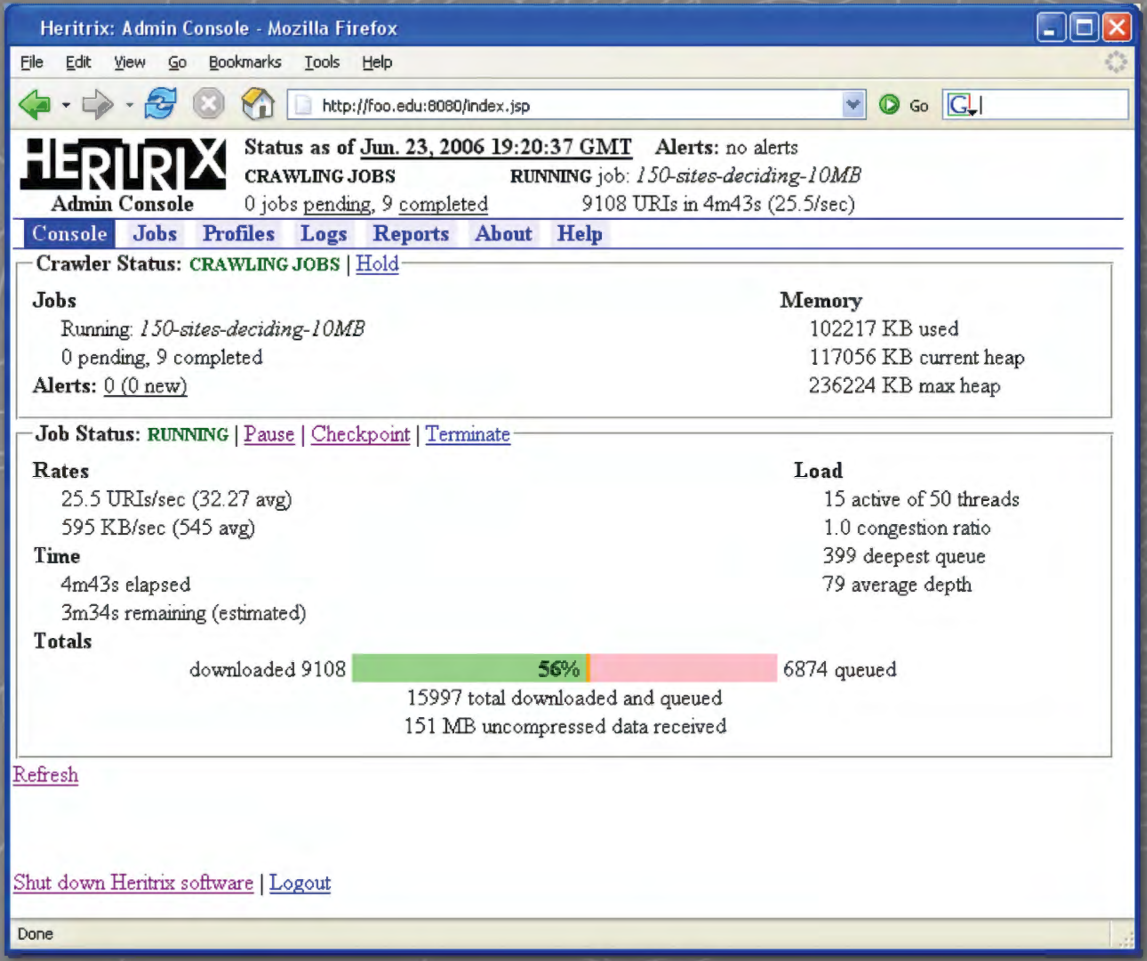
Archive-It

- Subscription-based web archiving service from the Internet Archive
- Helps organizations with their collections
- Over 400 partner organizations in 48 U.S. states and 16 countries worldwide
- Flexible costs and plans

Archive-It

- Orgs can determine the frequency and scope of crawls and manage metadata
- It can handle a wide range of content
- WARC files are stored in data centers
- LOCKSS and DuraCloud

Heritrix


Heritrix: Admin Console - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://foo.edu:8080/index.jsp
Go
HERITRIX
Admin Console
Console Jobs Profiles Logs Reports About Help

Status as of Jun. 23, 2006 19:20:37 GMT **Alerts:** no alerts

CRAWLING JOBS **RUNNING** job: *150-sites-deciding-10MB*

0 jobs pending, 9 completed 9108 URIs in 4m43s (25.5/sec)

Crawler Status: **CRAWLING JOBS** | Hold

Jobs

Running: *150-sites-deciding-10MB*

0 pending, 9 completed

Alerts: 0 (0 new)

Memory

102217 KB used

117056 KB current heap

236224 KB max heap

Job Status: **RUNNING** | Pause | Checkpoint | Terminate

Rates

25.5 URIs/sec (32.27 avg)

595 KB/sec (545 avg)

Time

4m43s elapsed

3m34s remaining (estimated)

Totals

downloaded 9108 56% 6874 queued

15997 total downloaded and queued

151 MB uncompressed data received

Load

15 active of 50 threads

1.0 congestion ratio

399 deepest queue

79 average depth

Refresh

Shut down Heritrix software | Logout

Done

Pros:

- Open-source
- Pluggable and customizable

Cons:

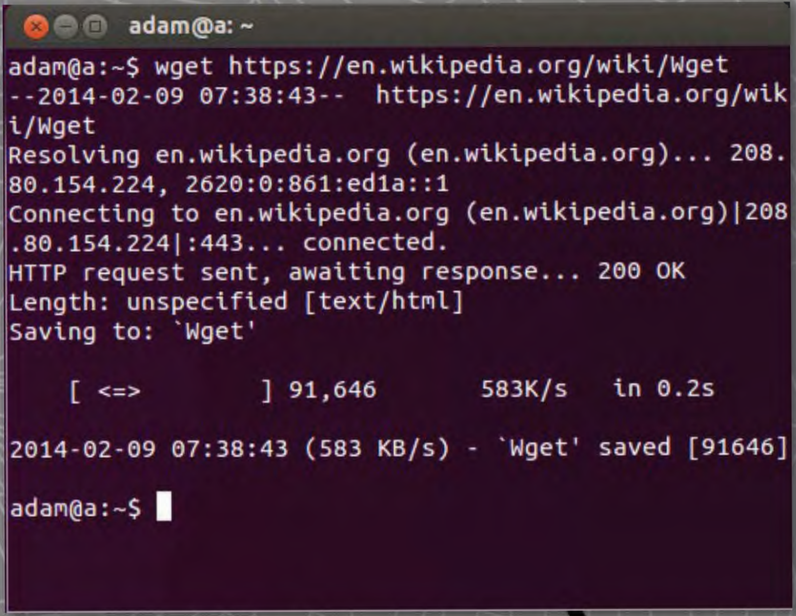
- Does not support continuous crawling
- Not dynamically scalable
- Lacks in flexibility?

A little bit on WARC

- Traditionally used in web crawls
- Four required fields of information
- Problem of scalability

HTTrack and Wget

- Characteristics
- How does it work?
- Limitations

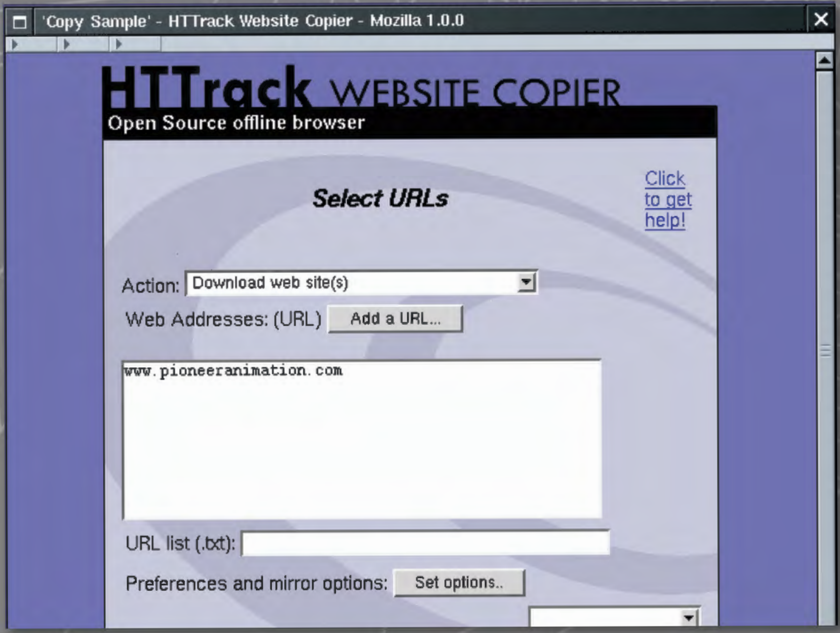


```
adam@a:~$ wget https://en.wikipedia.org/wiki/Wget
--2014-02-09 07:38:43--  https://en.wikipedia.org/wik
i/Wget
Resolving en.wikipedia.org (en.wikipedia.org)... 208.
80.154.224, 2620:0:861:ed1a::1
Connecting to en.wikipedia.org (en.wikipedia.org)|208
.80.154.224|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: `Wget'

    [ <=>          ] 91,646      583K/s   in 0.2s

2014-02-09 07:38:43 (583 KB/s) - `Wget' saved [91646]

adam@a:~$ █
```


'Copy Sample' - HTTrack Website Copier - Mozilla 1.0.0
HTTrack WEBSITE COPIER
Open Source offline browser
Select URLs
Click to get help!
Add a URL...
Set options..

Action: Download web site(s)

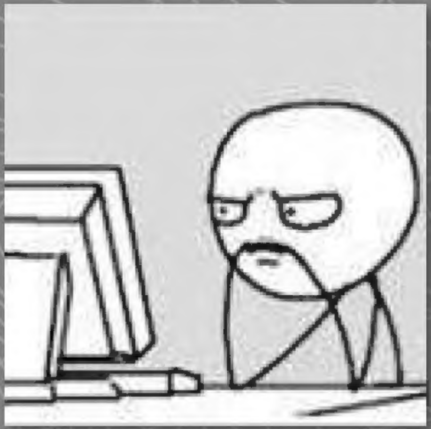
Web Addresses: (URL)

www.pioneeranimation.com

URL list (.txt):

Preferences and mirror options:

Brozzler



- Crawler + Browser
- Aims to solve the two main problems related to capturing a/v materials
- Youtube-dl + Chrome + RethinkDB

Webrecorder

- Released by Rhizome
- Free open-source software
- Human-centered
- Pros and Cons?

Challenges in Web Archiving

- Streaming and Downloadable Media
- Password-Protected Websites
- Form and Database-Driven Content
- Robots.txt Exclusions
- Dynamic Content

(One more time)