Crawling or Spidering

- How do crawlers work?
- What are Seeds?
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
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

Status as of Jun. 23, 2006 19:20:37 GMT  Alerts: no alerts
CRAWLING JOBS  RUNNING job: 150-sites-decoding-10MB
0 jobs pending, 9 completed
9108 URLs in 4m43s (25.5/sec)

<table>
<thead>
<tr>
<th>Console</th>
<th>Jobs</th>
<th>Profiles</th>
<th>Logs</th>
<th>Reports</th>
<th>About</th>
<th>Help</th>
</tr>
</thead>
</table>

**Jobs**
- Running: 150-sites-decoding-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 URLs/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
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)

WARC IT GIRL