Web Archiving in Context and Up Close

May 06, 2011
Digital Library Seminar
Columbia University Libraries
Mellon Project on Web Resource Collection Program Development
Presentation Overview

- How it works...
  1. Crawler: what it is, how it works
  2. Crawler + website, how they interact
  3. Challenges
  4. Products
  5. Reports
  6. Rendering

- Website walk-through
  1. Successfully crawled site/live site
  2. Site with navigation problems
  3. Site with search box
  4. Site with flash
  5. Blogs/linked content
  6. Social media
  7. News aggregators

- Tools/services
  Subscription
  Corporate/E-Discovery/Compliance

- Other web archiving tools
Tools, Resources, Process

Tools & Resources
- Crawler
- Storage for crawler-generated files
- Software to render files
- Access via portal, catalog, index, etc.

Process
- Select
- Test
- Scope
- Crawl
- Review
- Provide Access
Crawlers

Example: Heritrix

What does a crawler do?
Starting from a particular URL, a crawler will gather information about that starting page, while looking for additional URLs. When it finds a new URL, it will jump to each to that URL and repeat the processing of gathering information and looking for additional URLs.
What users see on the live site...

Citizenship Education.

Un nouveau cours est disponible.

Citizenship Education

sera dispensé par la section mauricienne

d'Amnesty International et d'Outremer

le 14 Mai 2011

pour plus de détails, cliquez ici.
What a crawler sees...

dns:amnesty mauritius.org
http://amnesty mauritius.org/
http://amnesty mauritius.org/ccw/downloads/didi1.mp3
http://amnesty mauritius.org/ccw/downloads/didi2.mp3
http://amnesty mauritius.org/ccw/downloads/didienglish.doc
http://amnesty mauritius.org/fovicn.ico
http://amnesty mauritius.org/french/1.2.6
http://amnesty mauritius.org/french/Microsoft.XMLHTTP
http://amnesty mauritius.org/french/application/x-www-form-urlencoded
http://amnesty mauritius.org/french/become.php
http://amnesty mauritius.org/french/becomemada.php
http://amnesty mauritius.org/french/contact.php
http://amnesty mauritius.org/french/downloads.php?cat_id=1
http://amnesty mauritius.org/french/downloads/CONSTITUTION_OF_AMS_AS_ADOPTED_BY_EGM_Feb09SP_RP.pdf
http://amnesty mauritius.org/french/downloads/statuteAmnesty.doc
http://amnesty mauritius.org/french/images/01.jpg
http://amnesty mauritius.org/french/images/02.jpg
http://amnesty mauritius.org/french/images/03.jpg
http://amnesty mauritius.org/french/images/04.jpg
http://amnesty mauritius.org/french/images/05.jpg
http://amnesty mauritius.org/french/images/06.jpg
http://amnesty mauritius.org/french/images/07.jpg
http://amnesty mauritius.org/french/images/08.jpg
WARC Files & Crawl Reports

• WARC file:
  o Web ARChive format. "Used for web-accessible content in archived state, representing the final form disseminated in final state over the web to a user agent (web browser)." (LC Digital Preservation)

• Archive-It Report
  o Summary
  o Hosts
  o Seed status
  o Seed source
  o File types
  o PDFs
  o Videos
Archive-It Crawl Report

**Human Rights**

**Semiannual (ID #20110422185929391)**

<table>
<thead>
<tr>
<th>Summary</th>
<th>Hosts</th>
<th>Seed Status</th>
<th>Seed Source</th>
<th>File Types</th>
<th>PDFs</th>
<th>Videos</th>
<th>QA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Statistics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Started</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Status</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Doc Rate</td>
<td>1.15 urls/sec</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average KB Rate</td>
<td>61.0 KB/s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Documents Crawled*</td>
<td>190,177</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Data Crawled</td>
<td>9.6 GB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Documents Archived</td>
<td>190,177</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New Data Archived*</td>
<td>5.2 GB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Started: April 22, 2011 2:59:32 PM
Completed: April 24, 2011 1:31:28 PM
## Hosts Report (Archive-It)

View only hosts containing [Filter]  [Clear]  

<table>
<thead>
<tr>
<th>Host</th>
<th>URLs</th>
<th>Data</th>
<th>New URLs</th>
<th>New Data</th>
<th>Queued</th>
<th>Robots.txt Blocked</th>
<th>Out of Scope</th>
</tr>
</thead>
<tbody>
<tr>
<td>amnesty.or.jp</td>
<td>95,661</td>
<td>2.4 GB</td>
<td>7,755</td>
<td>198.3 MB</td>
<td>0</td>
<td>0</td>
<td>781</td>
</tr>
<tr>
<td><a href="http://www.amnesty.ca">www.amnesty.ca</a></td>
<td>25,781</td>
<td>1.8 GB</td>
<td>13,359</td>
<td>604.6 MB</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.amnesty.ch">www.amnesty.ch</a></td>
<td>24,987</td>
<td>1.1 GB</td>
<td>12,993</td>
<td>561.8 MB</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>cdhrap.net</td>
<td>19,026</td>
<td>1.3 GB</td>
<td>17,149</td>
<td>1.1 GB</td>
<td>0</td>
<td>0</td>
<td>22,800</td>
</tr>
<tr>
<td>woeser.middle-way.net</td>
<td>9,850</td>
<td>1.4 GB</td>
<td>9,850</td>
<td>1.4 GB</td>
<td>0</td>
<td>81</td>
<td>1,033</td>
</tr>
<tr>
<td><a href="http://www.blogger.com">www.blogger.com</a></td>
<td>3,104</td>
<td>39.8 MB</td>
<td>3,097</td>
<td>39.7 MB</td>
<td>0</td>
<td>0</td>
<td>26,263</td>
</tr>
<tr>
<td>amnesty.ca</td>
<td>2,472</td>
<td>276.0 MB</td>
<td>945</td>
<td>19.3 MB</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2.bp.blogspot.com</td>
<td>1,447</td>
<td>27.6 MB</td>
<td>1,446</td>
<td>27.5 MB</td>
<td>0</td>
<td>0</td>
<td>1,146</td>
</tr>
<tr>
<td>3.bp.blogspot.com</td>
<td>1,437</td>
<td>27.7 MB</td>
<td>1,436</td>
<td>27.7 MB</td>
<td>0</td>
<td>0</td>
<td>1,122</td>
</tr>
<tr>
<td>4.bp.blogspot.com</td>
<td>1,401</td>
<td>27.8 MB</td>
<td>1,400</td>
<td>27.8 MB</td>
<td>0</td>
<td>0</td>
<td>1,110</td>
</tr>
<tr>
<td>1.bp.blogspot.com</td>
<td>1,383</td>
<td>27.1 MB</td>
<td>1,382</td>
<td>27.1 MB</td>
<td>0</td>
<td>0</td>
<td>1,096</td>
</tr>
<tr>
<td><a href="http://www.andalusitas.net">www.andalusitas.net</a></td>
<td>972</td>
<td>28.1 MB</td>
<td>972</td>
<td>28.1 MB</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.ombudsman.org.na">www.ombudsman.org.na</a></td>
<td>588</td>
<td>43.6 MB</td>
<td>588</td>
<td>43.6 MB</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tamaynut.org</td>
<td>346</td>
<td>5.1 MB</td>
<td>346</td>
<td>5.1 MB</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
“Alexa Internet, in cooperation with the Internet Archive, has designed a three dimensional index that allows browsing of web documents over multiple time periods, and turned this unique feature into the Wayback Machine.” (Internet Archive)
Anatomy of a Website

- Public-facing content (what a user is meant to see)
- Underlying content
  - Text
  - Images
  - Databases
  - Scripts
  - External software
- Text/html
- Documents: PDFs, .ppt, .xls, .doc,
- Audio/visual files
  - jpeg, gif, png, mpeg, avi, x-flv, mp4, x-ms-wmx
Challenges

- Dynamic pages
- Flash
- Javascript
- Characters/fonts
  - ex: URLs with Korean characters
- Databases
- News aggregators
- Robots.txt
- Crawler traps
- Social media
Successfully crawled site...

You are viewing an archived web page, collected at the request of Columbia University Libraries using Archive-It. This page was captured on 20:30:25 Mar 02, 2011, and is part of the Human Rights collection. The information on this web page may be out of date. See All versions of this archived page. Metadata

Civil Society Petition: General Assembly should suspend Libya’s UN Human Rights Council membership...

Publications

Towards a Democratic Legislation Supporting the Independence of Non Governmental Organizations (NGOs) | 15/03/2009

Editor: Essam El Din M. Hassan

The significance of the study becomes clear in light of the current moves by the government to amend the present law, something that will probably result in imposing additional restrictions on NGOs. In particular, the official move to amend the law came at a time that is witnessing extensive legislative attacks on liberties, and increasing tendency to oppress the rights of expression and association and other forms of political and social mobility. It is worth noting that the deliberations regarding amending the law came simultaneously as the closing down of two human rights organizations for the first time since the establishment of human rights organizations.
Characters/fonts

**EHREA** Eritrean Human Rights Electronic Archive

1991-2011 | Prisoners | PFDJ crimes | Testimonies | Links

**TESTIMONIES:**

Prisons and Torture Methods

- List of Known Detention Facilities in Eritrea
- The map of terror in Eritrea: A list containing prisons
- Indifference to suffering - unless it's the Palestinians
- Stories from Underground Prison in Eritrea: (Life under PFDJ)
- Owns abroad-groups
- Responsible for

2011/05/06 Fallon, Web Archiving Up Close
Navigation

Malawi Human Rights Commission

We believe that human rights are universal, indivisible, interdependent and inalienable.

The vision of the Commission is to see Malawi have a vibrant human rights culture.

Our mission is to develop and...

Study on Cultural Practices and Human Rights

In May 2006 the Commission released its extensive study on the impact of cultural practices in Malawi.

The study finds that some practices can be very harmful, especially for children, and that many practices impact girls and boys differently.

The Commission calls on all Malawians to protect children's rights and to promote neutral and positive cultural practices. It recommends empowerment of women and a leading role for...
Searching/Databases

You are viewing an archived web page, collected at the request of Columbia University Libraries using Archive-It. This page was captured on 17:17:24 Dec 21, 2010, and is part of the Human Rights collection. The information on this web page may be out of date. See All versions of this archived page.
Flash

The Palestinian Human Rights Monitoring Group

Tax Exemption in the U.S.A

Related Links
Press Release
Past Projects
Flash
Esa mañana, la máquina de regadío estaba parada en medio del terreno y parecía un alcalde de amplias alas atascado bajo el sol. Mis amigos y yo nos metimos en la cabina vacía, tomamos la palanca, los botones, el timón. Saltamos sobre el remolcado asiento y fantaseábamos con que aquel trozo de metal chirriante iba a echar andar y mojaríamos con su riego a todos los estudiantes. Nos reímos por anticipado, pero ni una sola gota salió de los larguísimos tubos que se extendían a ambos lados. Sin embargo, mientras husmeábamos aquí y allí nos tapamos una lata con unas frutas raras. Tenían la forma de un pimiento, pero el color iba del amarillo al anaranjado intenso y una semilla les colgaba por fuera. Jóvenes urbanas, atabadas entre las carencias del racionamiento y el colapso agrícola, no había forma de que superáramos que aquello era un "meraño".

Les hircamos el cliente de inmediato. Dulce, sueve y después, cuando la boca comenzó a resentirse, pensamos que no habíamos envenenado. Corrimos horrorizados, gritando. La carcajada del profesor duró largos minutos. Cuando la sensación astringente pasó, nos quedó el deseo de morir otra vez esa textura que ya había sido cantada en las décadas quebradas, mencionada por nuestros abuelos y pintada por algunos pinzuelos del siglo anterior. Quebr arriesgada por aquella fruta prohibida de nuestro paraíso socialista. Pasarían casi veinte años antes de la volver a encontrar.

La casona, el país

Mayo 3rd, 2011 | 283 comentarios | Imprimir
Robots.txt

http://www.bgcentar.org.rs/robots.txt

User-agent: *
Disallow: /administrator/
Disallow: /cache/
Disallow: /components/
Disallow: /images/
Disallow: /includes/
Disallow: /installation/
Disallow: /language/
Disallow: /libraries/
Disallow: /media/
Disallow: /modules/
Disallow: /plugins/
Disallow: /templates/
Disallow: /tmp/
Disallow: /xmlrpc/
Social media...

Mark Zuckerberg
Public Figure

Wall

Mark Zuckerberg
Check out some of the first grants made from Startup: Education.

5 NJ Schools Get Grants From $100M Facebook Gift
abcn.ws-g/5

5 NJ Schools Get Grants From $100M Facebook Gift

April 13 at 6:34pm • Share

9,690 people like this.

View all 4,842 comments

Mark Zuckerberg
Had a lot of fun on Saturday Night Live tonight! You can check out the clip here:

January 30 at 12:50am via iPhone

People You May Know

Arturito Tett
Add as friend

Xia Xna
Add as friend

Sponsored

New Policy in NEW YORK
allquotespro.com

Drivers with no DUIs in NEW YORK are urged to apply for full coverage car insurance at as low as $30 per month.

Goodyear $80 Rebate
tirebuyer.com

Get up to $80 rebate when you buy a set of Goodyear or Dunlop tires from TireBuyer.com from
## Services: Subscription and OS/Free

- **Subscription/Fee-based**
  - Archive-It
  - CDL-Web Archiving Service
  - Archivethe.net
  - OCLC Web Harvester

- **OS/Free; no support**
  - Web Curator Tool
  - Netarchive Suite
  - Heritrix + Wayback
Corporate/E-Discovery/Compliance

- Iterasi
- Hanzo
- Reed Technology (affil. LexisNexis)
Other tools

- **Scrapbook Plus**
  - Add-on
  - Annotate, select specific content
- **Diigo**
  - Bookmarks, screenshots and annotation
- **Read it Later**
  - Offline copy of site
  - Add-on
- **Zotero**
  - Snapshot, links to live page
  - Locate in Wayback Machine function
- **Bundlr, Delicious**
  - Bookmark collections, link aggregators
- **HTTrack**
  - Offline copy of site
- **GNU Wget (commandline tool)**
  - Local version of site