Making the Ephemeral Endure: Collecting the Web in Research Libraries
COLLECTING WEB RESOURCES: OVERVIEW

- Why
- Who
- Columbia context
- How
- Some Issues
- Questions
MOUVEMENT CONTRE LES ARMES LÉGÈRES EN AFRIQUE DE L’OUEST

ACTUALITÉS

Les femmes du Sénégal se mobilisent contre les armes légères
Des associations sénégalaises de femmes membres de la Société civile, ont plaidé, lundi à Dakar, pour la ratification de la Convention de la CEDEAO sur les Armes légères et de petit calibre (ALPC)....

Le Sénégal ratifie la convention de la CEDEAO sur les Armes Légères et de Petit calibre
Le Sénégal devient ainsi le cinquième pays à ratifier la Convention de la CEDEAO après le Niger, le Mali, le Burkina Faso et la Sierra Leone..

archives...

You are viewing an archived web page, collected at the request of Columbia University Libraries using Archive-It. This page was captured on 18:16:52 Jun 11, 2009, 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 The information on this web page may be out of date.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.
The Audi all road quattro set a unique standard

The Audi all road models may no longer be produced by Audi, but they are the image and standard that has been seen throughout the world. The all road, or Quattro, has been one of the cars that took dominance in the market during the 1980's through the early 1990's. This is a car that has been in a number of rallys and has gained a number of awards. It has been the spearhead that Audi used to not just gain attention in the market, but to gain a level of dominance in most of Europe and also parts of the United States. Considered to be one of the most sought after cars in Eastern Europe, the Quattro models have continued to have a high demand even after more than fifteen years of discontinued production.

The Quattro has been a car that has been highly sought after by people that want a car that has a good design, impressive appearance, and also a reputation as a car that will never give up. The Audi all road...
WHY IT MATTERS


-- citation from Human Rights Review, March 2009
Welcome to ugandacan.org

Related Searches
- Mountain Gorilla
- Cheap Air
- Uganda
- Volunteering
- Uganda Travel
- Hotel Deals
- Discount Airfare
- Luxury Car Rental
- Uganda Tours
- Cruise Vacation
- Vacation Package Deal

RELATED SEARCHES:
- Car Rental
- Travel Insurance
- Cheap Airfare
- Family Vacation Deals
WHY COLLECT WEB RESOURCES

• Libraries build research collections by selecting, acquiring, describing, organizing, managing, and preserving relevant resources.

• Libraries have stable models for collecting non-digital print resources—the roles of selectors, acquisition departments, catalogers, and preservation units are well-understood.
For commercial digital resources a different model has emerged:

- resource bundling
- licensed access rather than physical receipt
- vendor-supplied cataloging
- collective preservation efforts (LOCKSS, Portico)
  - Libraries’ financial investment in these resources has ensured that they are managed
WHY COLLECT WEB RESOURCES

What about non-commercial web resources?
- Many have high research value
- May supplement or replace existing print resources

But as yet we have no common model for:
- Identifying relevant resources
- Integrating access with other collections
- Securing permissions for harvesting
- Preservation
A Lot of Content
MUCH OF IT IS NOT COLLECTED

Refugees International

- 40 documents on web site
- 0 in Columbia collections
- 10 listed in OCLC
  - 1 held by more than 2 libraries
  - No library holds more than 3
### WHO: SOME KEY PROGRAMS

<table>
<thead>
<tr>
<th>Program</th>
<th>Description and Resources</th>
</tr>
</thead>
</table>
| International Internet Preservation Consortium (IIPC) | Members include over 30 international libraries and the Internet Archive.  
|                                            | [http://netpreserve.org](http://netpreserve.org) |
| Archive-IT (Internet Archive)              | Over 100 Institutions using Archive-IT software. Includes universities, schools, state libraries, museums …  
|                                            | [http://www.archive-it.org](http://www.archive-it.org) |
| Web Archiving Service (California Digital Library) | 16 partner institutions  
|                                            | [http://webarchives.cdlib.org](http://webarchives.cdlib.org) |
| Independent Initiatives                    | Commercial web archiving services (Hanzo, Iterasi)  
|                                            | National institutions (libraries, archives) |
MELLON PROJECT ON WEB RESOURCES COLLECTION PROGRAM DEVELOPMENT

**Collection Building**
Make non-commercial web resources of scholarly value an integral part of Columbia’s collection building

**Workflow**
Move web resource collection from a project-based activity to part of routine workflow

**Collaboration**
Develop complementary and collaborative approaches with other research institutions
TERMINOLOGY

- Project
- Program
- Web archiving
- Collecting web resources
CREDITS: COLUMBIA TEAM

- Bob Wolven (Associate University Librarian for Bibliographic Services and Collection Development)
- Stephen Davis (Director, Columbia Libraries Digital Program)
- Pamela Graham (Director, Area Studies and CHRDR)
- Kate Harcourt (Director, Original and Special Materials Cataloging)
- Alex Thurman (Web Collection Curator)
- Tessa Fallon (Web Collection Curator)
GENERAL WEB ARCHIVING WORKFLOW

Selection  Permissions  Crawling  Quality Review  Description  Access
Requirements

- **Crawler**
  - Tool(s) for capturing websites

- **Access/Rendering**
  - Mechanism for viewing captured websites

- **Storage**
  - Storage for data collected and created by the crawler
Selection

- Event based
  - “Arab Spring” websites 2011
  - Japan earthquake 2011
- Thematic
  - Human Rights
  - Blogging in Iran
- Institutional
  - Avery Library, Historic Preservation
  - Burke Library, NYC Religions
- Domain
  - Top level domains such as .uk
CUL EXPERIENCE

Selection models in use at Columbia:

- Subject specialists
- Public nomination form
- Internet Resource Cataloging Request
- Coordination with other library collections
  - Avery Fine Arts and Architecture Library
  - Burke Library/Union Theological Seminary
  - Rare Book and Manuscript Library
  - Columbia University Archives
Unlike other countries, there is no mandate for any US institution to archive websites.

- **Section 108 Study Group** recommendations for web archiving
- Internet Archive Model
- **Oakland Archive Policy**
CUL EXPERIENCE

• Permissions request created in consultation with legal counsel
• Request permission from site owners
• Response from site owners
• Complications
  – Identifying site owners
  – Third party copyright
  – Extent of permission
Crawling

- Collection of URLs required to reproduce a website
- Test crawls gauge size of sites and flag potential crawl issues
- Actual crawls may take hours or weeks
- Product: WARC files (ISO 28500)
SERVICES + OPEN SOURCE

Web Archiving Services
- Archive-It
- CDL-WAS
- Hanzo Archives
- Internet Memory Foundation
- Iterasi

Open Source/Free
- Heritrix + Wayback Machine
- Web Curator Tool
- NetarchiveSuite
- HTTrack
- GNU Wget
- WebCollect toolbar
**Services vs. Open Source**

**Web Archiving Services**
- Customer support and training
- External storage
- Development of interface
- Management of crawler

**Open Source**
- Customizable
- Free software
- Crawled sites are stored locally
Crawling for this collection has not yet been scheduled for the following frequencies: Annual.

You may wish to make adjustments to the crawl frequencies of your seeds before running the crawl. To make changes to individual seeds, use the Seed Management area.

To start crawling for all unscheduled frequencies, click the below button. A crawl will begin immediately and future crawls will be scheduled automatically according to the frequencies. Or use the ‘Start Crawl Now’ buttons at the bottom of the page to start crawling for individual frequencies. Note that One-Time crawls must be started using the 'Start Crawl Now' button.

Collection Management

Seed Management

by seed state:
- All (480)
- Active (343)
- Inactive (143)

by crawl frequency:
- Twice Daily (0)
- Daily (0)
- Weekly (0)
- Monthly (7)
- Bi-monthly (0)
- Quarterly (303)
- Semiannual (0)
- Annual (10)
- One-Time (23)

Crawling Activity

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Last Completed Crawl</th>
<th>Next Scheduled Crawl</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-Time</td>
<td>March 24, 2011 6:04:09 PM EDT [Test]</td>
<td></td>
</tr>
<tr>
<td>Annual</td>
<td>November 29, 2010 3:09:26 PM EST</td>
<td>No Crawls Scheduled</td>
</tr>
<tr>
<td>Quarterly</td>
<td>March 9, 2011 2:26:17 PM EST</td>
<td>June 2, 2011 2:52:23 PM EDT</td>
</tr>
</tbody>
</table>
**Profile basic_domain_scope_WARC:**

- **Description:** Domain scope, writes WARC files, contact email is culhrweb-dev
- **Crawl Operator:** Admin
- **Crawl Organization:**
- **Crawl Job Recipient:**

**crawl-order**

- **max-bytes-download:** 0
- **max-document-download:** 0
- **max-time-sec:** 0
- **max-toe-threads:** 50

**scope**

- **enabled:** True
- **max-link-hops:** 25
- **max-trans-hops:** 5
Quality Review

- Crawler-generated reports
- Crawler traps
- Robots.txt
- URLs captured
- Crawled sites
- Formatting/style
- Navigation
- Multimedia
### Hosts Report

The Hosts Report shows how many URLs were archived from each host, as well as the total amount of data for the collected documents. Other columns in this report provide more information about what was archived.

1. **URLs** refers to the number of documents crawled from each host. Click the number to view the 'URL Report' of exactly what URLs were crawled.
2. "**New URLs**" refers to documents that changed or were newly discovered since the previous crawl. Click the number to view the 'URL Report' of exactly what URLs were crawled.
3. "**Queued**" refers to the number of documents discovered but not crawled due to the crawl time limit.
4. "**Robots.txt Blocked**" refers to documents discovered but not crawled due to a robots.txt exclusion.
5. "**Out of Scope**" refers to documents that were discovered but not crawled as they were determined to be out of scope. The value in this column may be "n/a" before the report has been generated.

Click the number in each column to view more specific information. This information is available 24 hours after a crawl completes.

<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><a href="http://www.es.amnesty.org">www.es.amnesty.org</a></td>
<td>52,465</td>
<td>2.9 GB</td>
<td>52,449</td>
<td>2.9 GB</td>
<td>33,485</td>
<td>0</td>
<td>133,891</td>
</tr>
<tr>
<td><a href="http://www.amnesty.fi">www.amnesty.fi</a></td>
<td>15,318</td>
<td>598.7 MB</td>
<td>15,318</td>
<td>598.7 MB</td>
<td>0</td>
<td>0</td>
<td>6,325</td>
</tr>
<tr>
<td><a href="http://www.aprodeh.org.pe">www.aprodeh.org.pe</a></td>
<td>10,884</td>
<td>1.0 GB</td>
<td>10,882</td>
<td>1.0 GB</td>
<td>0</td>
<td>0</td>
<td>2,761</td>
</tr>
<tr>
<td><a href="http://www.amnesty.si">www.amnesty.si</a></td>
<td>5,785</td>
<td>375.7 MB</td>
<td>5,785</td>
<td>375.7 MB</td>
<td>0</td>
<td>0</td>
<td>5,380</td>
</tr>
<tr>
<td><a href="http://www.amnesty.dk">www.amnesty.dk</a></td>
<td>4,082</td>
<td>257.7 MB</td>
<td>4,080</td>
<td>257.7 MB</td>
<td>0</td>
<td>0</td>
<td>4,471</td>
</tr>
<tr>
<td><a href="http://www.amnesty.cz">www.amnesty.cz</a></td>
<td>3,481</td>
<td>122.6 MB</td>
<td>3,481</td>
<td>122.6 MB</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td><a href="http://www.amnesty.org.ph">www.amnesty.org.ph</a></td>
<td>2,907</td>
<td>245.5 MB</td>
<td>2,907</td>
<td>245.5 MB</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td><a href="http://www.amnesty.se">www.amnesty.se</a></td>
<td>2,001</td>
<td>21.5 MB</td>
<td>2,001</td>
<td>21.5 MB</td>
<td>0</td>
<td>0</td>
<td>918</td>
</tr>
<tr>
<td><a href="http://www.amnesty.at">www.amnesty.at</a></td>
<td>1,879</td>
<td>173.1 MB</td>
<td>1,879</td>
<td>173.1 MB</td>
<td>0</td>
<td>0</td>
<td>196</td>
</tr>
<tr>
<td><a href="http://www.amnesty.sk">www.amnesty.sk</a></td>
<td>1,564</td>
<td>57.1 MB</td>
<td>1,584</td>
<td>57.1 MB</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><a href="http://www.amnesty.org.gr">www.amnesty.org.gr</a></td>
<td>1,392</td>
<td>32.6 MB</td>
<td>1,389</td>
<td>32.6 MB</td>
<td>0</td>
<td>0</td>
<td>432</td>
</tr>
</tbody>
</table>
Scoping

- Testing phase or post-crawl
- Excluding out-of-scope URLs
- Expanding scope
  - Additional domains (common: blogs, other languages, subordinate sections of an organization)
- Excluding crawler traps
Access

- Wayback Machine or equivalent necessary to render the WARC files
- Description: metadata created by cataloging staff
- Access
  - Web archiving service
  - OPAC
  - Portal
  - Consortium
Challenges

- Rapidly changing technologies used in website development
- Dynamic pages, deep web, other inaccessible content
- Providing access across collections and avoiding web archive silos
- Aggregation of data: long-term storage and responsibility
- Long-term preservation challenges
ISSUES

Scale, sustainability
- Matching scale to program objectives
- Budgeting for storage, staffing

Scope; collection policy
- Limit to a few concentrations or broader?
- Defining by source (.org), format (.pdf), topic?
- What happens to resources excluded?

Collaboration, external
- Duplication/overlap with related initiatives
- Complementary approaches: frequency, access, scoping
- Role of Archive-IT partners, consortia (2CUL), NDSA
MORE ISSUES

Coordination, internal

- Relation to institutional repository, archival collections, earchives

Staffing, roles

- Centralized vs distributed effort
- Impact on selectors, cataloging, archivists, digital program

Impact on print collecting

- Potential for “e-only”
STILL MORE ISSUES

Technical

• Local vs. hosted storage
• Open source, local development vs externally-supported toolkit
• Moving from harvesting to archiving

Access, Use, Assessment

• Use cases for portals, cross-collection searching
• Disclosure outside local context
WHAT DO LIBRARIES DO?

• Libraries build research collections by selecting, acquiring, describing, organizing, managing, and preserving relevant resources
• Libraries manage business transactions necessary to provide access to resources needed for research
• Libraries preserve research resources to enable access to be restored if lost
ADDITIONAL INFORMATION

CUL Mellon Project on Web Resource Collection Development Program
- Project Information
- Human Rights Web Archive
- Archive-It Collections Page
- Human Rights Web Archive Delicious Survey

Other Web Archives
- Archive-It Partners
- IIPC
- Internet Archive
- Internet Memory Foundation
- Web Archiving Initiatives wiki

Services + Tools:
- Heritrix
- Wayback Machine (newest Beta version)
- Archive-It
- CDL-WAS
- NetarchiveSuite
- Internet Memory Foundation
- Web Curator Tool
- Web Collect Toolbar
- GNU wGet
- HTTrack