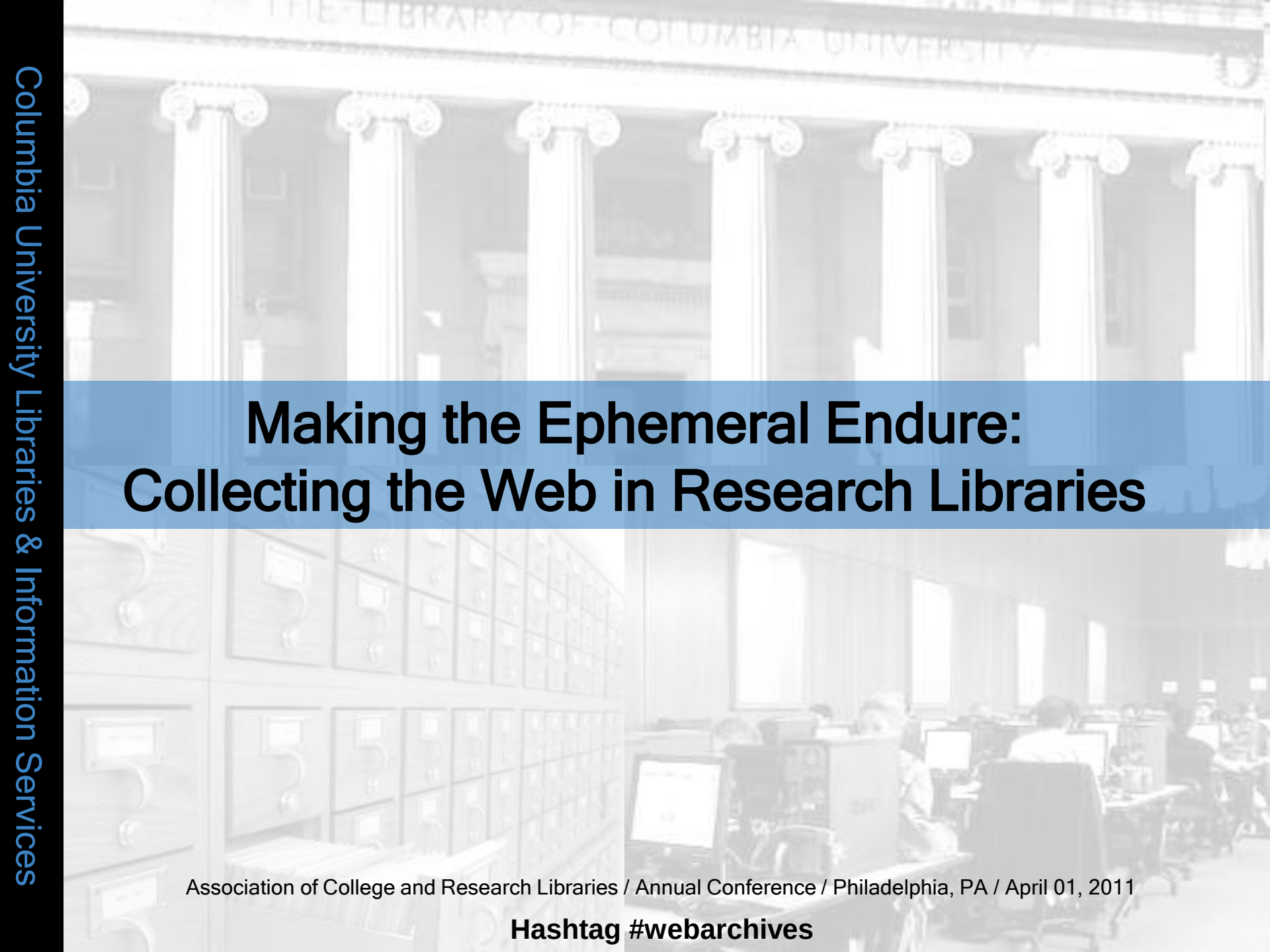


Columbia University Libraries & Information Services

Making the Ephemeral Endure: Collecting the Web in Research Libraries

Association of College and Research Libraries / Annual Conference / Philadelphia, PA / April 01, 2011

Hashtag #webarchives

Collecting Web Resources: Overview

- Why
- Who
- Columbia context
- How
- Some Issues
- Questions

WHY IT MATTERS…

THE LIVE SITE TODAY

OFF Road Adventure ÐÐ

Ads by Google | Rally | Video Rally WRC | Sturgis Rally Photos | Audi Cars | Biker Rally Pictures
Ads by Google | Audi Quattro | Audi S1 | Car Driving Rally | 3D Rally Race | Adventure Rally Suit

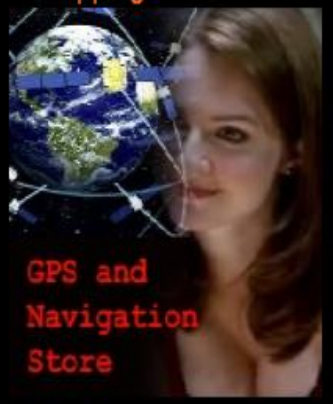
THE AUDI ALL ROAD QUATTRO SET A UNIQUE STANDARD

– Filed under: Rally Adventure

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 ro more »

GO Shopping...



Most Read

Off drive place win make To rally day tent line roads part class world way Don Lot age Star sport hand use special time ex Driving set thing Road speed driver end Racing vehicle start safety

Why It Matters

- “Joaquim Chissano Appointed UN Special Envoy for LRA-Affected Region 2007,” Uganda-CAN [article on-line]; available from http://www.ugandacan.org/item/1846; accessed 17 May 2007"

-- citation from Human Rights Review, March 2009

THE LIVE SITE TODAY

The domain ugandacan.org may be for sale. Click here for details.

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

Why Collect Web Resources

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

LSE Centre for the Study of Human Rights

international center for TRANSITIONAL JUSTICE

CEDHA
Center for Human Rights and Environment

Tibetan Centre for Human Rights and Democracy
བོད་ཀྱི་འགྲོ་བ་མིའི་ཐོབ་ཐང་དང་མང་གཙོ་འཕེལ་རྒྱས་ལྟེ་གནས་ཁང་།

hrea.org
Human Rights Education Associates

ХҮНИЙ ЭРХ ХӨГЖИЛ ТӨВ
CENTER FOR HUMAN RIGHTS AND DEVELOPMENT

Palestinian Centre for Human Rights
المركز الفلسطيني لحقوق الإنسان

Home Contact Search Site Map

- Special Consultive Status, ECOSOC
- Affiliate, Federation Internationale des Ligues des Droits de l'Homme
- Affiliate, International Commission of Jurists
- Member, Euro-Mediterranean Human Rights Network
- International Legal Assistance Consortium (ILAC)

MUCH OF IT IS NOT COLLECTED

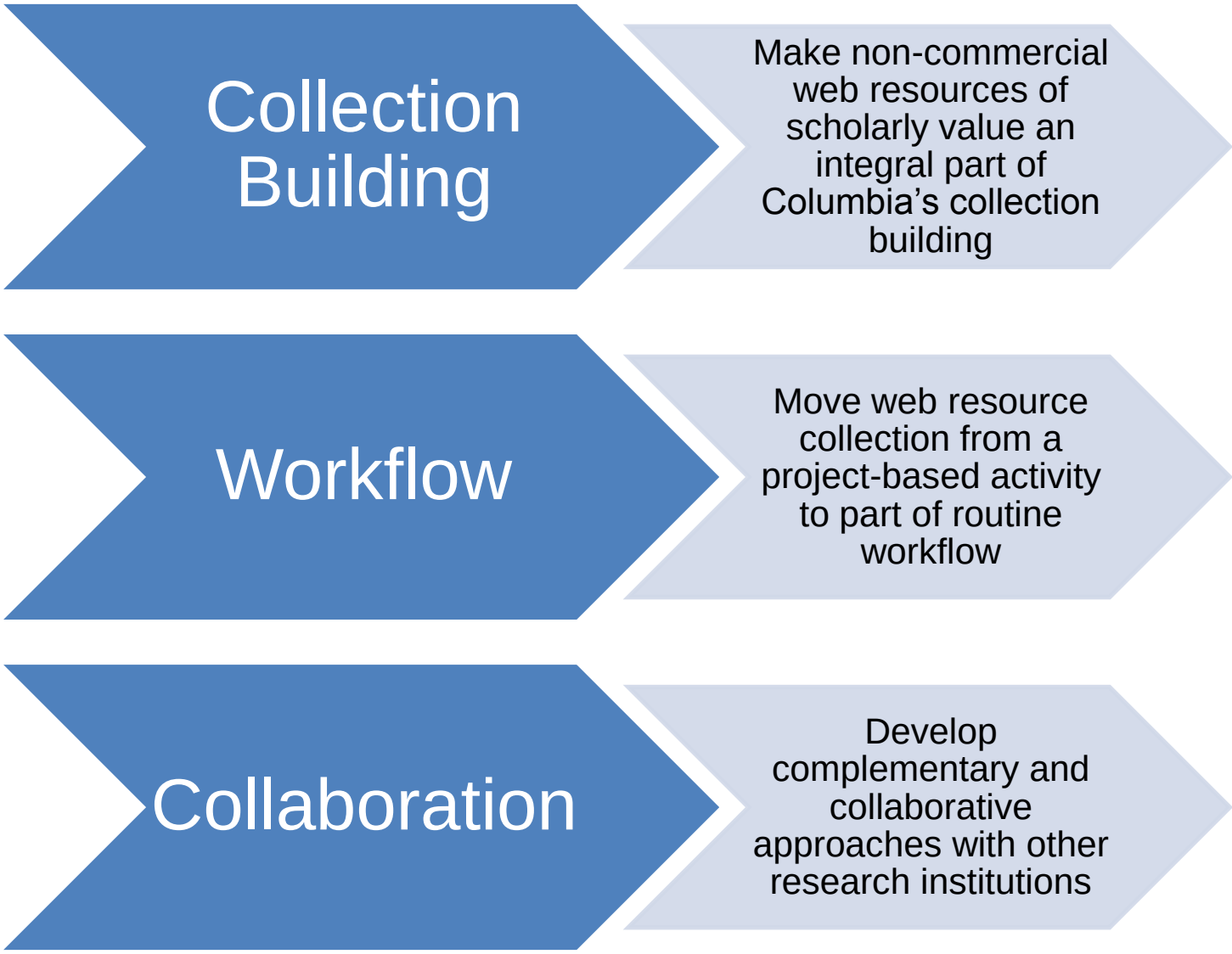
Refugees International

- 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

International Internet Preservation Consortium (IIPC)	• Members include over 30 international libraries and the Internet Archive. • 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
Web Archiving Service (California Digital Library)	• 16 partner institutions • http://webarchives.cdlib.org
Independent Initiatives	• Commercial web archiving services (Hanzo, Iterasi) • National institutions (libraries, archives)

Mellon Project on Web Resources Collection Program Development



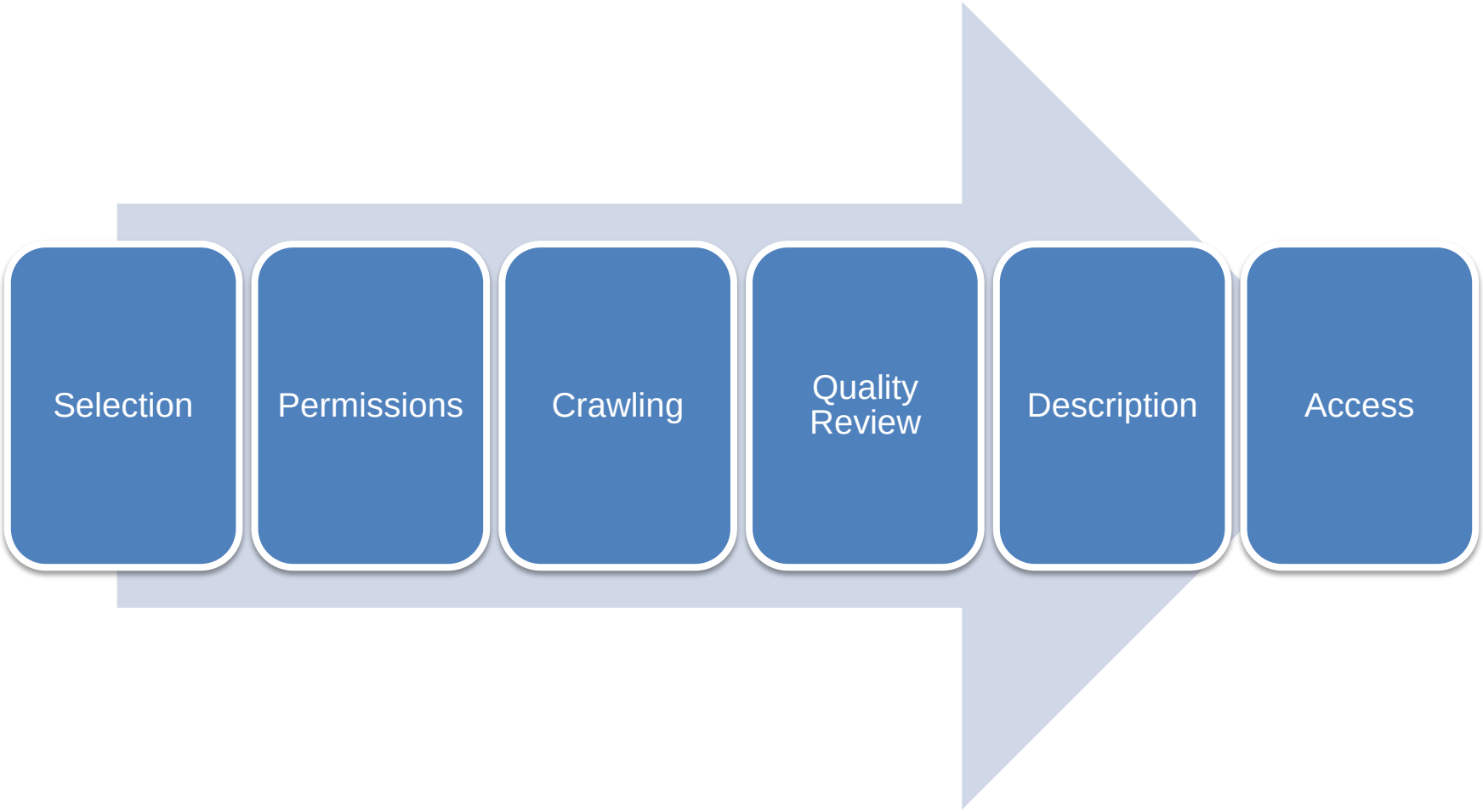
TERMINOLOGY

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)

Columbia University Libraries & Information Services

GENERAL WEB ARCHIVING WORKFLOW



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

Copyright

+

Permissions

- 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

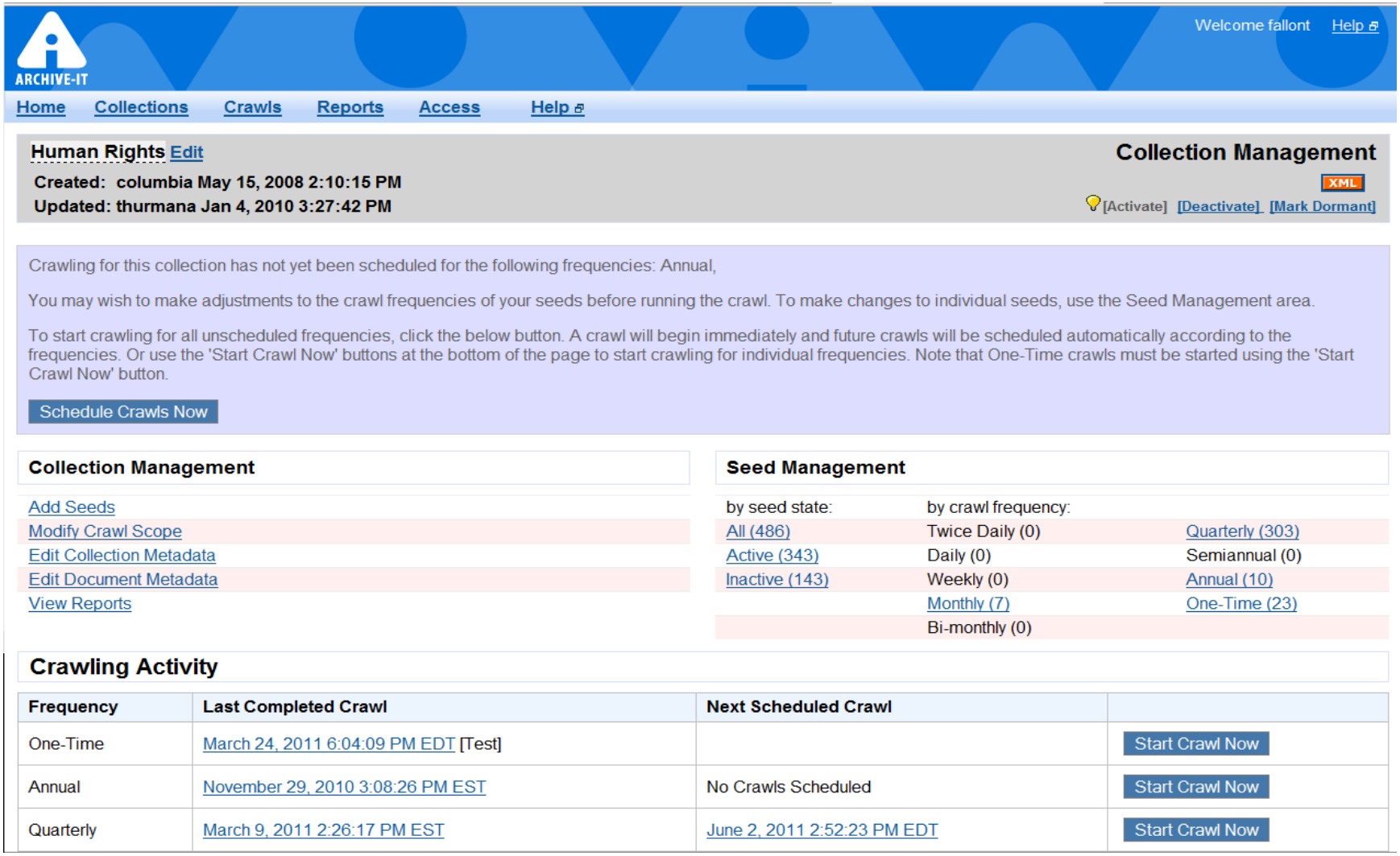
Web Archiving Services

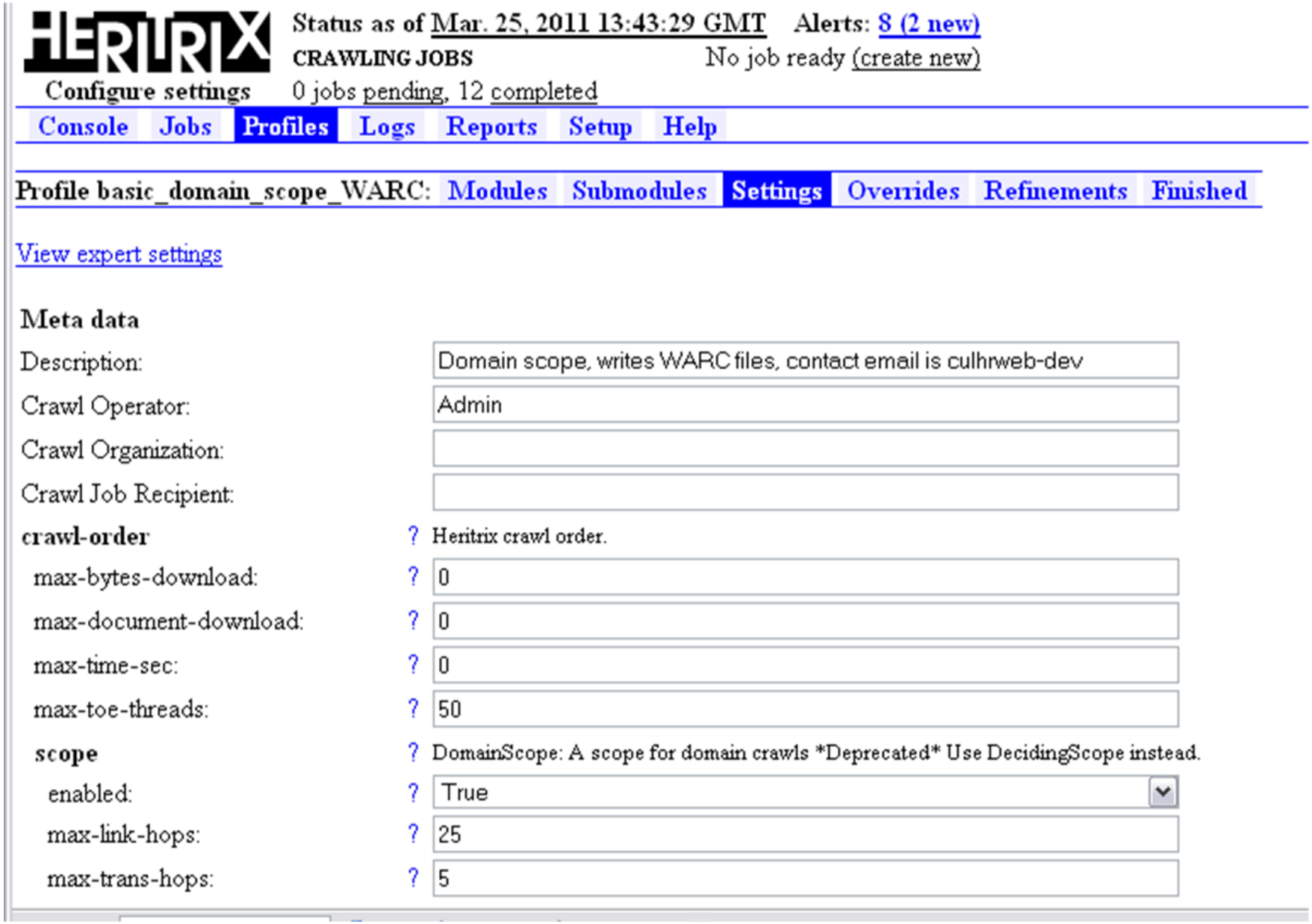
- Customer support and training
- External storage
- Development of interface
- Management of crawler

Open Source

- Customizable
- Free software
- Crawled sites are stored locally

CUL EXPERIENCE





Quality Review

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

CUL EXPERIENCE

Human Rights

One-Time (ID #20110317180036550)

Crawl Report

Started: March 17, 2011 2:00:37 PM
Completed: March 22, 2011 2:33:11 PM

<< Back to Reports | **Scope-It Crawl Explorer**

Summary | Hosts | Seed Status | Seed Source | File Types | PDFs | Videos | QA

Hosts **Download**

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.

View only hosts containing [ ] Filter Clear

« First « Previous Next » Last » 1 through 100 of 2674

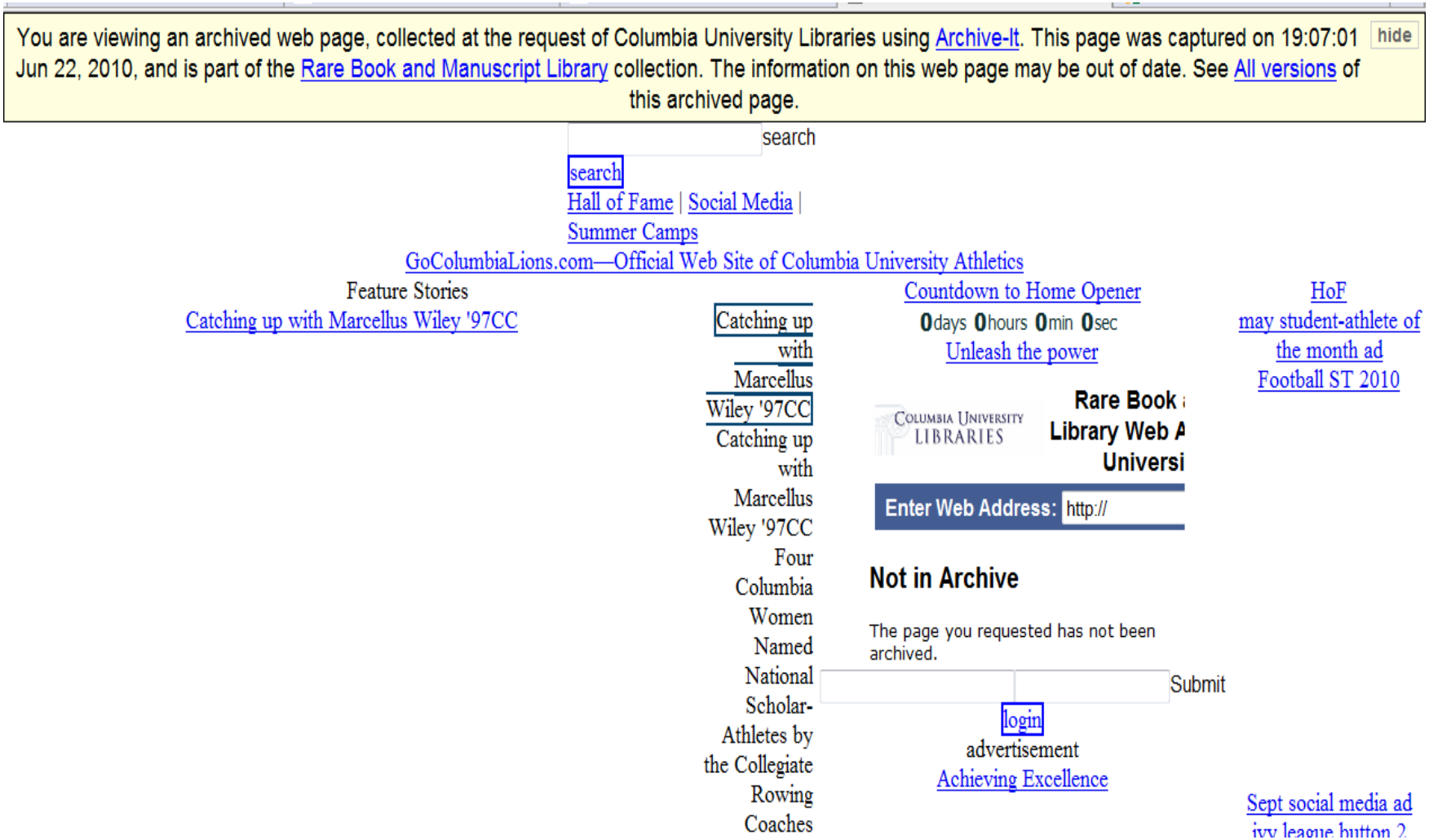
Host	URLs	Data	New URLs	New Data	Queued	Robots.txt Blocked	Out of Scope
www.es.amnesty.org	52,455	2.9 GB	52,449	2.9 GB	33,485	6,443	133,891
www.amnesty.fi	15,318	598.7 MB	15,318	598.7 MB	0	0	6,325
www.aprodeh.org.pe	10,884	1.0 GB	10,882	1.0 GB	0	0	2,761
www.amnesty.si	5,785	375.7 MB	5,785	375.7 MB	0	0	5,380
www.amnesty.dk	4,082	257.7 MB	4,080	257.7 MB	0	0	4,471
www.amnesty.cz	3,481	122.6 MB	3,481	122.6 MB	0	0	7
www.amnesty.org.ph	2,907	245.5 MB	2,907	245.5 MB	0	0	2
www.amnesty.se	2,001	21.5 MB	2,001	21.5 MB	0	0	918
www.amnesty.at	1,878	173.1 MB	1,878	173.1 MB	0	0	196
www.amnesty.sk	1,564	57.1 MB	1,564	57.1 MB	0	0	0
www.amnesty.org.gr	1,392	32.6 MB	1,389	32.6 MB	0	0	432

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

Columbia University Libraries & Information Services

CUL EXPERIENCE

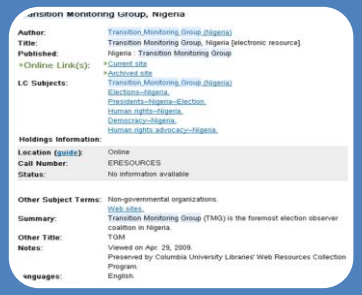


Access

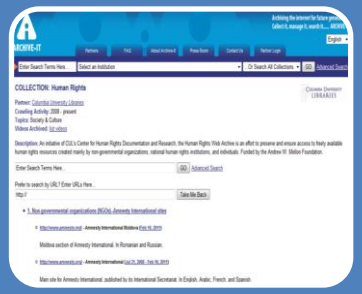
- Wayback Machine or equivalent necessary to render the WARC files
- Description: metadata created by cataloging staff
- Access
  - Web archiving service
  - OPAC
  - Portal
  - Consortium

Columbia University Libraries & Information Services

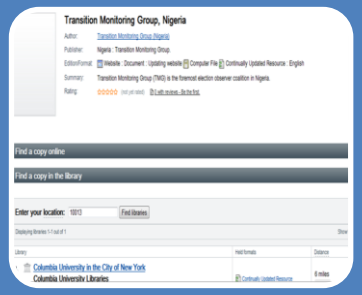
CUL EXPERIENCE



CLIO



Archive-It



WorldCat

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, e-archives

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