



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...

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 The information on this web page may be out of date.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.

[Metadata](#)

MOUVEMENT CONTRE LES ARMES LÉGÈRES EN AFRIQUE DE L'OUEST



[Accueil](#) | [Plan du site](#) | [Contacts](#)

Recherche

Ok

tes « légères et de petit calibre ».

ACTUALITES

MALAO

ARMES LÉGÈRES ET PAIX

RÉSEAUX

FEMMES

PROJETS

INFORMATIONS

PARTENAIRES



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 de fait, le cinquième pays à ratifier la Convention de la CEDEAO après le Niger, le Mali, le Burkina Faso et la Sierra Leone..

archives...

MA FORCE EST DANS LA PAIX



Site réalisé avec le soutien de :



COMMUNIQUES

DOCUMENTATION

EMPLOI

Communiqué de presse du 2 mai

Bulletins

Air pour le MALAO

THE LIVE SITE TODAY

OFF ROAD ADVENTURE

[Ads by Google](#)
[Ads by Google](#)

[Rally](#)
[Audi Quattro](#)

[Video Rally WRC](#)
[Audi S1](#)

[Sturgis Rally Photos](#)
[Car Driving Rally](#)

[Audi Cars](#)
[3D Rally Race](#)

[Biker Rally Pictures](#)
[Adventure Rally Suit](#)

THE AUDI ALL ROAD QUATTRO SET A UNIQUE STANDARD

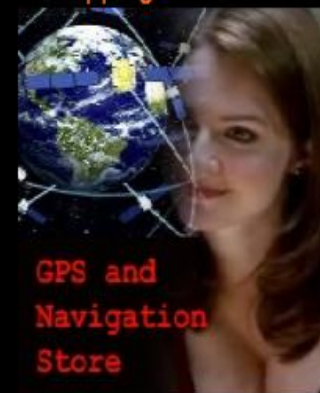
- Filed under: [Rally Adventure](#)



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 rallies 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...



[GPS and Navigation Store](#)

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



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་ལྗོངས་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

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

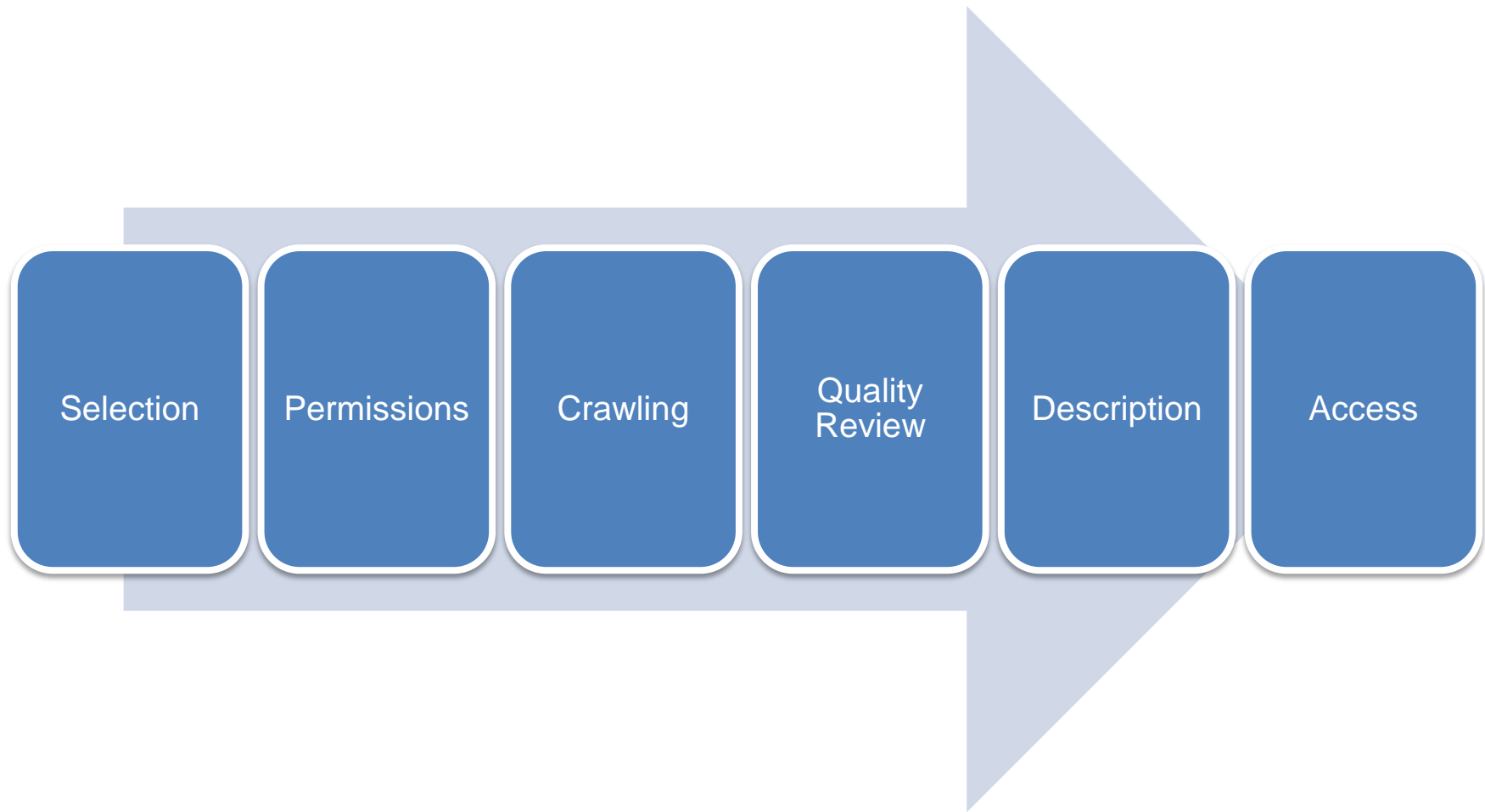
TERMINOLOGY



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



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



Welcome fallont [Help](#)

- [Home](#)
- [Collections](#)
- [Crawls](#)
- [Reports](#)
- [Access](#)
- [Help](#)

Human Rights [Edit](#)

Collection Management

Created: columbia May 15, 2008 2:10:15 PM

Updated: thurmana Jan 4, 2010 3:27:42 PM



[\[Activate\]](#) [\[Deactivate\]](#) [\[Mark Dormant\]](#)

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.

[Schedule Crawls Now](#)

Collection Management

[Add Seeds](#)

[Modify Crawl Scope](#)

[Edit Collection Metadata](#)

[Edit Document Metadata](#)

[View Reports](#)

Seed Management

by seed state:

[All \(486\)](#)

[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

Frequency	Last Completed Crawl	Next Scheduled Crawl	
One-Time	March 24, 2011 6:04:09 PM EDT [Test]		Start Crawl Now
Annual	November 29, 2010 3:08:26 PM EST	No Crawls Scheduled	Start Crawl Now
Quarterly	March 9, 2011 2:26:17 PM EST	June 2, 2011 2:52:23 PM EDT	Start Crawl Now



Status as of Mar. 25, 2011 13:43:29 GMT Alerts: [8 \(2 new\)](#)

CRAWLING JOBS

No job ready ([create new](#))

[Configure settings](#) 0 jobs [pending](#), 12 [completed](#)

[Console](#) [Jobs](#) [Profiles](#) [Logs](#) [Reports](#) [Setup](#) [Help](#)

Profile [basic_domain_scope_WARC](#): [Modules](#) [Submodules](#) [Settings](#) [Overrides](#) [Refinements](#) [Finished](#)

[View expert settings](#)

Meta data

Description:

Crawl Operator:

Crawl Organization:

Crawl Job Recipient:

crawl-order ? Heritrix crawl order.

max-bytes-download: ?

max-document-download: ?

max-time-sec: ?

max-toe-threads: ?

scope ? DomainScope: A scope for domain crawls *Deprecated* Use DecidingScope instead.

enabled: ?

max-link-hops: ?

max-trans-hops: ?

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

CUL EXPERIENCE

You are viewing an archived web page, collected at the request of Columbia University Libraries using [Archive-It](#). This page was captured on 19:07:01 Jun 22, 2010, and is part of the [Rare Book and Manuscript Library](#) collection. The information on this web page may be out of date. See [All versions](#) of this archived page.

hide

search

[search](#)

[Hall of Fame](#) | [Social Media](#) |

[Summer Camps](#)

[GoColumbiaLions.com—Official Web Site of Columbia University Athletics](#)

Feature Stories

[Catching up with Marcellus Wiley '97CC](#)

[Catching up with Marcellus Wiley '97CC](#)

Catching up with Marcellus Wiley '97CC

Four Columbia Women Named National Scholar-Athletes by the Collegiate Rowing Coaches

[Countdown to Home Opener](#)

0 days 0 hours 0 min 0 sec

[Unleash the power](#)



Rare Book
Library Web A
Universi

Enter Web Address:

Not in Archive

The page you requested has not been archived.

[login](#)

advertisement

[Achieving Excellence](#)

[HoF](#)

[may student-athlete of the month ad Football ST 2010](#)

[Sept social media ad](#)
ivv leaone button ?

Access

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

CUL EXPERIENCE

Transition Monitoring Group, Nigeria

Author: [Transition Monitoring Group \(Nigeria\)](#)

Title: [Transition Monitoring Group, Nigeria](#) (electronic resource)

Published: Nigeria : Transition Monitoring Group

Online Link(s): [Current site](#)
[Archived site](#)

LC Subject: [Transition Monitoring Group \(Nigeria\)](#)
[Elections--Nigeria](#)
[Transition--Nigeria--Election](#)
[Human rights--Nigeria](#)
[Democracy--Nigeria](#)
[Political parties--Nigeria](#)

Holdings Information:

Location (global): Online

Call Number: ERESOURCES

Status: No information available

Other Subject Terms: Non-governmental organizations
[Web sites](#).

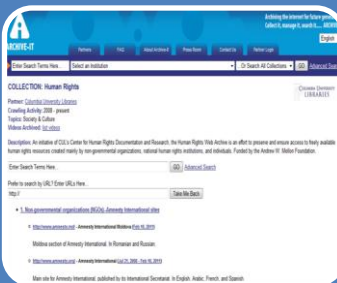
Summary: [Transition Monitoring Group \(TMG\) is the foremost election observer coalition in Nigeria.](#)

Other Title: TGM

Notes: Viewed on Apr. 28, 2008.
Preserved by Columbia University Libraries' Web Resources Collection Program.

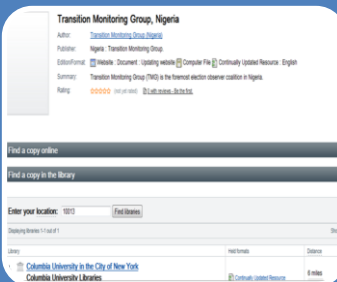
Language: English

CLIO



Archive-It interface showing the 'Human Rights' collection. The page displays a search bar, navigation tabs (Home, Help, About, Search), and a list of documents. The main content area shows the title 'Human Rights', the creator 'Columbia University Libraries', and the description: 'Description: An initiative of the Center for Human Rights Documentation and Research, the Human Rights Web Archive is an effort to preserve and ensure access to freely available human rights resources created freely by non-governmental organizations, national human rights institutions, and scholars. Funding by the Andrew W. Mellon Foundation.' Below the description is a search bar and a list of document titles, including '1. Non-governmental organizations (NGOs): Amnesty International site' and '2006/2007 Human Rights Report: Amnesty International Release Fall 6, 2007'.

Archive-It



WorldCat interface showing the 'Transition Monitoring Group, Nigeria' record. The record includes the author 'Transition Monitoring Group (Nigeria)', the publisher 'Nigeria : Transition Monitoring Group', and the title 'Transition Monitoring Group, Nigeria'. It also shows a summary: 'Summary: Transition Monitoring Group (TMG) is the foremost election observer coalition in Nigeria.' Below the record information is a search bar and a list of library locations, including 'Columbia University in the City of New York'.

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](#)