Collecting web resources : selecting, harvesting, cataloging

Plans for a web resources collection program at the Columbia University Libraries

Alex Thurman

NETSL 2009

Overview

- Why collect web resources?
- Why the focus on human rights?
- What we've done so far
- What we plan to do
- What we hope to accomplish

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, involving:
  - 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
  - Disclosure

Web archiving organizations

- International Internet Preservation Consortium (IIPC)
  - Members include over 30 international libraries and the Internet Archive. Has working groups devoted to Standards, Harvesting, Access, and Preservation.

    http://netpreserve.org
- Living Web Archives (LiWA)
  - Consists of 8 European institutions. Dedicated to technical advancements in web content capture, preservation, analysis.

    http://www.liwa-project.eu/

Web archiving projects

- Domain-wide
  - Internet Archive's Wayback Machine
  - National Library of Sweden's Kulturarw3
- Event-based
  - Library of Congress's Minerva (elections, Iraq war, Hurricane Katrina)
- Thematic
  - University of Texas's LAGDA (Latin American Government Documents Archive) & LAGRA
  - North Carolina State Government Web Site Archives

CUL program objectives

- Make non-commercial web resources an integral part of Columbia's collection building
- Move web resource collection from a project-based activity to part of routine workflow
- Begin with subject focus on human rights

Human rights

- Broad multidisciplinary subject—but main tenets are elaborated in the 30 articles of the Universal Declaration of Human Rights (UN, 1948)
- Freedom from slavery, torture, discrimination, arbitrary arrest
- Rights to equal protection, fair trial, movement, asylum, property, work, marriage, freedom of thought, freedom of expression

Human rights at Columbia

- Center for the Study of Human Rights

  http://hrcolumbia.org/

- Columbia Law School, Human Rights Institute

  http://www.law.columbia.edu/center_program/human_rights

- Center for Human Rights Documentation & Research (CHRDR)

  CHRDR houses the physical archives of Amnesty International USA, Human Rights Watch, and the Committee of Concerned Scientists

  http://www.columbia.edu/cu/lweb/indiv/humanrights/

Human rights web resources

- Sources
  - Governmental
  - Inter-Governmental Organizations (IGOs)
  - Non-Governmental Organizations (NGOs)
  - Academic institutes, libraries
  - Blogs, news sites
- Types of content
  - Annual reports, country reports, case studies, news bulletins, legal instruments, statistics, video, audio, images, maps

What we've done so far

- Secured 1-year planning grant from the Andrew W. Mellon Foundation for 2008, with grant partner the University of Maryland Libraries (their subject focus was historic preservation)
- The planning grant was used to fund project manager positions at both institutions

What we've done so far

- Surveyed existing web content related to human rights
- Tagged over 600 websites, mainly NGOs, at http://delicious.com/hrwebproject, recording both descriptive and administrative (project-related) metadata
- Researched existing harvesting tools and ran test crawls of over 80 sites using Archive-It

Web content tagging

- Descriptive metadata elements tagged for each site:
  - Content type
  - Organizational type
  - Organizational home (country)
  - Geographic focus (region and/or country)
  - Thematic focus
  - Language(s)
  - Organizational name (i.e. website name)
  - Authorized heading for organization (if in NAF)

Web content tagging

- Administrative aspects tagged:
  - Amount of print titles by organization already in our OPAC, CLIO
  - Website itself already in CLIO?
  - NAF record exists for organization?
  - Initials of selectors who submitted/approved site
  - Test crawls run?
  - Robots.txt restrictions found on site?
  - Record migrated into CLIO?

Migrating delicious.com data

- Mapped metadata from delicious.com web survey to access-level MARC records
- Migrated delicious-to-MARC records into CLIO
- Begun light revision of records, including establishment of NAF headings where lacking

Evaluating harvesting tools

- Commercial hosted services (combine crawling and archiving)
- Locally run tools, commercial or open source (allow more flexible crawling, require local technical support, do not address archiving)

Commercial hosted services

- Archive-It (mostly academic and government clients)

  http://www.archiveit.org

- Hanzo Archives (mostly corporate clients)

  http://www.hanzoarchives.com

- OCLC Web Harvester (bundled with CONTENTdm)

  http://www.oclc.org/webharvester/

- Web Archiving Service (California Digital Library tool currently limited to University of California partners)

  http://www.cdlib.org/inside/projects/preservation/webatrisk/web_archiving.html

Locally run harvesting tools

- Open source
  - Web Curator Tool (developed by IIPC)

    http://webcurator.sourceforge.net/
  - NetarchiveSuite (Danish national libraries)

    http://netarchive.dk/kildetekster/index-en.php
  - HTTrack (free offline browser utility)

    http://www.httrack.com/
- Commercial
  - WebCopier Pro (made by MaximumSoft)

    http://www.maximumsoft.com/

Archive-It trial

- Performed test Archive-It crawls on over 80 websites (yielded 1.8+ million documents [i.e.] objects with distinct urls, taking up 68 GB)
- Selected test sites based on factors such as importance and nature of content; country of origin; type of organization; overlap with print collections; perceived risk of site disappearance
- Most crawls successful; however, several sites had partial robots.txt restrictions, or technical issues that led to unsuccessful crawls

Archive-It trial

- Advantages
  - Hosted, requires no installation or local technical support
  - Provides crawling and long-term storage
  - Short learning curve—you can begin crawling sites immediately
  - Good customer support
  - Good help wiki
  - Software is actively upgraded, often based on partner feedback
  - Archived content can be migrated out of Archive-It into locally hosted options if desired
  - Includes partner page with public access to all archived content, and a customizable template for partners to use to create their own branded portal page

Archive-It trial

- Drawbacks
  - Lack of flexibility in crawl management
  - Crawls managed by frequency, not by seed
  - Reports based on crawls, not seeds
  - No built-in crawl quality assessment module
  - No permissions module
  - Archived content can't be moved from one "collection" to another
  - No automatic metadata extraction from crawled sites
  - Partner-applied metadata not used by archiveit.org search engine

Current program status

- One-year (2008) planning grant completed
- Three-year (2009-2012) implementation grant proposal submitted to Andrew W. Mellon Foundation
- Grant proposal requests funding of 2 three-year Web Curator positions and one two-year Digital Library Analyst position
- We'll see …

Project plan

- Selection
- Permissions
- Harvesting and archiving
- Description and organization
- Disclosure
- Making content available for use
- Assessment

Selection

- Continue adding to survey of relevant resources maintained on delicious.com page, via:
  - Submissions from Columbia area studies librarians and teaching faculty, CUL subject guides
  - Gathering, following links from web directories of human rights NGO content (Duke, Minnesota, GODORT)
  - Web-based form to solicit input from listservs of librarians and scholars

Selection

- Prioritize resources to determine appropriate treatment
  - Highest priority candidates for harvesting will be NGO-produced resources from countries without strong national archiving programs, particularly sites viewed most “at-risk”
  - Governmental, IGO, and academic websites, or NGO websites based in the USA, Western Europe, Australia or Canada will be cataloged and included in access and discovery venues but not harvested

Permissions

- Secure explicit permissions agreements with organizations for which Columbia holds physical archives
- Develop a generic Memorandum of Understanding for web harvesting with which to approach other selected organizations
- When contact with organizations is not feasible or site is deemed “at-risk” of disappearing, proceed with non-intrusive harvesting

Permissions

- Principles for non-intrusive harvesting (see Section 108 Study Group Report)
  - Respect robots.txt restrictions
  - Frame harvested content to clearly indicate its nature
  - Link to original site as well as archived content
  - Remove harvested content upon request by site owner

Harvesting and archiving

- Proceed with large-scale harvesting using Archive-It
- Develop procedures for regular quality assessment of web crawls
- Continue to test locally run tools such as Web Curator Tool and WebCopier Pro for more control when integrating selected web content into our local environment

Description and organization

Our approach to description will be multi-faceted and experimental

- Access-level MARC records for all selected sites will be generated from delicious.com metadata
- Harvested sites with complex content groups will receive finding aids, treating the sites as analogous to archival collections (as in the “Arizona model”)

Description and organization

- Some cross-organizational finding aids, highlighting specific topics, regions, or genres, will also be created
- Selected serials and documents from these sites will receive their own MARC records, for better integration with existing library collections
- Student interns will assist with creation of metadata for newly added sites and with updating the initial set of catalog records and finding aids as sites are re-crawled and new content

Disclosure

- No standard exists for establishing whether a particular website is being archived, and if so, with what frequency and depth
- Our program will attempt to disclose its work beyond Columbia's local systems
- MARC records for selected serials and documents will be exposed in Worldcat, and those harvested will be registered in OCLC's Registry of Digital Masters
- MODS will be used to generate collection-level and series-level records from finding aids to increase potential record sharing

Making content available for use

- To best integrate archived web resources into our campus search and discovery environment, we will combine use of Archive-it for full-website archiving with selective local archiving of document-like content in Columbia's Fedora-based repository
- Web Curator Tool, or a similar tool, will be used for targeted document-level harvesting

Making content available for use

- We will build on our relationships with those human rights organizations that have deposited their physical archive collections at Columbia
- Web resources from these three organizations will be most thoroughly exposed, forming the core of an evolving "Human Rights Electronic Reference Collection" to be hosted on the website of Columbia's Center for Human Rights Documentation and Research

Making content available for use

- For this core reference collection we will explore techniques for generating XML resource maps of harvested sites
- These maps could be displayed in conjunction with EAD finding aids, acting as the equivalent to an archival “container list”, but with direct links to the archived content
- Resource maps from each of the depositor organizations could be merged to create a composite map correlating geographical and thematic content across the three organizations

Assessment

- Input from scholars, librarians, archivists, and representatives of human rights NGOs will be regularly solicited in two key areas:
  - Selection of content for archiving
  - Usability of content presentation

Web Collection Program Goals

- Substantial collection of human rights web resources, and infrastructure to sustain and grow it
- Framework for resource discovery, access, and presentation
- Integration of procedures for collecting web resources with the routine work of selectors, e-resources staff, archival curators, and catalogers

Web Collection Program Goals

- Establish an infrastructure that could be extended into other subject areas important to Columbia's collections as needed
- Serve as a model for other libraries to use, adapt, and improve in their own web collecting activities

Acknowledgments

- This presentation is based on a program grant proposal authored by Robert Wolven with help from the other Steering Committee members

Program Steering Committee

- Robert Wolven (Associate University Librarian for Bibliographic Services and Collection Development)
- Stephen Davis (Director, Columbia Libraries Digital Program)
- Pamela Graham (Latin American & Iberian Studies Librarian)
- Kathryn Harcourt (Director, Original and Special Materials Cataloging)
- Alex Thurman (Catalog Librarian) [at2186@columbia.edu]
- Melanie Wacker (Metadata Coordinator)

Further reading

- Brown, Adrian. *Archiving websites : a practical guide for information professionals*. London: Facet, 2006.
- Pierce-Moses, Robert and J. Kaczmarek. "An Arizona Model for Preservation and Access of Web Documents." Dttp: Documents to the People. 33:1 (2005) 17-24.
- Section 108 Study Group Report

  http://www.section108.gov/docs/Sec108StudyGroupReport.pdf
- http://delicious.com/webarchiving