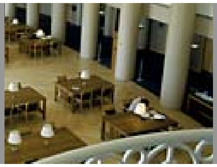
Web Archives Workbench

machine-assisted management of web harvesting

Taylor Surface, OCLC

Representing the *ECHO DEPository Project*

DLF Forum, November 2005



ECHO DEPository, Project Partners

- University of Illinois, Urbana-Champaign
  - Libraries, GSLIS, NCSA, WILL, DMI
- OCLC
- Tufts – Perseus Project
- Michigan State – Vincent Voice Library
- State Libraries of Arizona, Connecticut, Illinois, North Carolina and Wisconsin

... an NDIIPP project in partnership with the Library of Congress

ECHO DEPository – Overview

- Design selection methodology
- Develop software implementing theory
  - Machine-assisted
  - Open source
- Evaluate various repositories
  - Using content gathered from tools
  - Other content providers
- Study semantic preservation techniques

Three objectives

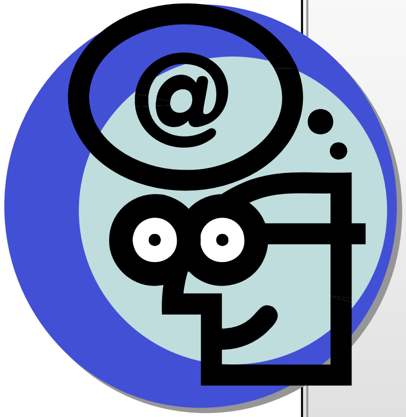
- Comparative test of repositories with various digital collections
- Development of Web Archives Workbench
- Investigations of semantic digital preservation and alternate applications of workbench tools

The Arizona Model

- Web domains as “archival collections”
- Creates efficiencies for …
  - Selection of “documents”
  - Name authority & other metadata
  - Browseable access

Arizona Model: a new approach

- Assumptions
  - Content creators won’t help
  - Item by item selection is unsatisfactory
  - Bulk harvesting is unsatisfactory
- An archival approach
  - Identifying groups of similar material (series)
  - Automatic identification of new series items
  - Series description
    - Item level description is possible if warranted
  - Ingest of documents into an archive



Web Archives Workbench (WAW)

Tools for curators …

- Discovery – identify & manage domains
- Properties – associate metadata, content, and providers
- Analysis – select content from structure
- Packager – package content & metadata

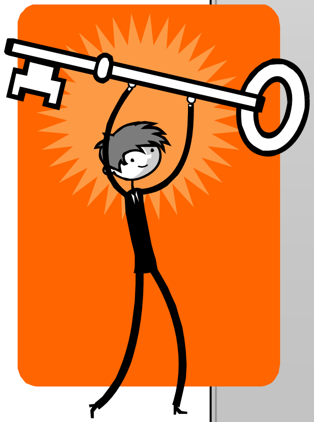
WAW - Discovery Tool

- Helps curators identify domains that are within their collecting scope
- Crawls web sites and extracts domains of possible interest from content
- Maintains lists of domains
- Monitors selected domains for changes

... currently available (May 2005)

WAW - Properties Tool

- Relates content providers to web sites
- Organizes a ‘group’ of web sites hierarchically
- Associates metadata to content providers and, later, to selected content
- Metadata can be subject headings, preferred names, aliases, etc.



… currently available (May 2005)

WAW - Analysis Tool

- Content selection at varying levels of granularity
  - Harvests an entire site or one document
  - Understands serials
- Scheduled harvesting of content
- Shows site structure
- Content automatically associated to content provider’s metadata

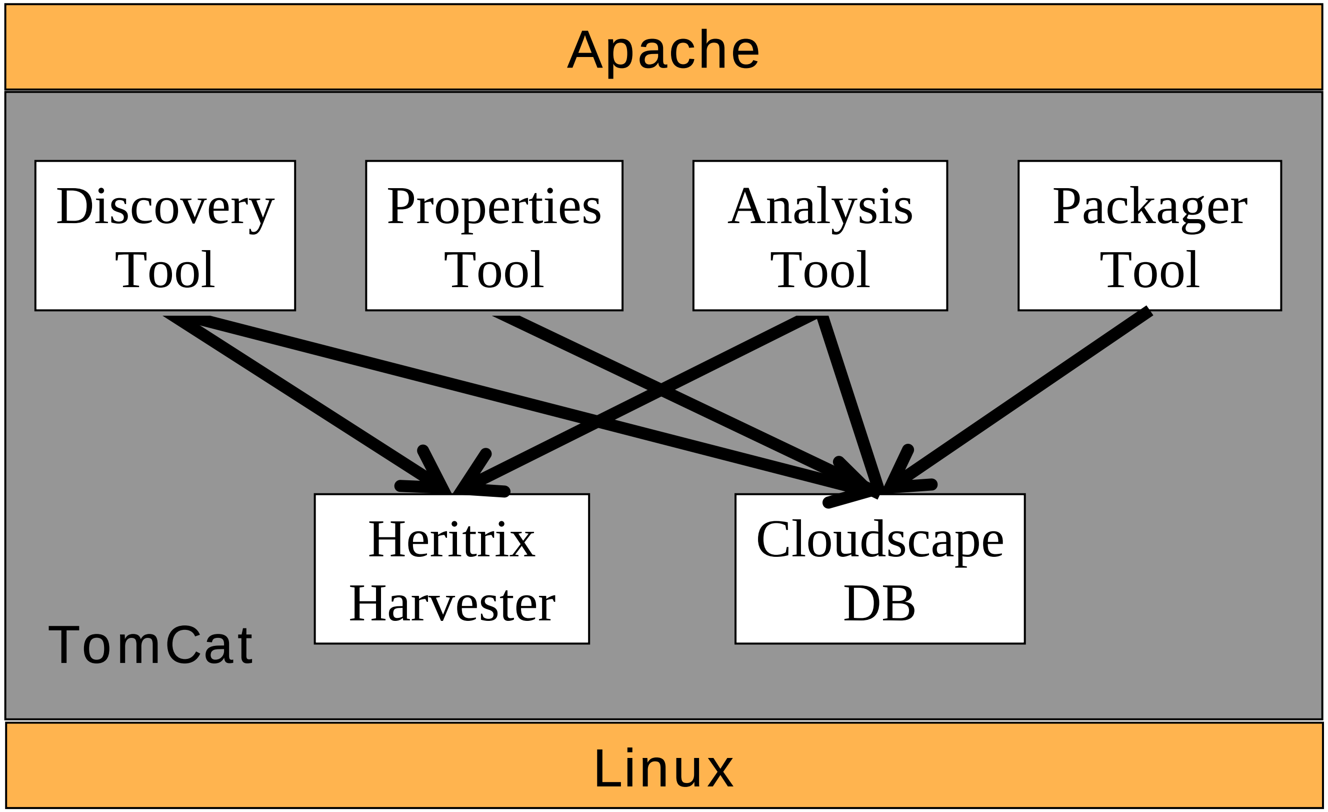
… available January 2006

WAW - Packager Tool

- Packages web content and metadata into an XML standard package (METS)
  - Combines descriptive metadata about content creator, series, and object
  - Creates administrative and preservation metadata
- Neutral format for ingest into OCLC archive and other repositories

... available January 2006

Web Archives Workbench



the
ECHO DEPository Project

A project of the University of Illinois at Urbana-Champaign and OCLC in partnership with the Library of Congress

ECHO DEPository project web site:

http://ndiipp.uiuc.edu

NDIIPP web site:
http://digitalpreservation.gov

Me: Taylor Surface, OCLC
taylor_surface@oclc.org