Digital Libraries and Long-term Access for the Rest of Us: The University of Alabama Acumen Experience

Tom Wilson, Jody DeRidder, Tonio Loewald
University of Alabama Libraries
April 13, 2010
HISTORY

- Digital program began ca. 2004.
- Alabama Mosaic
- Publishers Bindings Online
- Minimal/dormant until 2006.
- Significantly ramped up in 2008.
Along the way

- Staffing changes.
- Large monetary gift.
- Increased pressure to produce.
Meanwhile

- Unhappiness with solution.
  - UI management
  - Metadata management
  - Collection management
  - Software management
  - Licensing cost
- Conversations—what could be!
Whatever goes in, comes out.

URL-addressable collections

Independent utilities

Web services architecture

UI separated from the data

Data stands alone.

We don’t need no stinkin’ ingest!

Light-weight services on top

Get as much online as soon as possible.

Crawlers, trolls, & bots (oh my!)

Automated inferences
Guiding Principles

- Never make something more complex.
- Never develop functions that already exist.
- Sacrifice purity for what works.
- The user experience trumps all else.
- Use a modular architecture & API interfaces.
- Less work is better than more work.
- Digital endeavors are inherently collaborative.
File storage is a lot like a basement closet...

What happens when it's time to move???
A digital object may belong in MANY potential virtual collections...

- Slavery
- African Americans
- Sheet Music
- Tombigbee River
- Southern History
- ... and more

"Gum Tree Canoe," Published by G.P. Reed (Boston: 1847). Wade Hall collection of Southern History and Culture, Hoole Special Collections, University of Alabama Libraries.

... but it originated from ONE SINGLE ANALOG collection. Provenance trumps all!
Hierarchical classes approach

Concatenated meaningful segments in filename; Same segments mirrored in file structure.

- “Holder” and format type
  (example: Hoole Special Collections Manuscripts)
- Collection within those holdings
  (example: Gorgas Family Papers)
- Numbered item within that collection
- Sequence of delivery for pages
  (and subpages)

Archival File: u0003_0000023_0000007_0005.tif
u0003_0001980_0000001 is the first digitized item in the MSS 1980 collection
Create Clarity:
Information at the Applicable Level

Holder/Format

Documentation  Metadata

Collection 1  Collection 2  Collection 3

Documentation  Metadata

Item 1  Item 2  Item 3

Documentation  Metadata

Page 1  Page 2  Page 3
Dropping the Technical Metadata in… where it belongs

The directory for
u0003_0000002_0001.tif

Is simply:

u0003/
  0000002/
    0001/

Makes METS creation a Piece of Cake!
(and redundant!)
Simple, Clear Hierarchical Organization:

University of Alabama LOCKSS Manifest Page

Content from University of Alabama Hoole Special Collections
- Hoole Special Collection Image Collections u0001 Manifest Page
- Hoole Special Collection Text Collections u0002 Manifest Page
- Hoole Special Collections Manuscript Collections u0003 Manifest Page
- Hoole Special Collection Sheet Music Collections u0004 Manifest Page
- Hoole Special Collections Audio Collections u0008 Manifest Page

Other content from the University of Alabama
- University of Alabama Libraries Research Collections u0015 Manifest Page

Content from Nonprofit Organizations
- Nonprofit Organizations Manuscripts u0003 Manifest Page

Content from patron holdings
- Patron Images u0001 Manifest Page
- Patron Text u0002 Manifest Page

LOCKSS system has permission to collect, preserve, and serve.

University of Alabama Hoole Special Collections Images u0001 Manifest Page

The University of Alabama Encyclopedia u0001_1831001 Manifest Page
- Monroe L. Smith III Collection u0002_2007004 Manifest
- Eugene Allen Smith Collection u0003_2007005 Manifest
- Sylvia R. Steege Antebellum Architecture in Tuscaloosa u0001_2007006
- Perkins Family Papers u0001_2007007 Manifest
- Central Iron and Coal Photograph Collection u0001_2007009 Manifest
- Woodward Family photo collection u0001_2007010 Manifest
- Jefferson J. Coleman photo collection u0001_2008002 Manifest
- Educational Media photo collection u0001_2008003 Manifest
- Jubilee Singers of Alabama Collection u0001_2008004 Manifest
- George L. Bryant photograph album u0001_2008005 Manifest
- James Harry Cowan Photograph Scrapbook u0001_2008006 Manifest
- Album of Friendship Collection u0001_2008023 Manifest
- John Hargis, Jr. photo album u0001_2009170 Manifest
- Sesquicentennial Hall Photographs u0001_2008025 Manifest
- Roland McMillan Harper photo collection u0001_2008022 Manifest

LOCKSS system has permission to collect, preserve, and serve this Arcd
Bringing Content Up to the Level Of the WEB!!!

*Greater Usability and Access == Longer Life*

Protected archive area

Web accessible area

Images … ImageMagick:  http://www.imagemagick.org

Audio … LAME:  http://lame.sourceforge.net

OCR … TESSERACT:  http://code.google.com/p/tesseract-ocr/

(it’s free!) 😊
Paper entitled, "The Destruction of the University of the State of Alabama," by Tom S. Birdsong, April 25, 1975

The University of Alabama Libraries Digital Collections

Tom S. Birdsong paper

Abstract
A paper by Birdsong entitled "The Destruction of the University of the State of Alabama," prepared for presentation to the Alabama Historical Association in 1975.

Creator(s)
- Role: Creator
- Name: Birdsong, Tim S.
- Role: Distributor
- Name: Tuscaloosa, Alabama; The University of Alabama Libraries

Date Created
1975 April 25

Physical Description
14 p.

XML schema agnostic
No ingest
No metadata modifications
All content easily accessible
Open to search engines

(also free!)
Tonio Loewald
All we want is a digital library system that’s

• “Good”
• Completely free
• Really easy to use
do as little as possible

corollary: ask as little as possible of everyone else
The Idea

Librarians
Archivists
Search Engines
End Users

MAGIC!
Concrete Goals

- Automatic Indexing
- Automatic Search & Discovery UI
- Automatic Rendering
- No new tools/workflows to learn
- All URLs are stable and readable (no &foo%20bar=dogs%20gonads)
Ingestion

- Really simple
- There is no ingestion
Indexing

Repository (metadata, assets)  Summary XSL

INDEXER

Index Data

Files
Assets
Authorities
Inferring Relationships
Search

- Don’t assume our way is the best or only way
- Expose content to outside search engines
- We might want to index other people’s stuff too
- Even so: do it well
Presentation

- We currently render XML as HTML on the server using XSL

- Interactivity is implemented in JavaScript (with jQuery)

- We can vend XML + XSL direct to the end-user on demand (which we used to do by default) — so we don’t need to use an API to expose metadata, it’s all there
Challenges

• **Debugging** (generally poor debugging tools + scale + unrepeatability — our nastiest bugs have been caused by problems in our own instrumentation)

• Indexing **Speed** (it would be great if the automatic indexing were faster)

• **Resources** (we don’t have people to handle QA, documentation, etc.)

• **Browser incompatibilities** (e.g. lack of HTML5 support, IE6/7 installed base)

• **Feedback & Criticism** (we don’t get enough)
Outcomes (first year)

Indexing

Presentation

Search

Discovery

Admin
Demo
Thank you!

Q&A?

Tom Wilson tcwilson@ua.edu
Jody DeRidder jlderidder@ua.edu
Tonio Loewald taloewald@ua.edu
acumen.lib.ua.edu

Coalition for Networked Information, Spring 2010 Membership Meeting
Roadmap

- Manifest / Sitemap (exposure to search engines)
- Modular / plugin architecture
- User tools (e.g. user tagging, marginalia support)
- Infer metadata from user behavior
- Curatorial Tools
- Improved discovery interface, linkability, content rendering, search interface, admin tools, instrumentation
Assumptions

- Our metadata is (more-or-less) in standard well-defined formats
- Our metadata follows a naming convention
- A_B contains A_B_C (for now)
- Our assets can be converted into standard well-defined formats
- Our assets follow a naming convention
Input Formats

• Import your existing metadata.

• If your existing metadata doesn’t import it’s either a problem in your metadata (fix it) or our XSL (fix that once).

• Don’t need to organize your files in any special way — in fact you can reorganize them without breaking anything.
Technical Dependencies

- PHP 5.x
- MySQL 5.x
- Apache
- jQuery 1.4
- excanvas.js (to emulate `<canvas>` in IE)
- Realbasic (for certain desktop tools — nothing we can’t port to LAMP)
Buzzword Compliance

- Object-Oriented? Mostly. Hoping to be more OO.
- RESTful? Yes, but not by design.
- MVC? Partially — more so in v2.
- Scalable? Yes.

100k files, 30k assets living on a slice of a server with no optimization and no performance issues. Full auto index takes 10h running on the production server with no impact on users. Every component can “live” on a different box, and indexing could easily be split among several boxes.

- Free? As in beer and speech.