Building Digital Archives

Mark Phillips
Cathy Hartman

June 6, 2008

Roadmap

- Steps followed in the development of the UNT Libraries' Digital Library infrastructure.
- Lessons learned along the way.
- Opportunities available today

Beginnings - CyberCemetery

- CyberCemetery
  - Archive created by content being deposited by site owners, or hosting services
  - Early web harvesting with tools like Teleport Pro
  - Content mapped easily to simple file storage structures

Beginnings – Texas Register

Backlogs of files sent to UNT on disks, tapes

Content unpacked, sorted, organized, described and a web presence built around it.

Content stored on Web servers

Simple structures to update

Weekly deposits from agency

Beginnings - Digitization

State Documents

Government Documents

Music Scores

Some done in house, some outsourced

PDF used for final delivery of content

(need for standards was recognized)

Building Infrastructure

We began to recognize that the way we were keeping content archived was not adequate.

Began to investigate infrastructure for digital libraries/archives

We were awarded a grant, wrote a spec, got an RFP out and had some vendors contact us

Infrastructure

Timing coincided with the very first public releases of other popular digital repository tools (DSpace, Fedora)

We chose an open source vendor (IndexData) to work with for the repository.

Developed extensive documentation on our metadata standards that were divided up into Descriptive, Technical and Administrative

Building A System

We had a system delivered which was exactly what we asked for.

Problem was...

We didn't know what to ask for.

Customizing

- Because we chose an open source solution
  - We spent less
  - We were able to hire a programmer
  - We were able to start tinkering with the system
- If we had chosen the commercial product
  - We would have spent more
  - We wouldn't have been able to hire a programmer
  - We wouldn't be able to tinker with the system

What has changed

- Tons...
  - Data Model
  - Metadata Model
  - Granularity of Technical and Preservation Metadata
  - Object Viewing
  - Scaling for large objects
  - Page Turning
  - Audio, Video, Maps, Newspapers, ETDs, Web Content
  - Search system

Current - CyberCemetery

We capture with both HTTrack and Heritrix, one for testing and public display and one for archiving.

Content is stored on disk for access

Content is stored for long term archiving

In the process of planning for new search and access to the archival versions of captured content.

Current – Texas Register

- Still getting weekly updates
- We are looking at migrating content to Digital Library System
  - Which is a preservation environment
- Will provide better searching, display, browsing and access

Current - Digitization

Many projects running concurrently (17+)

All types of content

- Audio
- Video
- Government Publications
- Rare Music
- Yearbooks
- Newspapers
- Maps
- Photographs
- Letters
- Legal Cases

Current – Captured Content

We are getting more electronic publications

- CRS Archive
- Electronic Theses and Dissertations
- Texas Laws and Resolutions
- State, Regional, University Web content
- new projects to come...

Current - Documentation

- Metadata Implementation Document
  - Describes and defines our descriptive metadata set
  - Describes our technical and preservation metadata
  - Contains crosswalks to other popular formats like Dublin Core, Qualified Dublin Core, MODS, MARCXML
  - Used by metadata creators
- Standards for digitization
  - Describes the various specifications used in our lab
- Project and Workflow documentation
  - For most large projects, training documentation is created for the benefit of ourselves and others.

Future - Infrastructure

We are in the process of building a new infrastructure for the deliver of content

There is a new metadata management component also being developed

We are building a system that will scale in the ways we now know we will scale

If we did it again

Groups starting down this road are in a great position.

Many projects/products/tools are now in existence that weren't just a few years ago

There are several choices one can make now about digital archiving

Repository System

Fedora

- Has options for commercial support
- Strong developer and user group
- Foundation created for long term sustainability

DSpace

- Has options for commercial support
- Strong developer and user group
- Foundation created for long term sustainability

E-Prints

- Has options for commercial support
- Strong developer and user groups

Repositories (2)

ContentDM

- Many happy users and user groups
- commercial product with support

Other vendors

- Ex-Libris
- Innovative Interfaces
- Other ILS vendors have systems

Preservation Repositories

aDORe

OCLC Archive

Metadata

- Descriptive Metadata
  - Dublin Core
  - Mods
  - Marc
- Preservation Metadata
  - PREMIS
  - LMER

Metadata (2)

- Content Type Specific Metadata
  - MIX
  - TextMD
  - Audio MD
  - Video MD
  - ALTO
- Wrapper Metadata
  - METS
  - MPEG21 DIDL

Tools/Models

- METS Toolkits
- Format identification
- Globally Format Registries
- Preservation Assessment Tools
- Open Archival Information System
- Trusted Digital Repositories
- Multiple Identifier Types

Lots of Choices to Make

- But having choices is a great thing
  - You just have to pick something and go with it
- There are large communities forming around Digital Preservation/Archiving
- Mainstream industry is starting to notice the Digital Library/Archive world as a group needing solutions both software and hardware related
- Schools are slowly starting to do more education in this area.

But it is hard sometimes

- To pick a repository that will meet your institutions needs/budget/vision
- To pick the right people
- To pick the right content
- To provide access to content
- To make decisions that might be wrong when a new standard comes along

Questions?