Digital Library 2.0: Trends in Management, Access & Preservation

Tom Cramer
Chief Technology Strategist
Stanford University Libraries
September 22, 2010
Stanford University

“The University of Stanford”?

Leland Stanford
Junior University
Stanford University

- 15,000 students
  - 8,000 graduate
  - 7,000 undergraduate
- 2,000 faculty
- 35,000 total university community
- $3.4 billion annual operating budget
- $17.2 billion endowment
- Roots of Silicon Valley
- One of the world’s leading research universities
First generation digital libraries?
1st Generation Digital Libraries

- Small scale digitization, largely focused on text & images
- Purpose built systems for specific content types – application focus
- Highly theoretical approach to digital preservation
- Institutional focus
- Anemic UI’s
2nd Generation Digital Libraries

- Large scale digitization
  - With more content types
- Multi-pathway workflows
  - Content use & reuse in an integrated environment
  - Pragmatic approach to digital preservation & full lifecycle of objects
- Cross-institutional structures
- Infrastructure & service focus
Specific Trends

- Digitization
- Born digital materials
- Interinstitutional structures
- Digital library architecture
Specific Trends

- **Digitization**
- Born digital materials
- Interinstitutional structures
- Digital library architecture
Digitization Trends -- Drivers

- Boutique → Large scale
- Text & image → text, image, audio, video, software and more
- Refresh of 1st generation delivery systems with contemporary UI’s
Digitization Trends -- Responses

Replacing individual, handwrought schemes with workflow-based systems, largely automated, with QA, exception handling and reporting *that work for multiple content streams.*

Management of full lifecycle of object, from physical object management through capture, preservation & access
Lifecycle Management = Integration
Digitization & file processing are the easiest parts of any digitization initiative. Description, file management, collection management, access, and a holistic workflow uniting all pieces, always has and will be the challenge.
Objective:
- download,
- validate,
- manage,
- preserve,
- index,
- mine, and
- deliver
All Google-scanned books.
Specific Trends

• Digitization
• **Born digital materials**
• Interinstitutional structures
• Digital library architecture
Born Digital Materials

• Legacy computer media
  • CD, DVD, tape, hard drives

• Bulk digital collections
  • Licensed, legacy, digitized

• Electronic deposit of singleton files

• Personal digital archives
  • Cloud based files
Legacy Media & Digital Forensics

- Files, operating systems & software
  - mss, correspondence, images, records, data, etc.

- Steps:
  - Extraction
  - Forensic analysis
  - Archival processing & description
  - Access & emulation

- Paradigm shift for archivists, donors
Stanford’s Legacy Media Counts

<table>
<thead>
<tr>
<th>Year</th>
<th>Total # of Collections</th>
<th>Collections with Handheld Media</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>347</td>
<td>2</td>
<td>0.570%</td>
</tr>
<tr>
<td>2001</td>
<td>352</td>
<td>4</td>
<td>1.130%</td>
</tr>
<tr>
<td>2002</td>
<td>346</td>
<td>2</td>
<td>0.570%</td>
</tr>
<tr>
<td>2003</td>
<td>342</td>
<td>7</td>
<td>2.040%</td>
</tr>
<tr>
<td>2004</td>
<td>356</td>
<td>10</td>
<td>2.800%</td>
</tr>
<tr>
<td>2005</td>
<td>382</td>
<td>7</td>
<td>1.830%</td>
</tr>
<tr>
<td>2006</td>
<td>337</td>
<td>17</td>
<td>5.040%</td>
</tr>
<tr>
<td>2007</td>
<td>338</td>
<td>23</td>
<td>6.800%</td>
</tr>
</tbody>
</table>

More than 20,000 handheld media objects in Special Collections alone
Stories of AAAI—Before the Beginning and After: A love letter
SALT: Digital Archives

- Archiving unstructured and semi-structured data
- Allow access to semi-processed information,
  - with strong access & visibility controls
  - leveraging full text & entity extraction
- Ongoing enrichment of the archive
  - through self-annotation by the donor
  - through crowd-sourcing description and organization
Electronic Theses and Dissertations

Submit your dissertation to the Stanford Digital Repository

Dissertation ID: 0000000060

Read-only administrative view
This is a read-only view of the student's ETD submission

Submission approved
Your dissertation has been accepted by the University and is currently being processed by Stanford University Libraries.

» Download your submitted file (including the copyright and signature pages).

When processing is complete and it is released for online access, the dissertation will be accessible from http://purl.stanford.edu/rb7f9bf2333. External access (i.e., from outside Stanford) to the link may be limited pending any release delay applied at the time of submission.

1 Verify your citation details
This information establishes how your work will be cited, and how you will be credited.

Confirm: Yes, the information in this section is correct and complete.

If the information that appears here is incorrect, please consult with your department in order to correct the data in Axess.
Electronic Theses & Dissertation (ETD)

- Automatic deposit to library as part of degree conferral
- Built in digital collection building
- Better access for patrons
- Reduced expenses for students, University, library processing
- Increased visibility of and access to Stanford research via catalog & Google
- Built in preservation through Stanford Digital Repository
EEMs: Accessioning Born Digital Materials

Browser widget enables selector to capture the PDF, plus URL, title, author, copyright status, payment information, and comments, and route to Acquisitions.
Establishing a Digital Library

PURL:
- http://lib.stanford.edu/pasig

Language: eng

Created by: Michael Keller (person)

Direct link to PDF:

Local copy of PDF:

Citation/comments:
- White paper on the contents, technology and process of establishing a digital library, based on the experiences of Stanford's University Librarian in the 1990's and 2000's.

Copyright: Requires request

Purchase: Free

Notifications:

Action log

Dashboard enables item processing, ultimately leading to preservation in SDR and access via the catalog.
Specific Trends

- Digitization
- Born digital materials
- Digital Preservation
- Interinstitutional structures
- Digital library architecture
Interinstitutional Structures

• The institutional repository (IR) is dead
• Resource-, regional- and domain-specific repositories are thriving
  • Hathi Trust – text
  • Internet Archive – web crawls
  • Flickr, YouTube – images & video
  • NanoHub, CrystalEye, ArXiv – sciences
  • DuraSpace – Fedora repository
Digital Medieval Manuscripts
Parker -> DMS Interoperability
Growth in Disk and Computing at SULAIR
Specific Trends

- Digitization
- Born digital materials
- Interinstitutional structures
- Digital library architecture
Stanford’s Digital Library Ecosystem
Three Spheres: Management, Preservation and Access
Deposit & Management

Discovery & Delivery

Preservation
Component Based Architecture

Fedora as a metadata store
Well structured file system as data store
Solr index for rapid data access
Blacklight & Hydra: app logic & presentation

Atomic Services

“Robots”: simple, autonomous scripts, providing small units of work in reusable packages

“Services” provide common operations that support workflows across the environment

“WorkDo”: lightweight workflow to orchestrate cascade of services
Blacklight at Stanford

Next generation catalog
IR discovery (ETD’s and more)
Digital manuscript search
Digital archival collection front end
Repository reporting interface
Hydra

- Reusable solution for building user-facing digital library & institutional repository services

- Use cases:
  - Electronic Theses and Dissertations
  - Born digital electronic library materials (EEMs)
  - Open Access articles
  - Research data
  - SALT - Digital archival “papers”

- Bonus: provide a user interface to digital library systems:
  - Repository administrative interface
  - Digitization workflow UI & metadata editing tool
Hydra

- Joint development project among Stanford, University of Virginia, University of Hull and Fedora Commons
- Based on Fedora, Active Fedora and Ruby on Rails
- Reuse Blacklight for search & browse within a hydra application
Separate Solutions, Same Components

- Tremendous overlap in components across the solutions
  - Blacklight UI
  - Hydra Management Tools
  - Indexers
  - Shelvers
  - Workflow Service
  - ILS Integration
  - Preservation Core

- Slight variations in views, models and workflow tailor the systems to different content types and contexts
Google Books End-to-End Workflow

Objective:
- download,
- validate,
- manage,
- preserve,
- index,
- mine, and
- deliver

All Google-scanned books.
Complex Systems from Atomic Pieces

Phoenix Project: End-to-End Environment

Stage 1: Download Prep
Stage 2: Download & DOR Management
Stage 3: SDR
Stage 4: Discovery/Delivery

Diagram:

1.1 request to GRIN for all books available
1.2. analysis & selection of titles to download
1.3 create list of barcodes

2.1 register object
2.2 conversion requests
2.3 create descriptive metadata
2.4 download
2.5 create page objects
2.6 create PDF
2.7 request deposit
2.8 ready to shelve
2.9 cleanup

3.1 get deposit
3.2 ingest
3.3 confirm

4.1 get queue
4.2 get contentMap
4.3 create index
4.5 search text
4.6 detect text
4.7 read book online

SearchWorks Bib Index
SearchWorks App
SBooks Full Text Index
SBooks App

User

Richard DB
SURI
Page Map
PDF Maker
Make METS
Symphony

MARc to SOLR
process
update index

DOR
Robot

process
Conclusion

The next generation of Digital libraries will be complex ecosystems made up of simple components.

Separate systems for digitization, management, preservation and access will enable pieces to be mixed and matched, supporting content streams from a variety of sources, access by a variety of communities, services and tools, and a network of global network of resources.

Photo by Alun Salt. Used under CC Attribution-ShareAlike 2.0 Generic license.