Concocting a Modern Chimera

Lessons in Building a Digital Library from One of the World’s Great Research Universities

Tom Cramer
Associate Director
Digital Library Systems & Services
tcramer@stanford.edu
March 2009
Overview

1. Stanford
2. SULAIR
3. Digital Library, a definition
4. Six Lessons
5. Final Thoughts
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
The Stanford Libraries

8.5 million volumes in the Stanford Libraries
75,000 current serials
280,000 cartographic resources (maps, images)
59 million manuscript pages across Special Collections
120 countries from which we acquire
400 languages represented
110,000 new volumes acquired each year
#2 in value on the list of University assets
One Definition of a “Digital Library”

A knowledge enterprise that comprises:

1. substantial **online collections**
2. the **tools, environments and services** necessary to discover and use information resources
3. the back office **infrastructure** required to acquire, describe, preserve, manage and deliver digital assets
4. a staff of **information professionals**, that is ready, willing and able to operate in a digital information landscape
The Digital Library: Content, Services & Infrastructure

- Ebooks
- Google Books
- Digitized mss, texts, images, media
- Born-digital materials (data sets, theses, articles, new media, etc.)

- Services
- Discovery
- Delivery
- Use
- Analysis
- Annotation
- Citation
- Collaboration
- Publication

- Infrastructure
- Digital preservation infrastructure
- Content management & content middleware services
- Security, DRM
- Server, storage and data center facilities
Lesson 1

A digital library is like a three-legged stool: it must have all its legs to stand.
Digitization

• 10+ Years of experience with digitization

• Books, images, maps, manuscripts, audio, video, software

• Google partnership has shifted internal focus to rare materials and special collections

• Demonstrated need for holistic approach
The Digital Library: A Schematic View

Content Streams
- Digitized Materials
  - Images
  - Videos
  - Text
- Born Digital Content
  - Research data sets
  - Faculty & student writings, publications
- Purchased & Licensed Content
  - Books
  - Journals
  - Databases
- Federated Content
  - Resource content sharing agreements
  - Public domain materials

Services
- Delivery
  - Books, images, non-musical media streaming (iTunes, etc.)
- Analysis
  - Text & data mining
  - Information extraction
  - Printing & processing
- Collaboration
  - Data sharing
  - Group research & writing environments
- Discovery
  - Refining
  - Searching
  - Scanning
  - Agents & alerts
- Preservation
  - Archiving
  - Preservation
  - Dataset, research article hosting & citation services
- Teaching
  - Course management
  - E-learning environments
- Publication
  - Online indexing
  - Personal publications (blogs)

Infrastructure
- Digital Library Infrastructure (Content Management & Software)
  - Identity
  - Access
- SDR (Sandford Digital Repository)
  - Storage
  - Servers
  - Hardware infrastructure
  - Network
  - Data Center
5 Pillars of Stanford’s Digital Library

1. Digitization
2. Digital Resource Management
3. Digital Preservation
4. Discovery
5. Delivery
Lesson 2

A digital library is not a system. It’s an ecosystem.
The Challenge
Strategy 1: Hardware Approach

- Format Registries
- Authn
- Authz
- DRM

- Annotation Tools
  - Data Mining
  - Text mining

- Analytic Environments

- Image servers
  - Geospatial apps
  - Media players

- Delivery Apps

- Visualizations
  - Semantic indices
  - Full text indices

- Discovery Apps

- Delivery Apps

- Image servers
  - Geospatial apps
  - Media players

- Geospatial apps

- Media players

- Digitization Workflow

- Google BookSearch

- SFX

- ERM

- ILS

- Aquifer
  - ADL
  - Etc.

- Federated Content Stores

- SDR

- collections.stanford.edu

- SULAIR LOCKSS

- Sakai

- ebrary

- Luna

- EEMS
Strategy 2: MacGyver It

- Annotation Tools
  - Data Mining
  - Text mining

- Image servers
  - Geospatial apps
  - Media players

- Visualizations
  - Semantic indices
  - Full text indices

- Delivery Apps

- Discovery Apps

- Analytic Environments

- Delivery Environments

- Discovery Environments

- Semantic Environments

- Full text Environments

- Strategy 2: MacGyver It

- Format Registries

- Authn

- Authz

- DRM

- SFX

- ILS

- ERM

- Digitization Workflow

- Google BookSearch

- Aquifer

- ADL

- Etc.

- Federated Content Stores

- SDR

- collections.stanford.edu

- ebrary

- Luna

- SULAIR LOCKSS

- Sakal

- EEMS
Strategy 3: Library Middleware

Content & Service Middleware - conceptual boundary

Access Broker

Collections Registry

Content Registry

Reporting

Intelligence

Monitoring

Format Registries

Authn

Authz

DRM

Annotation Tools

Data Mining

Text mining

Analytic Envron

Image servers

Geospatial apps

Media players

Delivery Apps

Visualizations

Semantic indices

Full text indices

Discovery Apps

ERM

ILS

SFX

ERM

Google BookSearch

Content & Service Middleware - conceptual boundary

Digitization Workflow

SDR

collections.stanford.edu

Sakai

EEMS

ebrary

Luna

SULAIR LOCKSS

Ebrary

Lockss

Sakal

Aquifer ADL

Federated Content Stores

SDR

collections.stanford.edu

Sakai
An Ecosystem of Activities to the Digital Library

**Management Components**
- Deposit
- Capture
- Subscribe
- ILS (Integrated Library System)
- DOR (Digital Object Registry)
- Monitoring
- Workflow
- Reporting
- Maintenance
  - Preserve
  - Serve

**Discovery Environments**
- NGDE
- Fed Search
- Alerts
- OPAC
- Music
- GIS
- Gov Docs
- EAD
- Art
- Soc Sci

**Delivery Tools**
- Text
- MSS
- Web
- Data
- Image
- Media

**“Portals” and Platforms**
- CMS
- Collab
- Pubs
Lesson 3

If you can’t find it, you can’t use it.
Vast Information Resources

5,825,821

5 million monographic and serial works

25,000 E-journal subscriptions

800,000 maps & images

800 Licensed databases

20 Different libraries

1 Department of Special Collections
Search is not the problem.

*If you know what you are looking for, why are you looking?*

*And if you do not know what you are looking for, how can you find it?*

- Old Russian Proverb
Discovery is the need, specifically:

Relevance ranking, across
...heterogeneous information sources, along with
...browsing across works, and
...browsing within a work.
Discovery Is the Need.
**Search Results**

Search works (SULAIR) - Stanford University Libraries

**Search Query:**
- **slovakia**
- **Everything**

**Access:**
- Online (19)
- At the Library (14)

**Library:**
- Green (Humanities & Social Sciences) (10)
- Off-campus (SAL3) (5)
- Stanford University Libraries (5)

**Format:**
- Journal/Periodical (10)
- Book (7)
- Newspaper (2)
- Microformat (1)

**Number (Library of Congress):**
- J - World History (5)
- P - Language (2)
- B - Philosophy, Psychology, Religion (4)
- H - Social Sciences (1)
- J - Political Science (1)
- more...

**Author:**
- Bořa, Peter (1)
- Domenová, Marcela (1)
- Medvec'ky, Matej (1)
- Polák, Milan (1)

**Showing 1 - 19 of 19 results for slovakia**

### Národná obroda
- **Title:** Národná obroda (Bratislava, Slovakia : 1990)
- **Published:** 1990
- **Online Access:** Content available online via www.narodnaobroda.sk
- **Location:** Loading
- **Call Number:** Loading
- **Status:** Loading
- **Format:** Newspaper

### Slovensko; kultúro-spoločenský mesačník.
- **Published:** 1977
- **Online Access:** Find it @ Stanford
- **Location:** Loading
- **Call Number:** Loading
- **Status:** Loading
- **Format:** Journal/Periodical

### Pamiatky a múzea.
- **Published:** 1991
- **Online Access:** Find it @ Stanford
- **Location:** Loading
- **Call Number:** Loading
- **Status:** Loading
- **Format:** Journal/Periodical

### Slovenský národopis
- **Published:** 1953
- **Online Access:** Find it @ Stanford
Lesson 4

Appearance matters: Libraries need the equivalent of “the digital stacks”.
Delivery: Building the Digital Stacks

Need tools and environments for:

- Image Viewing
- Media Streaming
- MSS Viewing
- Book Viewing
- GIS Software
- Data delivery

In the not too distant future, will also need tools/interoperability for using these resources for:

- Annotation
- Analytics
- Collaboration
- Publishing
Delivery: Manuscript Gallery
Delivery: Image Gallery
Delivery: Text Delivery

Contents

Abstract vii
Currency and Equivalent Units viii
Executive Summary ix

Chapter 1: The Challenge of Recovery 1
Slovakia Under the Federation 2
Dissolution of the Federation 3
Monetary and Customs Union 4
The Velvet Divorce 5
Coping with Shocks 7
Loss of Markets: the Collapse of Foreign and Domestic Demand 7
Adjusting to a Loss of Fiscal Transfers 11
Monetary Policy: Credibility and Crowding Out 16
Establishing Credibility 16
Crowding Out 17
Prices, Wages, and the Exchange Rate 18
Restructuring for Recovery 20

Chapter 2: The Financial Sector 25
The Structure of the Financial System 26
DEWI: An Example of a Data Delivery Service

<table>
<thead>
<tr>
<th>Dataset Title</th>
<th>ICPSR</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pilot National Asian American Political Survey (PNAAPS), 2000-2001</td>
<td>3832</td>
<td>05-20-2004</td>
</tr>
<tr>
<td>Political Regimes and Regime Transitions in Africa, 1910-1994</td>
<td>6996</td>
<td>02-25-2004</td>
</tr>
<tr>
<td>Polity III: Regime Type and Political Authority, 1800-1994</td>
<td>6695</td>
<td>01-23-2003</td>
</tr>
<tr>
<td>Roll Calls of the Continental Congresses and the Congresses of the Confederation, 1777-1789</td>
<td>7537</td>
<td>10-22-2003</td>
</tr>
<tr>
<td>Schools and Staffing Survey (SASS), 1987-88</td>
<td>98-312</td>
<td>06-01-2000</td>
</tr>
<tr>
<td>Schools and Staffing Survey (SASS), 1990-91</td>
<td>98-312</td>
<td>06-05-2000</td>
</tr>
<tr>
<td>Slave Hires, 1775-1865</td>
<td>7422</td>
<td>10-15-2003</td>
</tr>
<tr>
<td>Slave Sales and Appraisals, 1775-1865</td>
<td>7421</td>
<td>10-15-2003</td>
</tr>
</tbody>
</table>
The Power of Digital Resources: Second Life
Lesson 5

Preservation doesn’t come easy.
SDR Serves As Common Preservation Infrastructure

while specialty archives and applications provide focused digital content collection, access and value-added services

- National Geospatial Digital Archive (NGDA)
  Geospatial data

- SUL Digital Bookshelves
  (Google Books, internally digitized, vendors' e-books)

- Digital Library Applications
  (images, mss, media, Special Collections showcases)

- Institutional Repository
  (faculty- and student submitted papers, data, websites, etc.)

Stanford Digital Repository (SDR): content agnostic, preservation repository
### SDR 1.0

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of objects:</td>
<td>&gt;18,000</td>
</tr>
<tr>
<td>number of files:</td>
<td>&gt;2,000,000</td>
</tr>
<tr>
<td>TB of unique content</td>
<td>33 TB</td>
</tr>
<tr>
<td>TB of managed data</td>
<td>~105 TB</td>
</tr>
<tr>
<td>number of content streams</td>
<td>4</td>
</tr>
<tr>
<td>number of collections</td>
<td>14 collections</td>
</tr>
<tr>
<td>number of access events</td>
<td>6</td>
</tr>
<tr>
<td>throughput</td>
<td>2.1 TB/wk</td>
</tr>
<tr>
<td>versions of software</td>
<td>7 major releases</td>
</tr>
<tr>
<td>number of hardware migrations</td>
<td>3.7</td>
</tr>
</tbody>
</table>

**Investment:** 4 years, team of ~6 FTE, lots of custom code
SDR 2.0

- Under development now.
- Refactor all existing code.
- Add / significantly enhance administration and access modules.
- Adopt Fedora as metadata management layer.
- Dramatically simplify and standardize the data model.
- Increase throughput, especially of the storage system.
Lesson 6

Be good to your IT provider.
Growth in Disk and Computing at SULAIR

Final Thoughts

“You ain’t seen nothing yet.”

• *Everything* we’re developing today will be gone within 10 years.

• Born digital onslaught is coming, and it’s going to make digitized materials seem like child’s play.

• “If we don’t hang together, we will surely hang apart.”
Thank you.