Minnesota Digital Library and HathiTrust

Prototype an Image Preservation Archive

5 April 2011
CNI Spring Task Force Meeting
John Butler, University of Minnesota
John Weise, University of Michigan
Eric Celeste, Consultant, MDL
Minnesota Digital Library, Butler
HathiTrust’s Interest and Role, Weise
Prototype and Lessons Learned, Celeste
Minnesota Digital Library Coalition

- Conceived in 2001
- Early years – LSTA paycheck to paycheck
- Minitex (UMN) now the administrative home
- Signature project — Minnesota Reflections
- 62K images, maps, documents
- 120 cultural heritage institutions
MDL Access
“...not possible without you”
MDL Access

“...we’ve got ours and we’re keeping it”
MDL Preservation

...something we can all agree upon

- Viewed as *common infrastructure*
- Economies/imperatives of scale
- Access need not be affected
  - but might it be?
- Attracting broader interest
  - including public radio & television
- A gift from Minnesotans
  - Arts and Cultural Heritage Funding
- A digital MLAC for the state
The Minnesota Digital Library (MDL) seeks to explore a **common infrastructure strategy** that will bring the state a **significantly enhanced capacity** for preserving and accessing its cultural heritage.

To move the discussion from the hypothetical to the practical, we should begin **building a prototype**. It should be **collaborative**, meeting the needs of the primary partners (MDL, UMN, MHS, Minitex) and **extensible to other partners** (e.g., MPR, TPT, county and local historical societies).
Nesting Consortia

HathiTrust

Mission
• Access
• Preservation
• Research foci

Policies
• Governance
• Rights
• Costs

Formats
• Books
• Journals
• Experiment: images & audio

Standards
• File Formats
• Procedures
HathiTrust’s Interest and Role
Contribute to the common good by...

- collecting,
- organizing,
- preserving,
- communicating,
- and sharing

...the record of human knowledge.
Preservation Philosophy

- Maximize...
  - partner contributions
  - use of available resources
Long Term Functional Objectives

- TRAC compliance
- Robust discovery mechanisms
- Open service definition (APIs)
- Support for formats beyond books & journals
- Data mining tools

- All functional objectives... http://www.hathitrust.org/objectives
HathiTrust’s Interest

• Help MDL find a solution
• Explore image support
• Leverage MDL resources
• Empower MDL in the process
• Draft ingest specifications
• Establish a viable model for support of MDL and similar preservation cooperatives.
Design

• Operational logistics at the scale of HathiTrust have led to solutions that favor:
  • consistency and standardization
  • simplicity over complexity
  • practicality over ideology
HathiTrust’s Role

- Scoping and guidance
- Lower barriers
- Raise bars
- Learn, grow and eventually provide
  - Object ingest specifications for images
  - Object preparation tools, esp. validation
All Together Now

- HathiTrust really is a collaborative effort.
- MDL and Michigan worked together under HathiTrust governance.
What did MDL actually send to HathiTrust?
What MDL sent to HathiTrust

<table>
<thead>
<tr>
<th>Packages</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Reflections Simple Contone</td>
<td>22,186</td>
</tr>
<tr>
<td>Reflections Compound Objects</td>
<td>888</td>
</tr>
<tr>
<td>Minnesota Historical Society</td>
<td>6,860</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>29,934</strong></td>
</tr>
</tbody>
</table>

- **Reflections Simple Contone**: 74%
- **Reflections Compound Objects**: 23%
- **Minnesota Historical Society**: 3%

Image of pie chart showing the distribution of packages.
## What MDL sent to HathiTrust

<table>
<thead>
<tr>
<th></th>
<th>Items</th>
<th>GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple JP2</td>
<td>22,186</td>
<td>429</td>
</tr>
<tr>
<td>Compound JP2</td>
<td>13,844</td>
<td>407</td>
</tr>
<tr>
<td>Compound Bitonal TIFF</td>
<td>13,272</td>
<td>1</td>
</tr>
<tr>
<td>JPEG</td>
<td>9,575</td>
<td>12</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>49,302</strong></td>
<td><strong>849</strong></td>
</tr>
</tbody>
</table>
What MDL sent to HathiTrust

Number of items transferred

- Simple JP2: 38%
- Compound JP2: 24%
- JPEG: 16%
- Bitonal TIFF: 23%

Amount of data transferred

- Simple JP2: 51%
- Compound JP2: 48%
- JPEG: 1%
An “object” sent to HathiTrust is a “Submission Information Package” and consists of many parts: a METS file with a variety of metadata, a set of image files, and a set of corresponding text files. The image files are the “items” sent.
What did we learn from working together?
What is a master?
Where is the identifier?
Metadata madness!

Self Portrait.
Trust us.
Who’s looking?
No free lunch.
No free lunch.

<table>
<thead>
<tr>
<th></th>
<th>Time as Producer</th>
<th>Cost as Producer</th>
<th>Time as Aggregator</th>
<th>Cost as Aggregator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programmer</td>
<td>4 weeks</td>
<td>$8,000</td>
<td>14 weeks</td>
<td>$28,000</td>
</tr>
<tr>
<td>Metadata Assistant</td>
<td>100 hours</td>
<td>$2,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Manager</td>
<td>40 hours</td>
<td>$2,800</td>
<td>30 hours</td>
<td>$2,100</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>$12,800</strong></td>
<td></td>
<td></td>
<td><strong>$30,100</strong></td>
</tr>
</tbody>
</table>
Next Steps for MDL?

- Trials with MetaArchive and OCLC Digital Archive.
- Paper and phone evaluations of a few others.
- Decisions in early summer about how to proceed.
Questions?