MDL Digital Image Preservation Prototype

Minnesota Digital Library & HathiTrust

17 March 2011
Library Technology Conference 2011
I am a consultant to the Minnesota Digital Library who has worked on some of the social initiatives of MDL and has recently been helping MDL think through its approach to digital preservation. You can reach me at efc@clst.org with any questions.
We all know MDL as at least Minnesota Reflections. But to get here, we first have to visit the scanner, where we make a digital copy. We scan these at very high resolutions, we capture a lot of detail. Most of that detail is not even show on Reflections most of the time. But in this process we make a “master” image (usually a TIFF or JPEG2000 image) that does keep all the details.
But MDL makes no promise to keep or store this master image. Instead, we burn it to a disc and return it to you for safekeeping. Do you know where all your master images are? Do you know if the discs are still readable, unscratched? (image: http://www.flickr.com/photos/darrenhester/3911531019)
MDL wants to keep the masters safe...
and so do other organizations around Minnesota.

Potentially not just digital image masters, but other formats as well. Audio. Video. Documents. We realized that the state could use a common preservation archiving strategy.
What do we need?

• Capability to handle multiple *formats*.

• Workflow for *partners* beyond Reflections.

• Bit *preservation*, plus some option for metadata and migration.

• Of two minds about *access*... dark? light? dim?

• Clear *governance* model, well defined roles.
Looking at partners in August 2010 we decided to embark on a prototype with HathiTrust. Working relationship with UMN in place already. Interest in moving beyond books. Well developed policy and governance structure.
Here’s what we sent to HT. Some of this data was from Reflections, some of it was from the Minnesota Historical Society Collections Online. A “contone” is a “continuous tone image” or what you would think of as a photograph.
Here’s what we sent to HT. Some of this data was from Reflections, some of it was from the Minnesota Historical Society Collections Online. Note that none of our masters were actually in JPEG 2000 format when we started, more about that later.

<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>
How much of what did we send? It all depends on how you look at it. In terms of objects, 77% Reflections, 23% MHS. In terms of items, 84% Reflections, 16% MHS. In terms of data, 99% Reflections, 1% MHS. Note also the bitonal TIFF images are 23% of items and add up to practically 0% of data! In terms of time spent, Simple, Compound, and MHS were roughly similar, about a month each.
What we sent to HathiTrust

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.

The HT SIP “object” is a way of wrapping together the “items” and their metadata. In fact, in the case of HT, even the items (those image files) had to include metadata (in the Adobe XMP format).
What did we learn from working with HathiTrust?

Looking at partners in August 2010 we decided to embark on a prototype with HathiTrust. Working relationship with UMN in place already. Interest in moving beyond books. Well developed policy and governance structure.
Get the full report at...

http://mndigital.org/projects/preservation/
What is a master?

Master images at local institutions are often not formatted as required by HT and require transformation and the addition of embedded metadata. Many of these master images also lack fixity checks. The format requirements of HT may present a high-bar for potential participants.

Image data of the sort in Minnesota Reflections is quite a bit larger than the page images of books currently found in HT collections. This creates challenges for data transfer and package ingest.
Where is the identifier?

Items in the preservation archive require unique identifiers and that identifier namespace needs continual careful attention as new collections are added. Local institutions should be prepared to share unique identifiers.
Mapping metadata from local systems is difficult to routinize and would require ongoing attention in a long-term preservation effort. Properly packaging items for HT can also be time consuming. The different perspectives of MDL and HT concerning metadata has resulted in differences of directional intent over the project period, some of which have been resolved during this pilot project, and others of which must be revisited before a long-term program is undertaken.

A programmer would be required in any long-term effort to integrate new collections, building scripts to do metadata mapping and packaging of objects.
Do you trust us?

A trusting relationship with well defined responsibilities is required to allow for pragmatic solutions to data transfer since the MDL will likely end up with access to more information than it needs to complete the archiving task.
Local institutions may be more sensitive about image dissemination than HT expects. Rights issues have posed key challenges for MDL and HT during this pilot project. The project partners currently hold different expectations and requirements regarding rights and display.
No free lunch.

This model cost about $1 per image up front and $0.10 per image in ongoing maintenance.
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>

This model cost about $1 per image up front and $0.10 per image in ongoing maintenance.
Next Steps?

- Trials with MetaArchive and OCLC Digital Archive.
- Paper and phone evaluations of a few others.
- Decisions in May about how to proceed.
- Talk to someone on the MDL Management Team if you have input you would like to share.
Questions?