Eric Celeste

Minnesota Digital Library and HathiTrust Image Preservation Prototype Project Report

31 December 2010

with minor edits on 26 January 2011

Prepared by Eric Celeste and Katherine Skinner

Eric Celeste 1993 Lincoln Avenue Saint Paul, MN 55105-1455 651-323-2009 efc@umn.edu http://eric.clst.org
Katherine Skinner Educopia Institute Atlanta, GA 404-783-2534 katherine.skinner@metaarchive.org http://www.educopia.org/

Table of Contents

0.1. Executive Summary

From September through December 2010 the Minnesota Digital Library (MDL) worked with HathiTrust (HT) to add content from Minnesota Reflections and the Minnesota Historical Society (MHS) to HathiTrust as a preservation archive. Nearly 50,000 images from Minnesota Reflections were shipped to HT before Christmas 2010 with another 8,000 MHS images due to be shipped soon after the holiday. HT plans to ingest these images and their metadata in early January.

While building this preservation archive MDL and HT learned a number of lessons:

- 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.

- Items in the preservation archive require unique identifiers and that identifier namespace needs continual careful attention as new collections are added.

- 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.

- 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.

- 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.

- 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.

- Once descriptions are ingested into HT, only the catalog information can usually be changed.

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

The delays imposed by the precise requirements of packaging pushed ingest into late-December, early January, well past the time this report was prepared. The lessons of ingest will have to be evaluated by the team once ingest has been completed. The lesson for the team is that whether the goals are accomplished or not, projects do have to come to an end. Both MDL and HT staff plan to continue work on ingest and retrieval into January without the project manager or preservation consultant directly involved.

Preservation services operate on a continuum from bit-level services to “full” preservation services that include the maintenance of a fully operational access-oriented content catalog. HT is in the highest end of this continuum. In this pilot project, MDL has explored the processes and workflows that would need to be actualized by a wide range of MN institutions in order to participate in HT preservation services. As MDL and HT continue to explore a potential longer-term relationship, it might be helpful to share pilot findings with representatives from across this range of institutions to see how their needs and abilities match up with this preservation service model.

To date, MDL has coordinated and hosted Minnesota Reflections and in this context, has worked with Minnesota-based institutions in a relatively informal manner. MDL now seeks to offer a new program consisting of preservation services to a Minnesota-based constituency. This project’s Sponsors Group has agreed that to offer preservation services, MDL will need to create an entity with (or perhaps build into MDL) a higher level of formality in its governance, policies, and documentation than has previously been engaged in the Minnesota Reflections project.

0.2. Background

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. The MDL senses a common need and opportunity in providing large-scale digital content repository services for Minnesota, and considers establishing a shared digital preservation service a valuable initial goal.

As stated in the summary of a January 2010 meeting of stakeholders: "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)."

MDL began this work by conducting a detailed digital preservation needs assessment with requisite initial focus on image data. Consultant Eric Celeste was contracted to conduct the assessment, which involved inputs from numerous current and prospective stakeholders and concluded in July 2010 with the final report, MDL Digital Preservation Demonstration Project: Digital Image Preservation Needs. This report positioned the MDL to take the project into its next phase – prototype and demonstration – which is the focus of the project described in this report.

The purpose of this project was to pilot and, therefore, demonstrate, the technological and organizational potential of a scalable digital preservation program and service for cultural heritage stewardship across Minnesota. We worked with a reputable partner, the HathiTrust Digital Library, through the auspices of the University of Minnesota's partner status with HathiTrust, and a well-scoped focus on the ingest of image data.

As the project launched, a project charter expressed the following assumptions:

1. HathiTrust has developed processes and standards for a particular set of partners dealing with a limited variety of digital content. This project tests the potential of HathiTrust in a new collaboration, with a different set of partners providing a different type of content.
2. To be successful, the project has to attempt to reconcile the HathiTrust's requirements and the MDL's capacities, determining what would be sufficient, sustainable and extensible over a longer term relationship.
3. The project is also the catalyst for an evaluation of the MDL's governance structure. The current framework is not adequate for a more ambitious and more complex digital preservation program.
4. HathiTrust as the solution for MDL is subject to evaluation along the lines of technological, organizational, and economic fitness for purpose.

0.1.1. Selected Acronyms

ACHF: Arts and Cultural Heritage Fund, also called “Legacy” funds

AIP: Archive Information Package of the Open Archival Information System reference model

DIP: Dissemination Information Package of the Open Archival Information System reference model

HT: HathiTrust

JP2: JPEG2000 is an image compression standard

MDL: Minnesota Digital Library Coalition

METS: Metadata Encoding and Transmission Standard, a kind of XML wrapper for all kinds of information about objects

MHS: Minnesota Historical Society

MPR: Minnesota Public Radio

PREMIS: PREservation Metadata: Implementation Strategies, a strategy for describing events in an objects history using XML, and in our case included as part of the METS file

SIP: Submission Information Package of the Open Archival Information System reference model

TPT: Twin Cities Public Television

UMN: University of Minnesota

XML: Extensible Markup Language, a syntax for the exchange of structured data

XMP: Extensible Metadata Platform, for storing information relating to the contents of a file

0.3. Project Participants

Sponsors of this project included: John Butler, AUL for Information Technology, University of Minnesota-Twin Cities; Bill DeJohn, Director, Minitex; Bob Horton, Director, Library, Publications & Collections, Minnesota Historical Society; and John Weise, Head, Digital Library Production Service, University of Michigan, and on behalf of HathiTrust.

The project manager was Eric Celeste, a consultant in Saint Paul, Minnesota. The digital preservation consultant was Katherine Skinner, Executive Director, Educopia Institute. Eric and Katherine are responsible for this report.

The primary developer on this project was Bill Tantzen, a software developer at the University of Minnesota Libraries. His supervisor, Jason Roy, Digital Content and Software Development Coordinator, University of Minnesota, was also a key member of the development team. Greta Bahnemann provided support for Minnesota Reflections.

As the project progressed, a number of staff from the University of Michigan Libraries joined the effort to represent the interests of HathiTrust. These included: Aaron Elkiss, Shane Beers, Tim Prettyman, Jeremy York, and Chris Powell.

As we embarked on the Minnesota Historical Society stage of the project, Karen Lovaas and Jane Wong of MHS also joined the team.

1. Execution

The project got underway in September 2010. The team extracted close to 40,000 master images from Minnesota Reflections and 8,000 images from the Minnesota Historical Society, converted them into formats suitable for HathiTrust, built submission information packages for them, and transferred them to Michigan. The HathiTrust staff are still working on ingesting these images, after which the Minnesota Digital Library will retrieve some to verify that a “round trip” is possible.

This section of the report details the process, technology, costs, and project management tools that were part of the project.

1.1. Process

The project timeline was ambitious, attempting to squeeze the whole prototype effort into just the last few months of 2010. MDL's goal was to develop the workflow to move digital image data and associated metadata from Minnesota into the HathiTrust, demonstrate that workflow by moving a defined set of images into HathiTrust, and work with HathiTrust to define the appropriate display for these images in that system. The team was given from September to the end of November to accomplish this goal, with December to report on progress. That schedule ended up slipping a few weeks, and not every goal was accomplished, but many lessons are nonetheless evident.

The project team settled on a workflow with six stages of processing: (1) **extracting** master images and metadata from current repositories, (2) **reformatting** the images to suit HathiTrust requirements, (3) **packaging** these binaries and associated metadata as required for ingest, (4) **transferring** these packages to HathiTrust, (5) **ingesting** these packages at HathiTrust, and providing for (6) **display & retrieval** of these images from HathiTrust. The first four of these were accomplished before this report was written, the latter two were not completed.

MDL wanted to address a variety of content types during the prototype, from simple continuous tone images, to compound objects made up of a series of images in a certain structural relationship, to images containing text and associated optical character recognition (OCR) derived text. Demonstration content would include the roughly 50,000 Minnesota Reflections images and a 10,000 image subset of the MHS collection management system. Note, while newspaper data was initially considered for inclusion, the sponsors agreed during their first meeting that the timeline did not allow for the preparation this would require. The project manager divided the effort into three stages: during *stage one* the team would transfer simple continuous tone images from the Minnesota Reflections system, during *stage two* the team would transfer the compound objects from Minnesota Reflections, and finally during *stage three* the team would transfer a set of images from the content management system at the Minnesota Historical Society. With each stage the developers and HT staff would learn lessons to apply to the following stages.

After a brief time in September during which the program manager shared and revised the workplan and guidelines (available in an appendix) with project team input, stage one launched in October 2010. Rather than recount the sequence of events as executed, this report will describe the workflow of a single "stage" as that workflow evolved.

1.1.1. Content Selection

The first step at each stage was to determine exactly what content would be part of the transfer. In the case of the two Minnesota Reflections stages, this consisted primarily of specifying how the programmer would distinguish the simple continuous tone images from those that were part of compound objects. In the case of MHS the project manager spent some time reviewing available MHS content to identify content that might be suitable for the very short window of opportunity available.

In any case, it is important to note that this prototype effort was very selective and quite narrow in its focus. Minnesota Reflections was well defined and easily accessible to the project team. The Minnesota Historical Society has all sorts of image content not included in this project, the Collections Online at MHS which we did select are some of the least complex, most fully described, and clearly public material available there. A more comprehensive preservation effort would encounter a much more diverse set of material.

Before each stage began the project manager also developed a specification document to guide the programmer and help the HT team know what to expect. These specifications, samples of which are in the appendices, defined in detail how identifiers and metadata would be defined and developed, and how the packages for HT would be structured. Each specification document was based on a set of guidelines the team agreed to at the beginning of the project. The guidelines noted that while the team would have to act as pragmatically as possible to meet the timeline, some of the pragmatic “shortcuts” acceptable for a prototype might have to be reconsidered for a longer-term project.

1.1.2. Extraction

Completing the prototype mission required both the master images from participating collections, and enough descriptive metadata to serve the needs of the HT catalog. The first two stages involved images stored at the University of Minnesota Libraries, making retrieval for the UMN-based MDL programmer straightforward. MHS decided to provide its images on a hard drive. In fact, MHS provided all the master images from its CMS because that was simpler than providing only the 8,000 or so required by the project. They trusted MDL to erase the hard disk when our job was done and not misuse the remaining images. The fact that all participants were so “local” made gathering the images quite simple.

Extracting metadata was a bit more of a challenge. Minnesota Reflections was accessible via the OAI-PMH protocol which provides Dublin Core descriptive metadata. Unfortunately the unqualified DC provided by CONTENTdm was missing many of the subtleties of the original descriptions. It was also difficult to acquire data for compound objects since this was typically excluded from the OAI harvest. Even though the project programmer worked in the same organization as the MDL staff managing the metadata, close coordination was required to pull off the harvests required without disrupting Minnesota Reflections.

In the case of the Minnesota Historical Society, no OAI Dublin Core data was readily available. Instead, MHS staff met with the project manager and allowed him to review their system and its output. Together they decided that the XML output from the cataloging side of their CMS would provide the best descriptive data and could be mapped to DC. This mapping eliminated some of the subtleties of MHS data in ways similar to what was experienced with Reflections data.

Each item also needed a unique identifier for use in building an identifier suitable for HT. Oddly enough, even this simple requirement turned out to be nontrivial. The managers of Minnesota Reflections realized during the extraction for this project that over 3,000 items had been assigned duplicate identifiers. They also found that no identifiers had been created at all for compound objects as a whole, only for their constituent parts. While there were no duplicates at MHS, the team did learn that two separate schemes had been used to identify their records. Choosing which identifier to use and creating new identifiers were unanticipated but vital steps in the workflow since the identifiers would be critical to later ingest into and retrieval from HT.

1.1.3. Reformatting

The biggest adjustment for the preservation project was also the first: HT could not preserve the master images already in our care. The team learned that HT instead required that our continuous tone TIFF images be turned into lossless JPEG2000 format, and even bitonal TIFF images or JPEG images in later stages required new XMP data in the binary file itself.

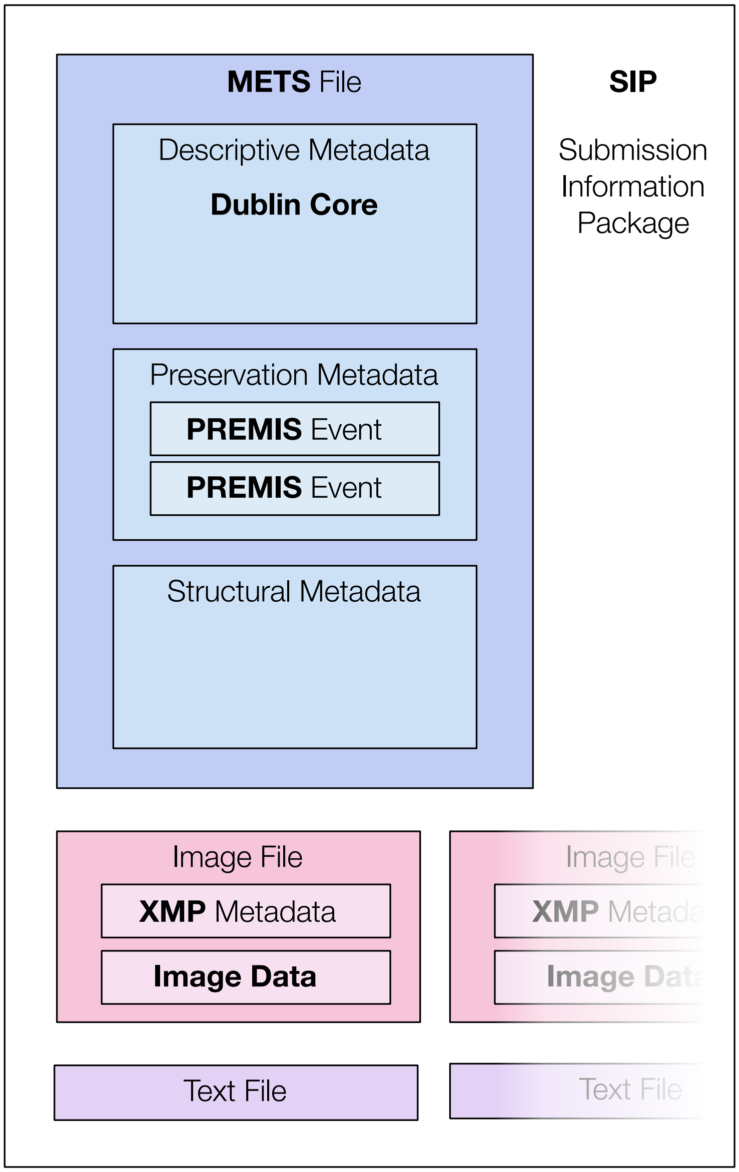
This requirement that every master image be transformed in some way increased the time needed to prepare images, required that we produce a new checksum for each image, and meant that we were preserving something fundamentally new instead of the original master. Some of the archival consequences of this requirement are discussed later in this report.

The requirement for HT-specified XMP metadata embedded in each binary also added a significant step to the workflow. HT has fairly clear guidelines for its XMP metadata, but even so the project programmer got plenty of “notes” about his attempts to create conforming XMP from HT staff. The team learned that it was important to review actual sample transformed images before producing a whole set for shipment to HT, because even the smallest change to XMP would require that the image be re-mastered and have a new checksum generated and stored.

1.1.4. Packaging

HT had not had much experience relying on an outside party to produce the submission information package (SIP) for ingestion into HT. While HT had worked with outside organizations before, HT staff had taken that content and produced the required SIPs. The project team spent a few weeks getting oriented to the guidelines and goals of the HT process. HT has good documentation and clear expectations, but even so the team found enough ambiguity to fuel many discussions.

The HT SIP is a combination of binary objects (the images), metadata (the description), and a few ancillary files (transcriptions or OCR, when available) zipped together into a single compressed (".tar.gz") file. The metadata is in the form of METS (Metadata Encoding and Transmission Standard) including a good deal of PREMIS (PREservation Metadata: Implementation Strategies) and took by far the greatest percentage of team time to work out.



Although HT initially requested that descriptive metadata be provided in MARC format, they quickly accepted unqualified Dublin Core as an alternative. MDL has not built any stores of MARC metadata and it is unlikely that any potential partners in future digital preservation efforts around Minnesota will have MARC descriptions available for this content. While future partners are also not likely to have DC data, at least DC is a much simpler target for mapping efforts. In fact, for the stage three of the prototype project we did have to map an MHS-specific XML format into unqualified DC. Such mapping removes many features of the native metadata, but what remains is sufficient for recalling items from the preservation archive.

While the descriptive DC data becomes one portion of the METS file, PREMIS events make up a much more contentious part. PREMIS events encode preservation events in the life of the object being described. For example, the initial scanning of the object might become a "capture" event while the reformatting from TIFF to JPEG2000 might be a "transformation" event. The team found that the PREMIS data dictionary and past HT practice left much open to interpretation and nuance.

For example, after weeks of struggling with PREMIS "eventOutcomeInformation" the HT staff suggested removing this information altogether. Even a decision by MDL to use MARC21 codes to identify PREMIS "linkingAgents" was a subject of much discussion. As staff become more familiar with METS and PREMIS these decisions might become simpler and even routine, but this project exposed many assumptions being made by the HT staff that needed explanation and discussion before MDL grasped the requirements or HT embraced the differences in MDL practice.

Getting all the details of packaging and metadata correct took by far the greatest portion of staff time during the prototype project. HT started with only one staff member on the project management Basecamp and ended with six staff members taking part in discussions there, mostly about building the SIP. The cost of an error that shipped to HT was very high since we would have to recall the hard drive and re-execute the scripts that generate our image files. Those scripts could take over a day to run. The team learned to share sample runs which allowed for evaluation before the full job runs. Unfortunately, much of the critique came through iteration, making it very difficult to cut determine when it was time to run the full job and produce the final data for HT.

In the end the team spent too much time on packaging, not arriving at the final steps until the holidays were at hand.

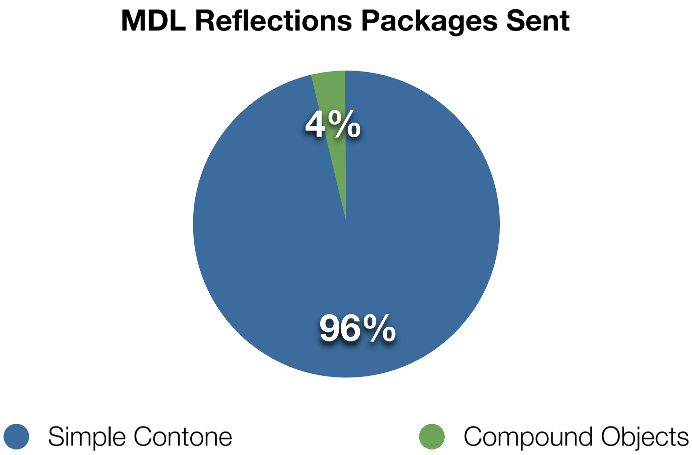
1.1.5. Transfer

Initially the team believed that some of the data transfer could take place over the net, particularly since both UMN and HT have access to the Internet2 network. However, it turned out to be more practical to use hard disks to transfer data.

Shared access to FTP sites and even file sharing through Basecamp was used to transfer samples of metadata and processed binaries. These samples were vastly smaller than the complete data sets, and sharing these samples helped facilitate comments about format and metadata issues.

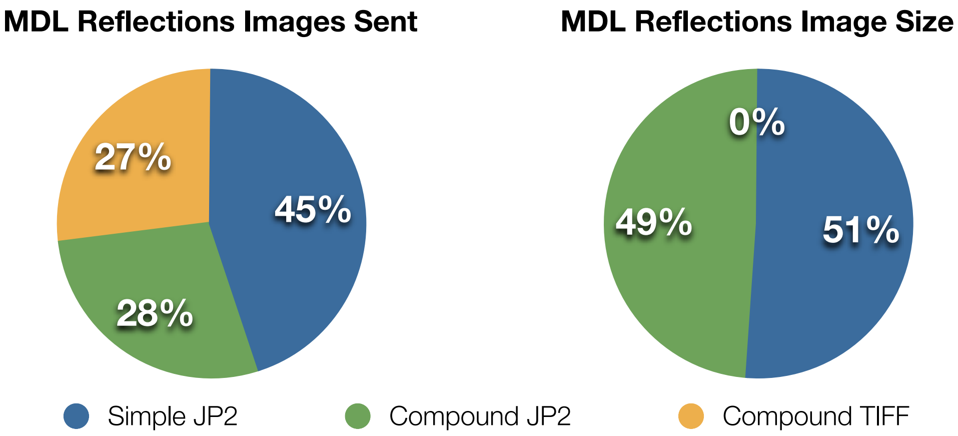
The final data sets were quite large, though. Unlike the book scans common to HT, these continuous tone image files were quite large individually and impractical to transfer in aggregate. In theory even such large transfers should be possible over the net, and big science projects routinely transfer even bigger datasets, but in practice using hard disks simplified the process. The team used one 2TB hard disk that made the roundtrip from Minnesota to Michigan a few times. A hard disk was even used to transfer images and metadata from MHS to UMN.

Shipped before Christmas	Packages
Simple Contone	22,186
Compound Objects	888
Total	23,074



Before Christmas 2010 the team shipped both the stage one “simple contone” (continuous tone) images and the stage two “compound objects” to HT. This made for a total of about 23,000 packages shipped.

Shipped by Christmas	Items	Bytes	GB
Simple JP2	22,186	460,327,111,143	429
Compound JP2	13,844	437,039,815,587	407
Compound TIFF	13,272	1,063,301,376	1
Total	49,302	898,430,228,106	837



Note that the 888 compound object packages made up 55% of the images sent. Some of those images, though, were very small “binary” (black and white) TIFF files, so the actual size of the stage one and stage two transfers was about the same.

1.1.6. Ingest

Unfortunately, as of the end of December 2010 it appears that no MDL data has been ingested in HT. The first attempt to transfer simple continuous tone images in October was rejected by HT because the files sent failed various validity tests. This failure inspired a more rigorous pre-transfer exchange of samples through November and early December. As of mid-December MDL has sent a new set of “stage one” continuous tone images and “stage two” compound objects to HT. The “stage three” MHS dataset will not be sent before the new year 2011 begins. HT does plan to ingest data from all three stages in January 2011.

1.1.7. Display & Retrieval

Obviously, with successful ingest still ahead of us, no retrieval is yet possible either. MDL will be very interested to see if HT can limit display derivatives of images to 1024 pixels on the longest dimension and provide links back to source system records by way of URLs shared as identifiers. MDL will also test to see whether access to retrieval of the master images can be restricted to a finite set IP ranges or account holders.

1.2. Technology

As it turned out, each stage of the project made different demands on the MDL programmer and he ended up developing separate, though related, approaches for each.

1.2.1. Stage One: Simple Contone

The process of creating the METS packages for the contone images can be broken down into three distinct steps.

The first step is the gathering of metadata. For this CONTENTdm hosted content, the programmer used the data as retrieved via an OAI harvest. The program for this step was written in Java because an OCLC library was available that supports OAI-PMH v2.0. The results of the harvest were stored in a MySQL database which was, at its heart, rows composed of identifier, field, and value attributes. The fields were the Dublin Core elements, plus a “local_identifier” field which was useful for locating the image associated with an asset.

The second step prepared the image derivatives. Using the Image::ExifTool libraries, metadata was extracted from the TIFF master files. This metadata was used to produce XMP data which was then attached to the compressed JPEG2000 files created by the freely available Kakadu demo programs.

The third step created a METS file based on the metadata stored in the MySQL database produced in step 1, gathered that file, and the compressed image derivative from step 2 into a directory, and then zipped and compressed that directory creating the final package.

1.2.2. Stage Two: Compound Objects

As data was collected from different sources, ideally only the first step needed to change. For instance, a particular collection’s metadata may be in an Excel spreadsheet or an Access database. The process of converting that data into MySQL tables and rows will differ, but steps 2 and 3 should not differ.

The database for the compound objects was different in that the individual images were associated in a parent-child relationship. In addition to the OAI harvested metadata, data describing the relationship between the records that composed a compound object had to be extracted from CONTENTdm. In this case, the CONTENTdm PHP API was used to expose data that OCLC does not make available through OAI-PMH.

The code from step 2 of the stage one records was leveraged without change, but step 3 differed, because the METS was quite different since it described a compound rather than a simple record.

1.2.3. Stage Three: Mixed JPEG Images from MHS

For the MHS data, the database was again very much the same, and only the process of getting data into it changed. In this case, the metadata was given to us in a single XML file from the MHS EMu system. This file was parsed, and its data inserted as the same series of identifiers, fields, and values.

Step 2 here differed in that there was no need to create image derivatives; the master images were already lossy JPEGs and HT agreed with MDL to leave them as they were rather than converting them to JP2s. However, the same code base and procedure was used to create the required XMP data to attach to each JPEG.

Step 3 differed slightly from the code for the stage two compound objects because the compound objects from MHS were not composed of individual records associated in a parent-child manner. For MHS, compound records were simply records with multiple images.

In each of the collections, the same 3 step process was followed, but because of differences between the collections, the details of these steps did differ. However the template remained the same, and the programmer estimates 80% of the code remained the same between stages. The programmer also believes that as the number of collections grows, and patterns emerge, these different patterns can be abstracted and a single generalized set of programs and scripts may result, at least for the second and third steps.

1.2.4. Catalog metadata

Along with the preparation of the SIPs for HT, the programmer also had to capture all the Dublin Core descriptive metadata that was included in the METS files and build a separate file of catalog metadata for HT. This catalog metadata became the foundation of access to this material at HT, and could also be reloaded later if MDL found corrections were warranted.

1.3. Costs

The costs for this project were simplified by the fact that HT was not charging UMN for participation in the prototype effort. Besides the consulting costs for the project manager and preservation consultant, which are not discussed in this report, there were minor technology costs, a hefty staff commitment by MDL (all UMN staff), and some time committed to the project by MHS.

The technology costs were relatively trivial. They included the cost of a 2TB hard disk and the postage required to ship it back and forth to Michigan a few times over. All workstations and local storage used were otherwise marginal to the work that regularly goes on at UMN.

Likewise, the staff of MHS were only required in the last stage of the project and the team purposely kept the requirements for MHS simple. Even so, two staff participated in weekly developer meetings for the last three weeks of the project, and staff spent time with the project manager reviewing content at MHS and selecting the records that would be used for the prototype effort. Finally, MHS staff had to prepare both metadata and binary exports from EMu from which MDL could build SIPs for HT.

The most significant expense of the project (outside HT's own commitments) was the staff time committed by MDL. Three UMN staff members were regular participants and their roles included both that of an information producer (exporting data from Minnesota Reflections) and information aggregator and transformer (preparing SIPs for HT).

The programmer spent full time on this effort. Given his salary scale and the occasional encroachment of other duties, this can be estimated to be a roughly $2,000 per week expense. The first few weeks were dedicated to extracting data from Minnesota Reflections into a database sandbox from which he could produce the SIPs, so it would probably be roughly fair to allocate four weeks of the programmer's time to the producer role for Minnesota Reflections. The remainder of his time, including the slip into December and January as the HT requirements have become more precise, should amount to another 14 weeks allocated to the aggregator and transformer role.

The metadata assistant spent approximately 100 hours at roughly $20 per hour on data remediation related to the extraction of data from Minnesota Reflections. This work included helping to resolve the 3,000 duplicate identifiers accidentally introduced into Minnesota Reflections and turning on and off certain data export options across all 120 collections in Reflections. Her time can all be allocated as producer time.

The manager at UMN also spent about 40 hours on the remediation task that can be allocated as producer time. However, another 30 hours of his time was spent supervising the programmer and participating in developer meetings in the aggregator and transformer role. The manager had billed his time at roughly $70 per hour for previous MDL projects.

	Time as Producer	Cost as Producer	Time as Aggregator	Cost as Aggregator
Programmer	4 weeks	$8,000	14 weeks	$28,000
Metadata Assistant	100 hours	$2,000		
Manager	40 hours	$2,800	30 hours	$2,100
Totals		$12,800		$30,100

The cost to MDL for its role as a producer of content to be archived was roughly $13,000. Much of this effort would also prepare MDL for including other CONTENTdm sites in the archive, since those extractions would be quite similar. Other significant portions of this cost arose as the result of the archiving process exposing shortcomings in existing MDL practice (the duplicate identifiers or lack of compound object identifiers, for example).

The cost to MDL for its role as an aggregator and transformer of data, the creator of SIPs that meet HT requirements, was roughly $30,000. Again, this expenditure has bought MDL some experience with CONTENTdm systems, but it would be prudent to assume that any new system introduced to the preservation mix would incur another roughly $10,000 of programming time to get integrated.

Given the roughly 49,000 images transferred before Christmas, the $42,900 in staff expenses at UMN cost about $0.86 per image transferred.

1.4. Project Management

Soon after this project began, the manager at UMN set up an instance of Basecamp at https://biomed.basecamphq.com/projects/5528974 to serve as the hub of team activity. This project used Basecamp heavily and a fairly full record the team's work can be found there.

In addition, the project manager called weekly developer meetings that included key staff from Minnesota and Michigan as well as the digital preservation consultant. These meetings focused on the details of progress of each stage, concerns with procedures, and reconciling MDL and HT practice.

Finally, the project manager also called monthly meetings of the project sponsors where issues raised by the developers, progress of the project, and concerns about the governance of a broader sustained digital preservation effort were discussed. These meetings were also open to other MDL and project staff.

These tools and meetings served the project well, giving HT and MHS staff who joined the process later a solid platform from which to review past work and access to the threads of conversation the team had developed over earlier weeks. Basecamp presented a few challenges of its own (getting new participants subscribed to ongoing conversations could be less than immediately obvious) but generally provided the team with a useful centralized collection of messages, files, examples, and meeting notes.

The process served well enough that the team intends to maintain the weekly meetings and Basecamp repository even now that the project manager is no longer engaged in the prototype.

2. Lessons

Although it was clear from the beginning that this project had an ambitious timeline, the intricacy of some of the tasks and difficulty of pulling them off with sufficient fidelity was sometimes a surprise. The project timeline slipped considerably and some significant goals were written off to complete the work.

This section of the report details some of the lessons we learned from the project.

2.1. Technical Issues

Many technical issues arose as the team worked the practical details of transforming the images and building the packages. The thorniness of some of these issues was attested to by the fact that HT staff attending to the conversations in the project Basecamp site rose from one to six. The technical issues were resolved enough to continue with the prototype, but may deserve reconsideration before continuing with a long-term digital preservation project.

2.1.1. Master of illusion

The most immediate and important lesson of the prototype effort was that the notion of a “master image” is somewhat of an illusion. Since the master image is the very essence of the preservation effort, this was discomforting, to say the least. What is to be preserved, if not the master image?

The master images on hand in Minnesota failed to qualify for preservation in HT on two grounds: some of the images were in formats that HT would not accept, and all of them failed to include certain metadata within the image binary that HT required. While some compromises were possible, for example HT came to accept JPEG as a format for inclusion in the archive, the MDL ended up having to remediate every master it sent to HT.

The Minnesota Reflections master images are in TIFF format, some of which are continuous tone (or “contone” color) TIFFs and others of which are binary (black and white) TIFFs. Luckily, the binary TIFFs were in a format acceptable to HT. But HT does not allow contone TIFFs in its collection (they waste space, being uncompressed), and demanded that these be transformed into “lossless” JPEG2000 (JP2) images. As the term “lossless” implies, this does not technically degrade the image in any way and is a theoretically completely reversible process. Even so, it means that MDL had to create new JP2 “master” images to share with HT, images that were otherwise not required by MDL.

In addition, HT requires that every image stored in the archive include certain basic descriptive and technical metadata right inside the image file itself, as XMP metadata. XMP metadata is not difficult to produce, but adding it to the image file changes that file, meaning that it is no longer the same file it was. Again, the notion of “master image” gives way to the practical requirements of the preservation archive.

The technical requirement to transform the master images in various ways before packaging them for shipment to the preservation archive added a significant step to this project and would make it very hard for many organizations to participate in a preservation archiving project without enlisting the services of an intermediary capable of such work.

2.1.2. Paper cuts can draw blood

While HT accepted the pragmatic approach laid out at the launch of the project, new HT staff became involved as the project progressed and called into question earlier decisions. The exchange of examples and corrections required many iterations on the part of MDL and HT staff. At times the back and forth resembled paper cuts to the project manager: none, individually, felt too bothersome. The issues being discussed and clarified were significant, but as they accumulated they began to draw real blood from the project, resulting in a drawn out process that significantly delayed progress.

Some of this delay may be reduced in the future as MDL staff internalize more of the HT requirements and become more fluent in the XML, METS, and PREMIS standards in use. Still, the delays evident in prototype development should raise a flag that an ongoing long-term process would likely continue to require close collaboration between HT and MDL staff. This is particularly true because MDL envisions bringing a stream of new data sources into the preservation archive as it matures. Each new source required considerable retuning of scripts, tools, and specifications during the prototype. There is little to suggest this would be much simplified over the course of a long term effort.

2.1.3. The missing fix

Although the specifications required fixity checking to so that we could assure the accuracy of copies made as masters were transferred to the project for processing, not all masters had the checksums required. A “fixity check” is something that allows a computer program to quickly determine that it has received an accurate copy of an item. In this case, no human was going to review every image, and yet the project was making copies of images to move them from site to site, onto and off of hard drives. Each act of copying is an opportunity for something to go wrong, and the image you get out the other end of this process is not necessarily the one you put in. A fixity check is usually a number the computer can get by running a calculation on the contents of the image file. If the result of that calculation matches a prerecorded value (a “checksum”) then the software can assume it got the file intact.

A significant minority of MHS binaries were missing checksums, but we proceeded anyway. This means we had no way of knowing whether we were sending corrupted images to HT from MHS. A sustained project might not want to accept this compromise, but it should be noted that some partners may not have the expertise required to prepare the required checksums.

2.1.4. Existing metadata is all over the map

Systems that don’t provide DC data require considerable attention to data mapping. The project manager spent time meeting with MHS staff and developing a mapping specification. Each time

MHS exported data, small and well intentioned tweaks to the XML format required that the mapping specification be reviewed and, in some cases, also tweaked. This task requires some degree of metadata expertise and a great deal of comfort making quick decisions including compromises uncomfortable for the typical cataloger.

Even systems that seem tailor made for this kind of sharing, such as OAI, ended up requiring significant staff time to whip into shape for this project. The programmer had to write scripts that could distinguish the records required for the project from those that were not; in some cases settings even had to be changed to produce different sets of data via OAI and then quickly reset so they would not interfere with other harvests.

2.1.5. Assumed identities

Most digital storage systems include the notion of identifiers for objects. Even before digital systems we used identifiers like call numbers to maintain control of objects. This project required MDL to, essentially, build a "namespace" within which identifiers provided to HT would be unique. This required consultation with the partner institution since in some cases MDL might have to modify the identifier in some way to meet HT requirements. The quality of identifier was also quite low, in particular, all the identifiers supplied by MDL to HT were missing any check-digit mechanism.

Managing this namespace of institutional identifiers would be an ongoing task requiring careful documentation. As we learned, even the considerably less complex task of simply maintaining unique identifiers within the Minnesota Reflections dataset was mishandled by MDL staff over the years, resulting in thousands of duplicate identifiers (a situation remediated during this project). A long-term commitment to a preservation archive would require greater care.

2.1.6. Revisiting code

Significant effort by the MDL programmer was required to develop scripts that could transform Minnesota Reflections and MHS EMu metadata into the kind of METS and PREMIS records HT requires. Though the team expected the effort of stage one to lay a foundation upon which stage two and three could be built, significant new work was required to handle new sources of data. Clearly a long-term preservation effort would continue to require the attention of a programmer to integrate new sources.

2.1.7. TMI

Sometimes we learn more about partners than we would like, we get "too much information"! During the prototype project, for instance, MHS found it easier to copy all their master images onto a drive for MDL than to write a script to sort out only those master images needed for the project.

Some of the extra images were of quite a sensitive nature, including images that MHS had promised to hold closely and not disseminate on the internet. The project manager promised MHS that MDL would erase the hard disk after extracting the masters required for this project, and MHS accepted this verbal assurance.

The lesson is that working closely on a project that reaches deeply into the digital collections of partners requires a degree of trust between the players. In this case a verbal assurance was sufficient, but a long-term preservation project may benefit from drawing up a basic agreement that documents the responsibilities of each party, especially regarding unintended lapses and the sharing of too much information.

2.1.8. Images are big

Unlike the book images, mostly bitonal TIFF images, that HT had mostly worked with to date, the contone images prevalent in this project were quite large (averaging about 24MB per image, or 300 times the size of the typical bitonal TIFF in our collection).

The data involved is large enough to make network transfers between institutions painful. The transfer of data from MHS to UMN was by hard disk. The transfers between UMN and UMich were also carried out by way of shipped hard disks. This requires a bit more planning, and the ability to purchase and part with hard disks for periods of days.

The data also caused HT to take a second look at some of our objects. For example, the *Oliver Iron Mining Company Mapbook* contains 435 images (mostly maps) scanned in full color resulting in a single SIP of over 18GB.

2.1.9. Distribution sensitivity

The missions of the MDL preservation archive and HT are not a perfect match. While the mission of the MDL preservation archive has yet to be spelled out, the public service aspects of the HT mission exposed areas of potential tension. This seems to stem primarily from HT's goal to "dramatically improve access to these materials in ways that, first and foremost, meet the needs of the co-owning institutions." This arose particularly concerning MDL desires to restrict distribution of derivative images to smaller sizes (at most 1024 pixels on the longest dimension), in some sensitive cases to no redistribution at all, and to restrict access to the masters images at HT to the contributing institution. John Weise with some consultation with John Wilkin at HT responded:

> HathiTrust is OK in principle with restricting sensitive content, and also with limiting image size to 1024 pixels in the longest dimension. These seem technically feasible as well. However, HathiTrust is maintaining a firm stance on openness when it comes to completely restricting access to content that can, legally, be unrestricted, but is constrained by contract or provider preference, including embargoed content. As

previously mentioned, including such content would “break” the current model. Formal consideration of policy change by the Collections Committee would be necessary to do anything different.

While the project did not reach the stage of testing these distribution restrictions, many questions remain in this regard. Definitions of “sensitive” and “legal” would need to be commonly understood, for example. Some method of indicating which material bear which restrictions would also have to be introduced into the metadata (and perhaps XMP data) so that HT could track these requirements.

2.1.10. Metadata manipulation

Once the SIP has been ingested by HT there are only limited options for modifying that data. The data in the derived from the SIP would not be modified. A copy of the descriptive metadata “at the time of ingest” is stored with the AIP (archival information package) and not modified after ingest. Most users would never see this, possibly out-of-date, version of the metadata. HT provides for reloading metadata to the HT catalog, which is used for search and display.

2.1.11. Projects do end

The delays imposed by the precise requirements of packaging pushed ingest into late December, early January, well past the time this report was prepared. The lessons of ingest will have to be evaluated by the team once ingest has been completed. The lesson for the team is that whether the goals are accomplished or not, projects do have to end. Both MDL and HT staff plan to continue work on ingest and retrieval into January without the project manager or preservation consultant directly involved.

Clearly the tight timeline recognized at the outset was a bit optimistic even after the elimination of newspapers from the objectives. The project manager tried to set a pace of two weeks per new collection being ingested. In fact, the pace set by a very hardworking team was closer to four weeks per new collection with another month of getting acquainted to the task and each other.

2.1.12. No free lunch

Staff costs of about $0.86 per image add up to a significant expense. Worse yet, much of this expense is incurred on a per-collection basis as each new collection is integrated with the preservation archive. As MDL reaches to participants with smaller and smaller collections, the per image cost of integration will actually rise, hopefully to be offset by a growing experience with the process that makes it a bit more routine.

HT also learned that the largely single continuous tone images present in our collections are not a good fit for their anticipated new pricing model. John Weise has said that HT would probably continue to use the existing pricing model for an ongoing MDL preservation archive collection. Though the one time and annual costs for participation during 2010 were waived as part of this prototype effort, these costs would eventually be assessed. The one time fee for our 837GB of data under the present model would have been $808 plus an annual cost of $3,231. That annual cost would continue to be charged in future years. To put this cost in perspective, it adds less than $0.10 per image to our staff costs.

Summarizing all these costs, this preservation archiving model costs about $1 per image to integrate a new collection and get it into the archive and under $0.10 per image per year in ongoing maintenance costs.

2.2. Archival Issues

As anticipated, multiple archival issues have arisen during the pilot project. As previously mentioned, the project's tight deadline necessitated that its participants work pragmatically rather than exhaustively as issues arose. This has meant that both MDL and HT have each made concessions in this project phase to resolve archival issues in ways that might or might not prove to be acceptable permanent solutions for each party. These issues and decisions need to be revisited and carefully evaluated before a full program is undertaken by MDL.

In this section, we highlight some of the major issues that have arisen and point to further work that may need to be done to resolve these archival issues prior to initiating a full preservation program in the future.

2.2.1. Format requirements

Prior to undertaking this pilot project, MDL and MHS anticipated preserving their master images, most of which are in TIFF format, with HT. Currently, HT accepts a narrow range of image formats that includes JP2, JPEG, and TIFF (binary only). Because HT only accepts binary TIFFs (and the majority of MDL and MHS images are not binary), the project began with MDL and MHS agreeing to convert their master TIFF images into JP2 images for ingest and preservation. As staff members from both MDL and MHS have evaluated this requirement, they have raised the question: what does it mean to archive and preserve items that are not the same as the institution's local master copies?

HT's restrictions around format type are intended to enable HT to ensure that they can properly manage and provide access to the materials they commit to preserve. HT has plans to provide preservation for a wider range of format types in the future, including non-binary TIFFs, but is not yet prepared to do so.

In accordance with HT's current format type restrictions, MDL has undertaken extra work to convert their master images into a supported format (JP2). If they need to recover their original master images from HT in the future, they will have to undertake a reversal of this work. In addition to adding this step to the SIP preparation process for MDL, undergoing two conversions (from TIFF to JP2 and back to TIFF) does introduce some level of risk for the MDL content. It may make it more difficult for MDL to keep track of the authenticity and accuracy of these original files over time.

Further, most of MDL's partners (as well as the larger group of Minnesota institutions that are in need of digital preservation services) do not have the staff time or expertise to undertake significant materials changes such as these conversion operations to prepare content for ingest. Unless MDL can provide infrastructure to support such conversions on behalf of partners, the format restrictions may present a high-bar that potential participants cannot reach at this time.

As MDL considers a long-term partnership with HT and vice versa, this is a key issue that will require further negotiation. This issue might be resolved through HT's acceptance and support of a broader range of formats. If not, MDL will need to carefully consider the format needs of its members and how they can best be met. MDL will also need to consider the burdens that will be incurred in a long-term project from both practical (time-based) and legal angles if MDL takes on the conversion of files on behalf of MN participants.

2.2.2. Metadata requirements

HT has documented its requirements in the HathiTrust *Digital Object Guidelines* and also on their website's "Getting Content into HathiTrust" page (http://www.hathitrust.org/ingest) and in the HT XMP checklist. As the project team built its familiarity with the HT Guidelines and other documentation at the beginning of the project, MDL anticipated mapping its existing Dublin Core metadata records into METS (and enhancing it as necessary using PREMIS) for the descriptive metadata. MDL also anticipated extracting and embedding the technical metadata (XMP) in the objects and in the METS record. MHS anticipated mapping its descriptive metadata (currently stored in the catalog side of a content management system termed EMu) to DC and then following a pathway similar to that charted by MDL (DC files embedded in METS files, enhanced using PREMIS, and additional XMP metadata extracted from or embedded in the objects and in the METS record). These were relatively straightforward processes that the project team expected to accomplish within the first month of project work for MDL (as the MDL collections were slated to be handled first, followed by the MHS content). However, HT's metadata requirements have been higher than anticipated at the project's onset and the metadata creation process has taken the majority of the project period to accomplish.

There are two key reasons for HT's recommendations and requirements. First, HT needs consistency across SIPs ingested from different Producers wherever achievable in order to keep its own processes and procedures as light as possible. For example, HT would rather not map SIP metadata to its AIP metadata because this requires them to maintain the SIP metadata, the AIP metadata, and a mapping. Second, HT is concerned about providing access to the materials it preserves through its own framework in a regularized fashion. Some of the metadata they require is metadata that allows HT to turn on all of the access features in HT's own system (e.g., make items findable through catalog search, eligible to be added to a collection, displayable, etc. If the metadata does not specify that an item is a photo, their application cannot know whether or not to offer a search box, text view, etc.).

There are likewise key reasons that MDL has questioned (and in some cases, sought to revise) some of these recommendations and requirements. MDL is a content curator; this pilot project positions MDL to take custody of content from across the state with the responsibility of preserving this content using a preservation solution (HT or otherwise). To accomplish this, they need to develop and maintain their own data management workflow for preservation purposes. Such a workflow should not be overly tied to whatever preservation solution the state uses at any given time, but rather should be based on internal management needs. Ideally, the practices MDL establishes on behalf of the state should provide a stable base that will be able to be mapped into multiple preservation solutions depending on the changing needs of the state over time.

The different perspectives of MDL and HT with regards to 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.

2.2.2.1. Descriptive Metadata

MDL has undertaken significant work to meet HT's requirements. First steps were to evaluate the HT requirements and the existing MDL metadata. MDL drafted an initial mapping of MDL DC metadata to METS in October, and also drafted a series of PREMIS events to account for three key preservation actions: 1) the ***capture*** of the original item (digitization of item from institution x by MDL); 2) the ***conversion*** of the item from TIFF to JP2 by MDL; and 3) the ***packaging*** of the JP2 and its associated METS file and its delivery to HT. This information also helped to establish the chain of custody of the item from its genesis to its ingest at HT.

The resulting MDL-METS profile was evaluated by HT and underwent multiple rounds of suggested changes. At the core of the resulting debates and eventual compromises was a reasonable difference in perspective: HT wanted to standardize the METS SIP profile to look as much like HT's AIP profile as possible in order to streamline their records-keeping practices; MDL wanted to ensure that the METS SIP profile met their local needs adequately.

Compromises reached included:

- **The language used for the PREMIS events:** MDL determined that there were three main PREMIS events that they wanted to record: "capture" for the initial creation of the digitized file, "conversion" for converting that master file from TIFF to JP2 for HT's SIP, and "packaging" for creating the zip packaging of the JP2 and METS for HT. HT requested that MDL use HT's current AIP taxonomy rather than creating its own terms for these three events, citing a particular concern about one term that was already in use by HT to describe a different event within their AIPs ("dissemination"). The compromise reached by the project team allowed MDL to define their own terms for their SIP metadata while also ensuring that those terms would not conflict with HT's AIP metadata taxonomy (e.g., they changed "disseminate" to "packaging". HT agreed to map the SIP metadata to the AIP metadata. This compromise has required that HT maintain one extra piece of documentation (the mapping); it also allows MDL to select terms that are meaningful within the local MDL environment and that may be used regardless of what preservation service(s) they work with over time.

- **The elimination of all “unknown” outcomes:** PREMIS events usually include outcomes that are recorded to indicate the status of the event’s completion. One of the PREMIS events used by MDL, “packaging,” results in an outcome that cannot be known by MDL; it will only become a known “success” as HT accepts and validates each package for ingest. As a result, MDL originally suggested using “unknown” as the outcome of this event in the SIP metadata. HT suggested that they instead use “success,” as the ingest or validation at HT will ultimately fail if it is *not* a successful packaging. MDL countered that they were not comfortable documenting “success” as an event outcome when this success could not be known. As a point of compromise, HT and MDL agreed that rather than documenting “unknown” as an event outcome, the outcome should be left out altogether if it is not known.

- **The use of MARC21 code in the METS records for agents:** For “agents” of PREMIS events (e.g., executor, destination), wherever feasible, MDL wanted to use a well-documented code rather than making up a code for the agents involved. MARC21 provided a way to accomplish this for at least some of the agents (e.g. MnU and MiU), and was selected by MDL for this purpose. HT suggested that instead, MDL should use HT-specific identifier codes. Again, MDL pointed to the need for the Producer/curator to have authority over internal metadata creation, and HT compromised to allow MDL to use MARC21 instead of HT identifiers for this purpose.

2.2.2.2. Technical Metadata

Prior to this project, neither MDL nor MHS had collected or recorded technical metadata for most of their collections. To satisfy HT’s technical metadata requirements, MDL has used the EXIF tool to extract XMP technical metadata for each item and have embedded that metadata in the JP2 files that they have created for HT.

MDL has found that the HT technical metadata recommendations have extended beyond those documented in the HT XMP checklist. They have also raised concerns about a few of the recommendations HT has made. For example, HT requested that MDL include deeper information about the make and model of the camera or scanner used to digitize items than the first XMP data contained. MDL reported concerns that some of this data is not reliable as it is delivered via the EXIF tool (e.g., resolution is given, but without any unit so that it is ultimately unclear what the value means). MDL indicated its preference to disallow data with unknown reliability, as this might unnecessarily compromise the accuracy of the technical metadata. HT has compromised in this pilot project and agreed to allow MDL to leave this information out where the values are unclear.

In keeping with HT's no redundancy or unnecessary duplication policy (which is used to streamline their own procedures, in this case so that they do not have to maintain the same metadata in multiple locations), the technical metadata has only been embedded in the object and has not been added to the METS file. The local master file (the TIFF) does not contain this metadata; only the JP2 that is prepared for HT currently holds this information. Further complicating this issue, changing XMP data also changes the binary object, which again means that a new "master" image is being generated for preservation purposes that will not match the local collections at MDL and MHS (this will be true even in the cases of the MHS binary TIFFs that HT is willing to take without a format conversion). One consideration for MDL as they continue to assess their preservation workflow and options is how to maintain local copies of this extracted technical metadata and associate it with their master TIFF images. This has heightened as-yet unresolved concerns within the project team about what it means to submit a file for preservation to HT that is not itself the Producer's master file.

2.2.3. Rights management

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. These have not been resolved, though they have been discussed and some compromises have been reached during the pilot project period, as described below.

2.2.3.1. MN Collections

MN content contributors, both within and beyond MDL, are working with diverse digital collections. Some of these collections are open access; others are not. There is a wide range of reasons for restricting access to collections, which include the following:

1. **Master images:** Most MDL participants provide derivative images to users, but restrict access to the master images for their collections. The access copies have a lower DPI and provided in smaller sizes (usually 1024 pixels in the longest dimension). This decision is made for myriad reasons, including protection of the curatorial institution's interest in and control over the images.

2. **Copyright:** Some MDL and MHS materials are clearly protected by copyright. Legally, MDL cannot provide access to these materials.

3. **Unknown copyright:** Because it is expensive and time consuming to research and reach the rights holders (even when those do exist), many MDL collections contain items for which copyright information is unknown and not likely to become known in the near future. Access to such materials is typically restricted in order to protect against unintentional infringement.

4. **Orphaned works:** MDL participants and MHS also have items in their collections for which the rights information is not identifiable. Where these are proven “orphaned works,” they are covered via Section 108 and may be displayed; however, different institutions have different levels of comfort with displaying orphaned items (in part because they have different procedures for establishing orphaned works, some of which are more rigorous than others).

5. **Contract-based restrictions:** Access to some content is restricted by contractual agreements, although the content itself is in the public domain. For example, MHS has worked with a vendor to digitize MHS materials where the vendor’s payment is a three-year period in which the vendor is the only entity that may have or provide access to the content. After that three-year embargo, MHS is able to provide access to the content. This has proven to be a helpful strategy for enabling digitization of content that MHS might not be able to otherwise digitize.

6. **Sensitive content:** There are materials that are in the public domain, but which may either endanger or offend users if they are freely shared. Curators must make decisions about what content is too sensitive for public access (e.g., lynching images, religious icons or images, hate-group materials, records that contain personal information that might be helpful to identity theft schemes, etc).

In order to encompass preservation of these types of collections as well as content that is in the public domain, MDL needs to have a preservation solution(s) that at least provides a mix of preservation levels, including highly restricted-access archiving, where the current Designated Community is the Producer (i.e., only the content contributor may have access to the DIP).

Further complicating the rights issues above, many institutions do not include item-level rights information in their metadata (and indeed, often do not know the rights status for much of their digital content). As a result, it is very difficult to accurately differentiate between the content that is in the public domain and the content that is not. For this reason, it seems likely that MN institutions will be unable to use open archiving strategies for content in mixed collections without investing a good deal of time into clearing and recording rights status at the item level.

Also, MDL and its partners have established that the public access images will need to be displayed with restricted image sizes (e.g., no more than 1024 pixels on the longest size). Because HT’s PageTurner application and API are not yet prepared to support this activity, it is not yet known what additional metadata/processes will be required from MDL in order to make this happen. HT has indicated its willingness to work through this issue, which will necessitate changing the access behavior of the PageTurner application and providing a solid authorization mechanism within the API so that only authorized users will be able to access the full object (original file and metadata). Early exchanges between MDL and HT on this matter have clarified that

MDL would like for HT to record the restriction requirement in the API rather than requiring MDL to indicate this in its source METS, and also that a simple IP authentication for authorization purposes is not an adequate solution.

2.2.3.2. HT philosophy

HT strongly believes in public access to public information. HT's preservation workflow includes their provision of access to all preserved content that is in the public domain via HT's own access portal. By policy, HT will not restrict access to content unless that content bears legal restrictions (e.g., is still protected by copyright law). They do not make exceptions for content that is constrained by contract or provider preference, including embargoed content. This policy, which can only be changed by HT's Collections Committee, is a core, defining philosophy for HT.

In the pilot project HT did agree to restrict access to sensitive materials (e.g., images of sacred objects) held by project participants. HT and MDL have discussed the challenges in defining "sensitive" materials on a large scale.

2.2.3.3. Future considerations

MDL and HT will need to continue to negotiate what materials may be maintained in a restricted access manner. If MDL participants will not be able to use HT to preserve parts of their collections due to rights issues (which will inevitably arise due to unknown rights status for content, contractual obligations, and other issues as described above), these participants will need to pursue multiple preservation service options. As MHS staff members expressed during this project, "long-term preservation becomes problematic when you cannot mingle restricted and unrestricted items (which are mingled in their local environments). Participants then need to sort those items and preserve them in different preservation solutions—something that is challenging administratively. Also, the restricted content is arguably the more vulnerable because it sits in the dark much of the time."

While the field-at-large already regularly cautions that institutions are likely to need to use different solutions for different content types, these cautions have mostly been concerned with different genres of material (e.g., e-journal content may require a different solution from special collections materials). Most practitioners have not recommended splitting special collections materials in multiple preservation repositories. This is an area that needs further study, both by MDL as it makes its determinations regarding how best to meet the state's digital preservation needs and within the broader digital preservation field, as practitioners determine what organizations might reasonably expect regarding the use of multiple repositories to meet the preservation needs of special collections materials.

MDL will need to determine its own capabilities for handling content from participants with regards to rights. Will MDL require that individual participants sort their content prior to sharing it with MDL (which might not be easily accomplished, especially for smaller organizations with less robust technical infrastructures)? Will MDL undertake this sorting on behalf of their participants (and if so, what are the liability issues that arise if a mistake is made)? Also, MDL will need to consider how and whether this might impact the workflow for the ongoing Reflections projects (e.g., does MDL need to ensure that rights issues are resolved prior to undertaking digitization, and inscribe rights data at the item level for all digitized items in anticipation of these preservation needs?).

2.2.4. Overall Archival Findings

Preservation services operate on a continuum from bit-level services to “full” preservation services that include the maintenance of a fully operational access-oriented content catalog. HT is in the highest end of this continuum. In this pilot project, MDL has explored the processes and workflows that would need to be actualized by a wide range of MN institutions in order to participate in HT preservation services. As MDL and HT continue to explore a potential longer-term relationship, it might be helpful to share pilot findings with representatives from across this range of institutions to see how their needs and abilities match up with this preservation service model.

2.3. Governance Issues

To date, MDL has coordinated and hosted Minnesota Reflections and in this context, has worked with Minnesota-based institutions in a relatively informal manner. As a program of Minitex (in turn, a program of University of Minnesota Libraries), the Minnesota Reflections repository has been offered as a free service to institutions that want to participate in a centrally hosted repository infrastructure. In order to participate, institutions respond to a CFP to suggest content that they would like to digitize and/or contribute through this program. If they contribute, they sign a document specifying that they "allow the use of (their) institution's project images for non-profit educational purposes." As further described in the MDLC *Policy on Digital Rights and Ownership*, the contributor retains ownership of content, and the MDLC Reflections program provides access to this content "for non-commercial, personal, or research use only."

As an informal program comprised of multiple projects, Minnesota Reflections has not formally documented its governance structure or the roles, responsibilities, and expectations of MDL and its participants. The current governance structure of the MDL and Minnesota Reflections includes a two-tier committee structure (Steering and Management Committees) and one employee that serves as the program's Outreach Coordinator (Marian Rengel). All committee members are volunteers (usually from the universities that perform most of the digitization work that occurs under the program), and there is no documentation concerning how to become a committee member. Minnesota Reflections has thrived to date using this informal model for its organizations.

MDL now seeks to offer a new program consisting of preservation services to a Minnesota-based constituency. In this pilot project period, the MDL has begun to study what governance structure and documentation are needed to undertake this long-term preservation work from a stable organizational base. This project's Sponsors Group has agreed that in order to offer preservation services, they will need to create an entity with (or perhaps build into MDL) a higher level of formality in its governance, policies, and documentation than has previously been engaged in the Minnesota Reflections project. As one MDL stakeholder stated, "we have an informal governance structure now because we're working on a limited mission and a year-to-year grant. As we move toward a more ambitious set of services and need to have a larger budget with corresponding implications, we see a need to have a more formal governance and reporting structure in place."

2.3.1. Issues and Needs

In initial conversations with MDL stakeholders, this prototype project has raised the following key issues and needs that should be addressed for this new program. In several cases, particularly under "services," the questions need to be thought about, decided, and negotiated internally to ensure they can meet the state’s digital preservation needs (e.g., metadata, SIP prep, formats accepted, open vs. restricted material). After the needs of the state are well understood and the governance model has been established, smaller components could be negotiated with HT.

I. Definition of mission and governance structure

A. What is the purpose

B. Who is the legal sponsor and host (Minitex?)

C. What is the governance model?

1. Who makes decisions?

2. Who does work? (including documentation)

3. If there is a board or working group, how are positions allocated?

a) Term length?

b) Rolling vs. rotating?

c) Number of representatives?

d) Voting policies?

4. How are members’ voices represented?

II. Definition of services

A. Will the solution be available to both closed and open access content (as institutions want to preserve content that they either cannot or do not want to make open access)?

B. What is the relationship between the preservation service and the Minnesota Reflections repository?

C. How much metadata will be required and who can invest the time to write the scripts to make this conform to a SIP spec (HathiTrust or otherwise)?

D. What resolution of images will be displayed via the preservation solution?

E. What are Minnesota’s statewide needs in terms of formats, and what formats will the solution address (HathiTrust or otherwise)?

F. Define long-term rights of access. What happens if MDL requests to have an image or a collection removed from HT’s access portal? What happens if MDL and HT part ways at some point in the future?

III. Definition of roles and responsibilities of sponsor and participants or members

A. What are the legal implications for the sponsor or host and for the participants or members?

B. Possible movement to a membership model (as differentiated from the Reflections project's current participant model)

C. Determine and document representation opportunities for participants or members in the governance

D. Account for differences between participants or members (e.g., perhaps weighted votes to account for differences between the large institutions that run the service and the small institutions that contribute content for preservation)

E. Account for the role that University of Minnesota will play as the conduit to HathiTrust if the state uses HathiTrust as its preservation solution

F. Ensure that the governance structure will provide stability in the event of administrative changes at any of the participating or member institutions

G. Define length of terms, withdrawal policies, replacement of committee members

IV. Definition of financial structure and long-term plans

A. Ensure that the financial structure will provide stability beyond grant funding and limited-term awards.

B. Consider potential charges to participants or members for ongoing services to ensure no interruption of preservation services

C. Think through issues of free-ridership

D. Think through the roles played by institutions and what happens if any of the participants that play central roles in SIP preparation, etc, were to drop out

3. Alternatives

The field of preservation is still in its early phase of development, however, there are already multiple options available to those seeking preservation services. Some of these options are available as “brokered” services, where MDL would contribute content to an external service provider for preservation (research institution-based options include HathiTrust, Chronopolis; vendors include OCLC’s Digital Archive). Others are available as open source solutions that MDL could run on behalf of its membership (e.g., LOCKSS, DAITSS). Still others are available as collaborative initiatives, where MDL would participate in the preservation process with a community of practitioners (e.g., MetaArchive). There are costs and benefits associated with each pathway; there are also differing services and requirements across these options (e.g., closed vs. open access, format type restrictions, metadata services, etc.).

As MDL completes this pilot project, it will have a good idea of how participation in HathiTrust might be structured and what services and requirements it will include. What might other options look like, and could any of them provide a close fit for the state’s needs? The information below is intended to provide a brief overview of some of the leading preservation solutions in the U.S. context.

3.1. Research institution-based external service providers

3.1.1. Chronopolis *(text by Katherine Skinner and David Minor)*

Founded in 2007 with funding from the NDIIPP program of the Library of Congress, Chronopolis is a digital preservation data grid framework developed by the San Diego Supercomputer Center (SDSC) at UC San Diego, the UC San Diego Libraries, the National Center for Atmospheric Research (NCAR), and the University of Maryland's Institute for Advanced Computer Studies (UMIACS). A key goal of the Chronopolis framework is to provide cross-domain collection sharing for long-term preservation. Using existing high-speed educational and research networks and mass-scale storage infrastructure investments, the partnership is designed to leverage the data storage capabilities at SDSC, NCAR and UMIACS to provide a preservation data grid that emphasizes heterogeneous and highly redundant data storage systems.

Each Chronopolis member operates a node containing at least 100 TB of storage capacity for digital collections. The Chronopolis methodology employs a minimum of three geographically distributed copies of the data collections, while enabling curatorial audit reporting and access for preservation clients. The key underlying technology for managing data within Chronopolis is the Integrated Rule-Oriented Data System (iRODS), a preservation middleware software package that allows for robust management of data. The partnership is also developing best practices for the worldwide preservation community for data packaging and transmission among heterogeneous digital archive systems.

Chronopolis is now offering its preservation services as a fee-based service to organizations in need. This service is available for immediate use by institutions who need a mature preservation environment but don't want to create the infrastructure needed on their own.

Ingest into Chronopolis requires the transfer of files either via hard drive or network transfer (using BagIt or a similar mechanism). SIP structure is established by the Producer; Chronopolis manages and regularly audits the data, and upon request, provides preservation copies back to the Producer in the same structure that Chronopolis originally received. Chronopolis is open to all file formats.

Pricing for the Chronopolis preservation service is available from David Minor, david@sdsc.edu.

3.2. Open Source solutions for internal service hosting

3.2.1. DAITSS *(text by Katherine Skinner and Priscilla Caplan)*

DAITSS2 is an open source software solution that is slated for release in the first quarter of 2011. Developed by the Florida Digital Archive (FDA) under the leadership of Priscilla Caplan, this open source package has been in use since 2006 for the Florida university system. FDA is interested in helping other states and other collaboratives implement DAITSS in 2011.

To run DAITSS, FDA currently supports two positions, a full-time manager and an operations technician. The estimated annual cost of hosting this program is less than $130,000, and the program is preserving 63 TB of content to date.

FDA Model

The FDA is a "dark archive" with no public access and no functionality beyond long-term preservation. University libraries in the state system that participate in the preservation repository use various applications for their institutional repositories, "digital library" or digital asset management systems, ETD systems, and so on. Regardless of how a digital resource was created or is made available on campus, if the resource is selected for preservation a copy must be sent to the FDA in a prescribed submission package format (available here: http://www.fcla.edu/digitalArchive/daInfo.htm). The FDA promises to deliver back to the library on request, at any point in time, a copy that is bit-wise identical to the original resource. If all files comprising the resource are in supported formats, the FDA will also deliver a version guaranteed to be renderable with tools available at the time of the request.

In order to archive materials in the FDA, a library must first negotiate a binding Agreement with FCLA. The Agreement lays out the responsibilities, liabilities and warranties of both parties, and is signed by the FCLA director and either the library director or university counsel on behalf of the university board of trustees. Once the Agreement is signed, the library becomes an affiliate of the FDA. The term "affiliate" is used instead of "submitter" or "customer" to emphasize that the FDA and the libraries work in partnership. The library deans are de facto the governing board of the FDA, and responsibility for long-term preservation is shared between the FDA and the affiliates.

In this model of shared responsibility, the affiliate is responsible for

- selecting content to be archived;
- securing rights to archive and preserve the content;
- describing the content adequately for its own purposes;

- submitting packages in format required by the FDA;
- maintaining local records of what it archived;
- withdrawing content that should no longer be archived;
- requesting disseminations when needed;
- providing access to disseminated content.

The FDA is responsible for

- accounting for every package submitted with an Ingest or Rejection report;
- providing useful counts and reporting information on ingested materials;
- implementing preservation strategies as described in the FDA policy guide;
- preserving original files exactly as submitted, with demonstrated integrity, viability and authenticity;
- providing a renderable version of all supported formats;
- providing disseminations on request;
- attempting to achieve and maintain certification as a trustworthy repository.

Brief description of technical achievements

The DAITSS application that underlies the FDA is locally written software designed to implement the OAIS reference model and perform active preservation strategies based on format transformation. DAITSS attempts to identify and describe all files, and for any file in a supported format, will create a normalized or migrated version (or both) when possible and desirable.

Some of the hallmarks of DAITSS are:

- The application does preservation and nothing else, and as such must function as a "back end" to other systems for acquisition and user access.
- It depends heavily on well-known standards, including OAIS, METS and PREMIS.
- Archived content is stored with all of its metadata, although some metadata is also replicated in a database for fast access, so the archival store could be interpreted even without the application.
- All format-based processing, including migration and normalization, is done inside the system, which keeps a rigorous record of digital provenance.
- Format-based processing takes place at the time of ingest and as part of a process called "refresh"; packages are refreshed before dissemination to ensure that they are fully up-to-date.

3.2.2. LOCKSS

The Lots of Copies Keep Stuff Safe (LOCKSS) open source software package was developed by the LOCKSS team at the Stanford University Libraries in the late 1990s. Currently, there are 11 Private LOCKSS Networks actively preserving content around the world.

A Private LOCKSS Network is a closed group of geographically distributed servers that are configured to run the open source LOCKSS software package. This software makes use of the Internet to connect these servers with each other and with the websites that host content that is contributed to the network for preservation. LOCKSS is format agnostic, and as such, may be used to preserve and monitor any format type.

In PLNs, every server has the same rights and responsibilities. There is no lead server equipped with special powers or features. After a server is up and running, it can continue to run even if it loses contact with the other servers. Such a peer-to-peer technological structure is especially robust against failures. If any server in the network fails, others can take over. If a server is compromised or corrupted, any other server in the network can be used to repair its copies. Since all servers are functionally alike, the work of maintaining the preservation network is truly distributed among all of the partners in the network. This is one of the great strengths of the distributed preservation approach.

The LOCKSS servers of a PLN perform an ongoing set of preservation-oriented functions:

- They ingest submitted content and store it on their local disks
- They conduct polls, comparing all cached copies of content to arrive at regular consensus network-wide on the authenticity and accuracy of content
- They repair any content that is deemed corrupt through the network polling process
- They re-ingest content from its original location (so long as it is available) in order to discover new or changed content, and they preserve any modifications alongside the original version
- They retrieve and provide a copy of the content to authorized recipients
- They provide information about their stored content for auditing and reporting purposes
- They migrate out-of-date file format types to specified formats, storing both the original and the migrated file in a versioned manner

PLNs range in size from very small groups with seven distributed servers to large groups with 24 servers. To date, the PLN methodology has scaled well from both an organizational and technical perspective, with the largest PLN currently hosting a 270TB capacity. PLNs typically operate as “dark” archives, meaning that access to content is limited to authorized representatives from that content’s Producer site.

PLNs are inexpensive to run, as demonstrated by the COPPUL, ADPNet, Synergies, and other PLNs. They do not require central staffing (many are run via volunteer directorships), and the network’s core technical infrastructure may be managed directly by the LOCKSS team at Stanford University.

Many groups are currently building tools that layer on top of LOCKSS to enhance the curatorial and auditing functions of PLNs, including MetaArchive (produced the Conspectus in 2004 for metadata capture and content monitoring purposes; currently working on a variety of web services, including format validation, for release in 2011), Data-PASS (working on an audit framework that will be released in 2011), and LuKII (bridging the leading German open source data management package with the LOCKSS software).

Documentation regarding how to run a PLN (both in terms of governance and the technical structure) abounds; see for example *A Guide to Distributed Digital Preservation* (eds. Katherine Skinner and Matt Schultz; Educopia Institute: 2010) and a wide range of articles (for clusters, see *Library Trends* Volume 57, Number 3, Winter 2009; and *Library Hi Tech* Volume 28 issue 2, 2010).

3.3. Collaborative solutions for community-based preservation services

3.3.1. MetaArchive Cooperative

The MetaArchive Cooperative (http://metaarchive.org) is an international preservation network comprised of research institutions. Established in 2004 through the National Digital Information Infrastructure and Preservation Program (NDIIPP) of the Library of Congress, the MetaArchive model focuses on sharing responsibility, sharing expertise, and sharing cyberinfrastructure to enable libraries, archives, centers, and museums to accomplish their preservation goals as a distributed community. Its members bring their collective strength to bear on the preservation challenge, not by outsourcing or centralizing operations, but rather by building knowledge and infrastructure in local institutional environments. MetaArchive's technologies are open source, and its curation and preservation services ensure the long-term accessibility of authentic content. Together, the membership preserves a broad range of digital assets, including ETDs, newspapers, journals, and archival holdings (including video, audio, image, and other media types), as well as digital creations from the digital humanities, social sciences, and sciences (such as datasets, databases, portals, and other resources).

Organizational Model

MetaArchive is a cooperative membership organization with three membership categories: Sustaining Members, Preservation Members, and Collaborative Members. Each member runs a server for the MetaArchive network and prepares its own content for ingest in consultation with our central staff.

- *Sustaining Members* are institutions that make the highest financial and technical commitment to the cooperative. They provide the Cooperative's leadership and commit to a pioneering role in the emerging field of distributed digital preservation. Each Sustaining Member has one voting representative on the MetaArchive Steering Committee. They offer input and guidance on future development paths for the MetaArchive, including the creation of new data curation tools and reporting tools. Sustaining members pay an annual membership fee of $5,500.

- *Preservation Members* are institutions that preserve content in the Cooperative and support its overall infrastructure through running and maintaining a network server. Preservation members pay an annual membership fee of $3,000

- *Collaborative Members* are groups of institutions that run a single, shared, centralized repository and preserve this shared content in the MetaArchive network. They also help to support the Cooperative's infrastructure through running and maintaining one of our network's servers. Collaborative members pay an annual membership fee of $2,500 plus $100/participating institution.

So long as MDL runs a central server through which the state's collections are processed for ingest, MDL would qualify to be a Collaborative Member of the Cooperative.

Governance and Staffing

The Cooperative is an intentionally lightweight organization that focuses on building infrastructure within member organizations, not within a central service agency. Most of the work of the Cooperative is achieved by member institutions, including SIP preparation, ingest, AIP monitoring, and DIP provision when necessary.

MetaArchive is governed by a Steering Committee comprised of one voting representative from each Sustaining Member and one representative each from the Preservation and Collaborative Member categories. Leadership of the Steering Committee is determined by nomination and simple majority vote by the Steering Committee. The Steering Committee directs the activities of the Cooperative through weekly phone calls and an annual meeting that is held at a member site.

MetaArchive's work is also guided by three committees, the Content Committee, Preservation Committee, and Technical Committee.

- The *Content Committee* is responsible for organizing, developing, and documenting content selection practices and MetaArchive SIP preparation guidelines. The Content Committee also recommends prioritization of new subject- and genre-based archives for the preservation network (e.g., ETD Archive, Newspaper Archive, Southern Digital Culture Archive).

- The *Preservation Committee* is responsible for researching, developing, documenting, and disseminating policies, procedures, and evaluative means for enhancing the MetaArchive Cooperative's practice of trustworthy distributed digital preservation.

- The *Technical Committee* is responsible for developing and maintaining technical specifications and coming to agreements on hardware, software, and networking protocols; overall server architecture; application development; and software maintenance.

The Cooperative is currently staffed by 3.5 positions, a Program Manager, a Collaborative Services Librarian, a Systems Administrator, and a Software Engineer. These positions provide the foundation for the Cooperative's ongoing work. The Program Manager oversees the day-to-day operations of the MetaArchive Cooperative. The Collaborative Services Librarian trains new members and assists all members as they manage their network servers, prepare their content for ingest, and monitor their content. The Systems Administrator oversees the core network functions, monitors content, and helps to train members as they bring up and monitor their own servers. The Software Engineer assists all members with their SIP development, ensuring that the SIP conforms to MetaArchive data management guidelines.

The Cooperative and its membership believes that local staff need to be actively involved in preservation activities in order to maintain a vibrant and knowledgeable preservation community. The Cooperative therefore intentionally depends upon distributed staffing located at each member institution. Each member runs an autonomous server for the network, ensuring that every copy of content is maintained by a different system administrator and that there is therefore no one point of human failure within the network. Each member also works in concert with the central staff to prepare its content for ingest. Most members further contribute to the Cooperative through Committee assignments. These roles are imperative for the Cooperative and its membership, both in practical and philosophical terms.

Technical Model

MetaArchive's technical model is grounded in the principle of distributed digital preservation. The central assertion of the Cooperative is that research institutions can and should take responsibility for managing their digital collections, and that such institutions can realize many advantages in collaborative, distributed long-term preservation strategies.

Historically, the most effective preservation efforts have succeeded through some strategy (intentional or not) of distributing copies of content in secure, distributed locations over time. Many of the threats to obsolescence in the digital arena are the same (natural disasters, intentional attacks, accidental destruction) as those faced in other eras. To be sure, there are additional challenges that we must meet with regards to managing digital content for long-term preservation, but these can be accomplished within a distributed network environment.

Implementing this strategy requires an investment in a distributed array of servers capable of storing and managing digital collections in a pre-coordinated manner. A single research institution is unlikely to have the capacity to operate multiple, geographically dispersed and securely maintained servers; MetaArchive enables research institutions to benefit from the shared network capacity made possible and financially reasonable through community-based work. We use the open-source LOCKSS software, developed at Stanford University Libraries, for our network base, and are layering data management tools on this foundation to accomplish our full preservation aims.

Replications and integrity

Our distributed network enables each member's content to be preserved at multiple (currently, at least six) geographically distinct sites across two continents. These replicated copies do not merely function as back-ups to be consulted in the event of data loss, but rather are regularly compared with one another to ensure that data integrity remains consistent across all replications all the time.

Ingest and versioning

All content is ingested via http, either through its access-based address (which can be open or secure) or through temporary mounting on a staging server (for collections that are not regularly available on line). The ingest pathway works well with a wide variety of repository infrastructures, including DSpace, ETD-db, CONTENTdm, Fedora, and other leading solutions. For those collections maintained in an access-based address, the network servers that preserve the content regularly revisit that content to ingest any changes or additions made to it. These versions are stored alongside the original, and all versions are available to the content producer and designated community upon request. For staged collections, versioning is accomplished through iterative remounting of the content for re-crawl according to a revision schedule established by the content producer.

Access

Access to content preserved within the network is restricted to only authorized representatives from the institution that contributed that content.

Formats and SIP requirements

MetaArchive works on the principle that content contributors (Producers) are curatorial experts and works in partnership with these experts to conduct sensible preservation practices that are designed to match member needs. The central staff consults with all member institutions regarding the collections that they select for preservation prior to ingest. During this process, the central staff makes recommendations to members that can help them to establish solid data management practices locally to support their long-term preservation aims. The central staff also helps members properly document their curatorial decisions and workflows in order to provide essential documentation to aid in the reconstitution of collections. MetaArchive does not restrict ingest based on format type or metadata standards. LOCKSS is format agnostic, which means that MetaArchive is able to accept and preserve all formats that its content curators deem worthy of preservation.

4. Appendices

A number of documents were prepared as part of this project. These appendices gather together some of these works for reference and for the record.

4.1. Governance Models for Collaborative Preservation

At this time, state-based and collaborative preservation repositories are using a range of governance models. Three of these are of particular interest in the Minnesota context: 1) the Florida Center for Library Automation's Florida Digital Archive (http://www.fcla.edu/digitalArchive/daInfo.htm), 2) the Alabama Digital Preservation Network (ADPNet) (http://www.adpn.org/resources.html), and 3) the MetaArchive Cooperative (http://metaarchive.org/prospective). Below is an outline of each of these governance models and its potential relevance to the development of a governance plan for the Prairie State Repository project.

1. FCLA-FDA (http://www.fcla.edu/digitalArchive/daInfo.htm)
Mission
To provide a cost-effective, long-term preservation repository for digital materials in support of teaching and learning, scholarship, and research in the state of Florida. In support of this mission, the Florida Digital Archive guarantees that all files deposited by agreement with its affiliates remain available, unaltered, and readable from media. For those materials designated to receive full preservation treatment, the Florida Digital Archive will maintain a usable version using the best format migration tools available.

Governance
Host: Florida Center for Library Automation, a system-wide center of the state universities attached to the University of Florida for administrative purposes
Advisory Board: FCLA Advisory Board (directors of 11 public university library directors, 1 rep of the state university system, 1 rep from the division of community colleges, and the state librarian of Florida
FDA staff: assistant director, technical staff member, administrative assistant

Eligibility
Public universities in the state university system, PALMM partners (institutions partnering with FL state university libraries), and others as approved on a case-by-case basis by the Board

Services
- Ingest of SIPs
- Secure storage and management of AIPs
- Full preservation (normalization, migration) for supported formats
- Dissemination
- Withdrawal (including to update information in previously submitted package—existing AIP must be withdrawn and new SIP submitted for ingest)
- Reporting

Roles/Responsibilities
Shared between FDA and Affiliates

Affiliate responsibilities:
- Negotiate an agreement for the use of FDA (details contact people, itemized description of all classes of materials to be deposited and the level of preservation desired)
- Select content for archiving and ensure adequate descriptive metadata is maintained locally
- Ensure rights to deposit and give FDA all permissions it needs
- Submit content in format required by the FDA SIP specification (SIP contains only one SIP descriptor and at least one other file; SIP descriptor must be a valid METS document that conforms to DAITSS SIP Profile, SIP descriptor must reference all files in the SIP that are meant to be archived, SIP must be contained in a single folder with a name of up to 32 characters, name of SIP

descriptor must be the same as that of the package, SIPs cannot be bundled or compressed or digitally signed)

- Maintain records of what they have archived with FDA
- Review reports
- Work with FDA staff to resolve problems
- Alert FDA if content no longer needs to be archived
- Request dissemination when needed

FDA responsibilities:

- Ingest and store materials in accordance with Affiliate's Agreement
- Restrict authorization to submit, withdraw, disseminate, or receive reports on materials to individuals specified in Affiliate's Agreement
- Provide detailed Ingest or Error info for every SIP
- Preserve original files exactly as submitted, with demonstrated integrity, viability, and authenticity
- Use appropriate preservation strategies for files in supported formats to ensure renderable version of files
- Provide DIPs on request (always contains original; may also contain modified renderable versions that have been normalized and/or migrated
- Provide appropriate reports to Affiliates
- Achieve and maintain certification as a TDR

Costs

Actual costs, staffing (around $150K/year)

Thus far, no charge for use. Specifies that this may change and that billing will be instituted with consent of FCLA Advisory Board with 180 days or more notice to Affiliates

2. MetaArchive Cooperative (http://metaarchive.org)

Mission

The mission of the MetaArchive Cooperative is to foster better understanding of distributed digital preservation methods and to create enduring and stable, geographically dispersed "dark archives" of digital materials that can, if necessary, be drawn upon to restore collections at the contributing organizations.

Governance

Host: Educopia Institute, a 501c3 organization

Steering Committee: one rep from every Sustaining Member plus an elected Chair

Steering Committee meetings are also attended by one non-voting rep for each member category

Staff: Program Director, Collaborative Services Librarian, Systems Administrator, .5FTE Software Engineer

Eligibility

Any digital memory organization or collaborative; approved by the Steering Committee on a case-by-case basis

Services

- Assistance to members as they form their preservation policies and strategies
- Advice and assistance to members as they prepare content for ingest (including extensive testing)
- Assistance to members as they bring up their MetaArchive server
- Facilitation of content ingest (including ongoing ingests of changed/new content which is stored in a versioning system)
- Secure storage and management of content in a distributed network with at least 7 copies of every file in distinct geographical locations
- Full preservation (normalization, migration) for supported formats when necessary

- Dissemination when necessary
- Reporting

Roles/Responsibilities
Shared between MetaArchive and Members

Member responsibilities:

- Maintain membership in good standing by fully complying with this Charter and the definition of membership herein and by acknowledging and agreeing to processes, procedures, and standards of governance found in the Charter and with technical requirements identified in the Charter
- Support at its own expense any and all costs incurred by participating in the Cooperative, including but not limited to paying membership fees, travel to required meetings, and other costs of participating in the Cooperative;
- Implement appropriate standards for addressing copyright and other issues related to contributed content in order to comply with local, state, federal, and international law, including the use of exemptions set forth within U.S. copyright law at Section 107, 108, and elsewhere and permissions through "deeds of gift" or other clearances;
- Represent and warrant that to the best of its knowledge the Member is not contributing content to the Preservation Network that would infringe the rights of others and that the Member holds sufficient rights to License the Cooperative and Members to use the content consistent with the requirements of a multi-site preservation strategy;
- Hold the Cooperative and other Members harmless in the event of infringement, claims of infringement, loss of data, interoperability, and any other technical standards and governance claims by waiving any rights of recovery for any costs or damages associated with its relationships and Agreement with the Cooperative;
- Indemnify the Cooperative to the extent permitted by law for any losses and costs incurred by the Cooperative and Members such as but not limited to legal fees, costs, and damage awards arising from infringement or other claims directly related to its activities in working with the Cooperative and Members;
- Cure any material breaches of the contract within a 90-day period unless the Cooperative agrees in writing to an extension of the cure period;
- Create and begin maintaining a preservation site, comprised of a MetaArchive-LOCKSS cache;
- Make the MetaArchive-LOCKSS cache available for testing new software and other MetaArchive developments as needed;
- Participate actively in the MetaArchive Preservation Network by contributing, ingesting, and monitoring content from the Cooperative;
- Join and maintain membership in good standing with the LOCKSS Alliance to promote the development and strength of the community (http://www.lockss.org/lockss/LOCKSS_Alliance);
- Employ an implementation of LOCKSS software that complies with all requirements in the current and subsequent versions of the Private LOCKSS Network software;
- Design and implement system features ensuring compliance with Cooperative security requirements and content validation, including but not limited to integrity checking as well as metadata analysis and tracking;
- Install and maintain any other software that may be required for participation in the Cooperative
- Provide technical and administrative contact information as necessary to enable communication with other Cooperative participants as needed or upon request by the Cooperative.

MetaArchive responsibilities:

- Retrieval of the Member's content in case of a catastrophic loss at that Member's organization;
- Use of a web-based tool – the Conspectus – under an appropriate open source license that enables Members to record information about submitted collections;
- Distributed archiving of Members' digital collections across multiple preservation sites;
- Reports and statistical information about Members' submitted content and statistical reports about the overall Preservation Network;

- Service opportunities within Cooperative Working Groups;
- Attendance and participation at Cooperative conferences and workshops at a discounted rate;
- Opportunities to collaborate with and/or learn from experienced digital preservation administrators, librarians, technologists, and others who work with the Cooperative;
- Storage space that can be purchased on an as-needed basis;
- Additional preservation services that can be purchased at a contract rate (e.g., consulting and training around preservation issues);
- Assistance with the installation and maintenance of LOCKSS software for private networks and any other improvements and ancillary software developed by the Cooperative, documentation of processes and technical standards, and technical support;
- In the case of catastrophic circumstances, the ability to request technical and financial assistance with the restoration of a Preservation Site's servers, software, and collections by the MetaArchive Cooperative.
- Access to the technical knowledge and expertise of Cooperative Members and technical support to establish and maintain preservation sites in compliance with the Cooperative's technical standards

Costs
Actual costs, staffing (around $300K/year)
Three tiers of membership:

- Sustaining Members: $5500/year plus $1/GB/year storage fees
- Preservation Members: $3,000/year plus $1/GB/year storage fees
- Collaborative Members: $2500/year plus $100/participating institution plus $1/GB/year storage fees

All members run one server for the network ($5K investment locally at the beginning of each 3-year membership term)

3. ADPNet (http://www.adpnet.org)
Mission
The mission of ADPNet is to manage and sustain a reliable, low-cost "dark archive"2 for the long-term preservation of locally-created digital resources in Alabama. ADPNet seeks to foster better understanding of distributed digital preservation methods in the state and to create a stable, geographically dispersed dark archive of digital content that can be drawn upon if necessary to restore collections at the Member institutions.

Governance
Host: Network of Alabama Academic Libraries (NAAL) at the Alabama Commission on Higher Education
Steering Committee: comprised of one voting rep appointed by each member. Term of service is 1 year, may be reappointed
Staff: Chair (elected to 1-year terms), administrator, LOCKSS staff at Stanford

Eligibility
Any Alabama cultural heritage institution creating publicly-available digital assets whose activities and objectives are consistent with the Alabama Digital Preservation Network's mission and principles may join ADPNet. This includes but is not limited to universities, libraries, museums, historical societies, and agencies of state government, as well as consortia of organizations and individual projects.

Services

- Highlight the importance of preserving significant digital assets to the academic community, state agencies, and other cultural heritage institutions in Alabama.
- Manage and sustain a distributed, low-cost network for the long-term archival storage and preservation of digital assets created by all types of institutions in Alabama.

- Sustain a collaborative administrative structure to manage the storage network and assure its long-term viability.
- Demonstrate that a distributed long-term storage repository for digital assets can support different types and sizes of collections from different kinds of institutions.

Roles/Responsibilities
Shared between ADPNet and Members

Member responsibilities:

- Meets the criteria for membership described in Paragraph 2.0.
- Agrees to install and and maintain a LOCKSS server in the Network and make that server available to support ADPNet initiatives and programs. The server must satisfy the ADPNet technical requirements
- Agrees to contribute locally-created and publicly available digital content to the Network and harvest digital content from other member institutions.
- Commits to joining the LOCKSS Alliance for the duration of its membership in ADPNet. No other dues or membership fees apart from the LOCKSS Alliance fee are assessed for membership in ADPNet.
- Commits to a membership term in ADPNet of no less than three-years' duration, with a one-year notice to cancel membership thereafter.

ADPNet responsibilities:

- Highlight the importance of preserving significant digital assets to the academic community, state agencies, and other cultural heritage institutions in Alabama.
- Manage and sustain a distributed, low-cost network for the long-term archival storage and preservation of digital assets created by all types of institutions in Alabama.
- Sustain a collaborative administrative structure to manage the storage network and assure its long-term viability.
- Demonstrate that a distributed long-term storage repository for digital assets can support different types and sizes of collections from different kinds of institutions.

Costs
Actual costs, administrator (unknown)
Beginning to charge for use this year (hard figures not yet available)
LOCKSS Alliance fees
Install and run a LOCKSS preservation node ($5K or so each 3-year term)

4.2. MDL-HT Hot Spots

This document tracks “hot spots” that MDL and partners need to discuss. The discussion itself won’t be evident in this document, but the issue, its status, and a brief note on its resolution will be included here. If you think an issue should be included or updated, either leave a comment below, or edit the document yourself following the model set out in existing entries.

Unresolved Hot Spots

These issues are still under active discussion.

{A} Redistribution of digital masters

Issue: MDL members whose masters are being moved to HathiTrust as part of this preservation effort would feel betrayed if the digital masters were shared with the world. HathiTrust believes anything in the public domain should be shared. Can we devise a mechanism, both technical and policy, that allows us to restrict the distribution of MDL digital masters from HT to the MDL?

101027: JTB asks, will MDL link a policy of open access to digital masters to the participation in an effort? Or open access of things in the public domain? Does something like Minnesota Reflections become an open access repository.

Status: Discussed without resolution at 100922 and 101027 sponsors meetings.

{B} Restricted resolution for display of digital images.

Issue: MDL members scanned images at high resolution, but only intended to mount lower resolution versions of those images on the web. Would HathiTrust be willing to limit the resolution of derivative images shared on the web from MDL contributions?

Status: Discussed without resolution at 100922 sponsors meeting.

{C} Limiting contributed images to JPEG2000.

Issue: HathiTrust has only collected JPEG2000 images to date. In particular, HT has avoided inclusion of TIFF images. MDL partners, MHS in particular, hold master images in TIFF format. Would MDL be open to migrating these images to JPEG2000 as part of the preservation effort?

1009: Brought up with regard to the work plan as an open question. “Future-proof formats. In this project we pretty much envision taking our master images as they come, in other words, we have no opportunity to demand that these masters adhere to strict format specifications. Should we build a set of specifications for each content type that would facilitate sustainability, migration, and general future-proofing of the preservation archive?”

101001: Discussed by EFC and JW after consultation with JR and BT. Agreed to use only JPEG2000 for prototype project, but noted that this is a hot spot issue for future work with HathiTrust. HT has good reasons to be conservative w/r/t formats accepted, but the degree of conservatism being demonstrated by an inability to ingest TIFF is worrisome for a project that will eventually have to deal with input from all around the state of Minnesota.

101027: (JTB) Our historical MDL practice of using TIFFs is not a deep philosophical position. We've been interested in what the broader HT community would think of standardizing on JP2 for image data. The notion of on-the-fly derivation of the TIFF. We are discovering the promises we can make.

Status: Resolved for prototype, but should still be discussed w/r/t wider impact on sustainability of "live" project.

Resolved Hot Spots

When some of the issues noted above are resolved, we'll move then to this section for the record.

{D} Dynamic metadata.

Issue: It has been noted that the metadata shipped to HT may, from time to time, require updating for accuracy. One example from recent MDL history was the move of the JJHill collection to MHS and the corresponding changes in ownership metadata required. What provision can be made for updating or reloading metadata into HT?

100927: Discussed on 100927 with JW. HT provides for reloading metadata to the HT catalog, which is used for search and display. This is what users see. However, there is also a copy of the descriptive metadata "at the time of ingest" which is stored with the ingest package and not modified after ingest. Most users would never see this, possibly out-of-date, version of the metadata.

Status: Resolved. MDL can accept the potential of correcting catalog metadata as sufficient for our purposes.

4.3. MDL-HT Image Ingest Prototype Guidelines

This document seeks to define the interactions between the Minnesota Digital Library (MDL) and the participants in the MDL-HT Image Ingest Prototype project. It serves as a model for future requirements of an ongoing process of Minnesota digital image preservation via HathiTrust (HT). It includes a brief description of the project and its goals, then turns to defining expectations for the extraction, packaging, transfer, ingestion, and display of content for this project.

Background

Our goal is to develop the workflow to move digital image data and associated metadata from Minnesota into the HathiTrust, demonstrate that workflow by moving a defined set of images into HathiTrust, and work with HathiTrust to define the appropriate display for these images in that system. The University of Minnesota, a HathiTrust participant already, is serving as the conduit for this exploration.

Accomplishing this will require attention to six stages of processing: **extracting** master images and metadata from current repositories, **reformatting** the images to suit the preservation archive, **packaging** these binaries and associated metadata as required for ingest, **transferring** these packages to HathiTrust, seeing that these packages are **ingested** by HathiTrust, and providing for **display** of these images at HathiTrust and retrieval of the masters via API calls.

We plan to address a variety of content types, from simple continuous tone images, to compound objects made up of a series of images in a certain structural relationship, to images containing text and associated optical character recognition (OCR) derived text.

[0.] Principles

[0.1.] The project is on an extremely tight timeline, so of necessity it will require that we **act as pragmatically as possible**. We will deal with content as it exists in local systems, without any expectation that either binaries or metadata be remediated to meet our standards. We will strive to meet the HathiTrust Guidelines for Digital Object Deposit, but we also realize that we are dealing with a new type of content and a much less sophisticated set of partners. Part of the test of this prototype is finding the minimum criteria that can provide successful HathiTrust participation.

[0.2.] We want the product we produce through this process to be **trustworthy and transparent**. Toward this end we will do what we can to document our decisions and the provenance of objects both in external documentation and within the metadata and packages produced by the project.

[0.3.] We are mindful that this project is a prototype, intended to develop and demonstrate a workflow for a future sustained effort, but not necessarily carrying out every element expected from such a sustained effort. In other words, we accept the need for **shortcuts** demanded by our tight timeframe, though we will discuss and document them as such and make suggestions about more appropriate long term strategies for future implementations of the workflow.

[1.] Content

[1.1.] We seek only content that is of **cultural heritage value** to Minnesota. Given the lack of time to make sophisticated judgements, we are accepting that content in MDL Reflections and the Minnesota Historical Society (MHS) content management systems do have such value.

[1.2.] We seek only content that has been **professionally produced** to meet basic archival standards. This means that we expect digital images to have as much fidelity to the original object as possible, represented by JPEG2000, and produced using well maintained scanning equipment. Again, we are accepting that MDL and MHS meet this standard without further review.

[1.3.] We require some form of **validity check** be present at the source so that we can ensure the digital object being conveyed to HT is the same digital object being held locally. This would probably be the presence of a local MD5 checksum for each binary object, though we would be open to other forms of validity check.

[2.] Extract

[2.1.] We require that metadata be provided in **XML format**. We seek as much metadata as we can get from local systems, including any descriptive, technical, or administrative metadata present.

[2.2.] Metadata may be extracted either via a **pull** process, such as an OAI-PMH (Open Archives Initiative Protocol for Metadata Harvest) harvest, or via a **push** mechanism, such as an XML report generated from a local system and sent to the project.

[2.3.] Metadata must include some **reference to binary objects** that identifies those objects uniquely.

[2.4.] Binary objects may be extracted via a **pull** process over the network or via a **push** process such as a hard disk sent to the project from a local system.

[2.5.] Binary objects must be named in such a way as to be **uniquely identifiable** and matched to their metadata.

[2.6.] A **unique identifier** must be provided for each item described in the metadata. This identifier may or may not be the same as that used to link binaries to the metadata.

[3.] Reformat

[3.0.] Binary objects must be provided in **JPEG2000** format. This means that MDL will have to reformat masters that are currently in TIFF format for this prototype.

> *[3.0.] This point has been resolved for the prototype, but the implications of JPEG2000-only remain severe for a longer term effort. See "hot spot" issue {C} for further discussion.*

[3.1.] Each image will be reformatted into JPEG2000 images suitable for HathiTrust.

[3.2.] Technical metadata will be created for each image and recorded in an XMP package that will be embedded in that same image.

[3.3.] A new MD5 checksum of the resulting JPEG2000 image will be generated.

[4.] Package

[4.1.] MDL will generate an **HathiTrust identifier** for each digital image or complex object that is unique to the whole of HathiTrust, using the namespace “mdl” assigned by HathiTrust. This identifier will be generated according to the HathiTrust ingest checklist. This identifier will be suitable for recalling a specific object from the HT system in the future.

[4.2.] MDL will generate a file of **Dublin Core (DC) metadata** for all objects being ingested. This file will be suitable to load into the HathiTrust catalog prior to the ingest of individual objects. At a minimum, these DC records will include:

- [4.2.1.] **Title**: a brief, though possibly not unique, descriptor of the item;
- [4.2.2.] **Other identifier**: identifiers used by the local institutions to retrieve the item, may not be unique; also any existing OCLC numbers will be included here;
- [4.2.3.] **Link**: a URL that can be used to navigate to this item within the local system from which it was extracted;
- [4.2.4.] **Description**: all other elements of DC present for an object will be passed through to HT;

[4.3.] We will include any **technical metadata** we get from local sources or from probes of the binary objects and their embedded metadata. This will be provided as **embedded XMP metadata** in each JPEG2000 in [3.2.].

[4.4.] All **descriptive and administrative metadata** will be wrapped into a **single METS file** for each digital image. In cases where multiple digital images comprise a single compound object, one METS file will apply to the whole compound object. The METS file will include the DC metadata described in [4.2.] as well as:

- [4.4.1.] **HT identifier**: built in [4.1.] uniquely identifying an object;
- [4.2.5.] **Provenance**: a set of PREMIS events that describe the chain of custody of this object;

[4.5.] The METS file and all associated binary files will be **ZIPed into a single package**. One such package will be created for each digital image or compound object.

[5.] Transfer

[5.1.] Collections of packages will be delivered to HT either over the net or via shipped hard drives.

[5.2.] Any hard drives shipped to HT will be returned for reuse once the transfer is complete.

[5.3.] MDL will retain copies of shipped packages until HT confirms receipt and successful ingest.

[6.] Ingest

[6.1.] Staff of the HT will be responsible for ingest of our packages.

[6.2.] We expect to be notified of the success or failure of ingest of each package.

[7.] Display & Retrieval

[7.1.] Images will not be displayed with greater than 1024 pixel resolution on their longest side on the HT system.

> *[7.1.] See “hot spot” issue {A} for further discussion.*

[7.2.] The HT description of the item will always include a link back to the item within the local system from which it came.

[7.3.] The original binary file associated with an item will be accessible via a HT API when given the HT or identifier for that item, with the following restriction:

- [7.3.1.] The original binary will only be made available outside the HT when a request is made presenting authentication that identifies the query as originating from theMDL.

[7.3.1.] This point may be particularly problematic given HT stance on sharing, yet it is important to smaller cultural heritage institutions. See "hot spot" issue {B} for further discussion.

[8.] Governance

[8.1.] The development team at MDL and HT will have authority to make decisions regarding the prototype on a day to day basis. All decisions will be recorded on the project management website.

[8.2.] The sponsors group will meeting monthly via phone conference to review progress of the project and affirm decisions made by the development team. Any decision not affirmed will be returned to the team for reconsideration. The sponsors, however, acknowledge that on a project with so little turnaround time, such reconsideration may not occur in time to have a great deal of impact on the outcome of the project.

[8.3.] The University of Minnesota is the sole voice of the project within the HathiTrust governance structure. The University agrees to represent interests of the sponsors in HT discussions.

4.4. MDL-HT Specifications for Reflections Continuous Tone Images

This document seeks to specify the details required for the packaging of MDL Reflections continuous tone images for the MDL-HT Image Prototype project. It is meant to be used in conjunction with the MDL-HT Image Ingest Prototype Guidelines which provide a definition of the interactions between the MDL and HathiTrust. This document will use the same numbering sequence, but only address items where more detail is needed than that supplied in the guidelines. Any “missing” numbers are likely found in the guidelines.

Background

MDL Reflections is a CONTENTdm system hosted by OCLC used by the Minnesota Digital Library as its primary catalog of content from around the state. While CONTENTdm does hold the descriptive and technical metadata associate with the images in the collection, it does not hold the master images themselves. Those are in a separate store on a server managed by the University of Minnesota Libraries. These master images are stored in uncompressed TIFF format.

MDL Reflections includes both simple continuous tone images with individual descriptions and more complex “compound objects” where a single description applies to a set of two or more images. At this stage of the MDL-HT Image Ingest Prototype project we are only seeking to transfer the simple “contone” images to HathiTrust.

Accomplishing this will require attention to six stages of processing: **extracting** master images and metadata from current repositories, **reformatting** the images to suit the preservation archive, **packaging** these binaries and associated metadata as required for ingest, **transferring** these packages to HathiTrust, seeing that these packages are **ingested** by HathiTrust, and providing for **display** of these images at HathiTrust and retrieval of the masters via API calls.

When “we” is used in this document it refers to MDL, and more specifically, to the development team of Bill, Jason, and Eric. In practice it almost always means Bill Tantzen. Thanks, Bill!

[1.] Content

[1.2.] The masters we have are uncompressed TIFF images. They will need to be transformed into JPEG2000 (JP2) images for this project. This will make them, effectively, no longer the “master” images used by MDL Reflections. This is acceptable for the prototype. See “hot spot” issue {C} for further discussion.

[1.3.] The UMN Libraries systems office has apparently generated MD5 checksums for theTIFF masters as part of its own integrity checking on the existing store. Since we must transcode to JP2 format, however, these checksums will have no validity for this project. We will have to generate new checksums for the JP2 versions.

[2.] Extract

[2.1.] Descriptive metadata available in Dublin Core (DC) format from an OAI-PMH harvest. Technical metadata is found in EXIF data with each image. Some further technical metadata recorded in CONTENTdm is not exported via OAI-PMH and will be ignored.

[2.2.] MDL Reflections descriptive metadata can be retrieved in XML format via an OAI-PMH harvest. While the EXIF data is not in XML format, it is quite accessible via ImageMagick and other tools.

[2.3.] The “MDL identifier” is available from the OAI data as one of the generic identifiers in DC. It can be identified as a three character lowercase code followed by a set of numerical digits.

[2.3.] Jason will provide Bill with a comprehensive list of the three character codes used in MDL identifiers.

[2.4.] The TIFF masters are available via on-campus file sharing from the store managed by the systems office. They will be copied via file sharing.

[2.5.] The TIFF masters are named using the MDL identifiers.

[2.6.] Prefix the MDL identifier with “reflections.” to create the unique project identifier for each item.

[3.] Reformat

[3.1.] Convert TIFF masters into JPEG2000 via ImageMagick. Use the MDL identifier as the filename, for example “umn123.jp2”. Confirm tool. What about kakadu?

[3.2.] Technical metadata will be created for each image and recorded in an XMP package that will be embedded in that same image via ImageMagick. Details will be saved for inclusion in a PREMIS event record, see [4.2.5.]. Confirm tool.

[3.3.] A new MD5 checksum of the resulting JPEG2000 image will be generated via unix command line md5 and saved for inclusion in PREMIS event record, see [4.2.5.]. Confirm tool.

[4.] Package

[4.1.] The “HathiTrust identifier” will be generated by adding the namespace “mdl.” as a prefix to the project identifier from [2.6.].

[4.2.] The DC metadata extracted via OAI-PMH will be passed through as the descriptive metadata for each item. This will include:

- [4.2.1.] The title as a element and will be left as-is.
- [4.2.2.] The identifiers present as elements and will be left as-is.
- [4.2.3.] The link will be one of many elements.

[4.2.3.] Note that HT will have to recognize that any such element that contains a URL should become an actionable link in the HT catalog. This could be limited to any identifier starting with the string “http://reflections.mndigital.org”.

- [4.2.4.] Other dc elements found in Reflections data include: description, date, source, format, subject, coverage, relation, publisher, and rights. All will be included as-is.

[4.4.] The METS file will be created using the MDL identifier as the file name, for example “umn123.xml”. This METS file will include:

- [4.4.1.] The HathiTrust identifier from [4.1.] will be used as the OBJID attribute value and as the ID attribute.
- [4.4.2.] A PREMIS event representing the initial ingest into MDL Reflections as a TIFF.
- [4.4.3.] A PREMIS event representing the conversion to JPEG2000.
- [4.4.4.] A PREMIS event representing the contribution of the item to HT.
- [4.4.5.] A section describing the structure of the contribution, which simply contains a pointer to the single image file.
- [4.4.6.] A section describing the single image file, including its type, creation date, size, and checksum.

[4.5.] Put the associated image and METS files (for example, "umn123.jp2" and "umn123.xml") into a single directory (for example, "umn123.package") and ZIP it using gzip into a single file (for example "umn123.package.zip"). Each zip file must be greater than 128KB in size. If any ZIPed file is smaller, then save the directory in question to bundle with other undersized objects into a single ZIP file.

[5.] Transfer

[5.1.] Files will be transferred via FAT32 formatted hard drives with USB interfaces.

[6.] Ingest

[6.2.] HT will notify Bill of the results of each ingest via email.

4.5. MDL-HT Specifications for Reflections Compound Objects

This document seeks to specify the details required for the packaging of MDL Reflections compound objects for the MDL-HT Image Prototype project. It is meant to be used in conjunction with the MDL-HT Image Ingest Prototype Guidelines which provide a definition of the interactions between the MDL and HathiTrust. This document will use the same numbering sequence, but only address items where more detail is needed than that supplied in the guidelines. Any "missing" numbers are likely found in the guidelines.

Background

MDL Reflections is a CONTENTdm system hosted by OCLC used by the Minnesota Digital Library as its primary catalog of content from around the state. While CONTENTdm does hold the descriptive and technical metadata associate with the images in the collection, it does not hold the master images themselves. Those are in a separate store on a server managed by the University of Minnesota Libraries. These master images are stored in uncompressed TIFF format.

MDL Reflections includes both simple continuous tone images with individual descriptions and more complex "compound objects" where a single description applies to a set of two or more images. We dealt with the simple continuous tone images in the last stage. At this stage of the MDL-HT Image Ingest Prototype project we are only seeking to transfer "compound" images to HathiTrust.

Accomplishing this will require attention to six stages of processing: **extracting** master images and metadata from current repositories, **reformatting** the images to suit the preservation archive, **packaging** these binaries and associated metadata as required for ingest, **transferring** these packages to HathiTrust, seeing that these packages are **ingested** by HathiTrust, and providing for **display** of these images at HathiTrust and retrieval of the masters via API calls.

When "we" is used in this document it refers to MDL, and more specifically, to the development team of Bill, Jason, and Eric. In practice it almost always means Bill Tantzen. Thanks, Bill!

[1.] Content

[1.2.] The masters we have are uncompressed TIFF images. They will need to be transformed into JPEG2000 (JP2) images for this project. This will make them, effectively, no longer the "master" images used by MDL Reflections. This is acceptable for the prototype. See "hot spot" issue {C} for further discussion.

[1.3.] The UMN Libraries systems office has apparently generated MD5 checksums for theTIFF masters as part of its own integrity checking on the existing store. Since we must transcode to JP2 format, however, these checksums will have no validity for this project. We will have to generate new checksums for the JP2 versions.

[2.] Extract

[2.1.] Descriptive metadata available in Dublin Core (DC) format from an OAI-PMH harvest. Technical metadata is found in EXIF data with each image. Some further technical metadata recorded in CONTENTdm is not exported via OAI-PMH and will be ignored.

[2.2.] MDL Reflections descriptive metadata can be retrieved in XML format via an OAI-PMH harvest. While the EXIF data is not in XML format, it is quite accessible via Kakadu and other tools.

[2.3.] The "MDL identifier" is available from the OAI data as one of the generic identifiers in DC. It can be identified as a three character lowercase code followed by a set of numerical digits.

[2.4.] The TIFF masters are available via on-campus file sharing from the store managed by the systems office. They will be copied via file sharing.

[2.5.] The TIFF masters are named using the MDL identifiers.

[2.5.] Note that each image has its own MDL identifier. Unfortunately the whole multi-page object does *not* have an MDL identifier. We will use the MDL identifier of the first page to represent the object in [2.6.].

[2.6.] Add "-all" to the end of the MDL identifier of the first page of the compound object to create a new MDL identifier for each compound object.

[2.6.] So if the first page had the MDL identifier "umn123" then MDL identifier for this compound object would be "umn123-all".

[2.7.] Prefix the MDL identifier for the compound object with "reflections." to create the unique project identifier for each compound object.

[2.7.] Continuing the example from [2.6.] the project identifier for the compound object would be "reflections.umn123-all".

[3.] Reformat

[3.1.] Convert continuous tone TIFF masters into JPEG2000 via Kakadu. Leave bi-tonal TIFF masters in TIFF format, but make sure to use G4 (fax) compression on these images, which may require reformatting via ImageMagick or some other tool. Use the MDL identifier as the filename, for example "umn123.jp2".

[3.2.] Technical metadata will be created for each image and recorded in an XMP package that will be embedded in that same image via Kakadu. Some details will be saved for inclusion in a PREMIS event record, see [4.2.5.].

[3.3.] A new MD5 checksum of the resulting JPEG2000 or TIFF image will be generated via unix command line md5 and saved for inclusion in PREMIS event record, see [4.2.5.].

[4.] Package

[4.1.] The "HathiTrust identifier" will be generated by adding the namespace "mdl." as a prefix to the project identifier from [2.6.].

[4.1.] Continuing our example from [2.6.] the HathiTrust identifier for the compound object would be "mdl.reflections.umn123-all".

[4.2.] The DC metadata extracted via OAI-PMH will be passed through as the descriptive metadata for each compound object. This will include:

- [4.2.1.] The title as a element and will be left as-is.
- [4.2.2.] The identifiers present as elements and will be left as-is.
- [4.2.3.] The link will be one of many elements.

[4.2.3.] Note that HT will have to recognize that any such element that contains a URL should become an actionable link in the HT catalog. This could be limited to any identifier starting with the string "http://reflections.mndigital.org".

- [4.2.4.] Other dc elements found in Reflections data include: description, date, source, format, subject, coverage, relation, publisher, and rights. All will be included as-is.

[4.3.] We will include any technical metadata we get from local sources or from probes of the binary objects and their embedded metadata. This will be provided as embedded XMP metadata in each JPEG2000 or TIFF in [3.2.].

[4.4.] The METS file will be created using the MDL identifier of the compound object as the file name, for example "umn123-all.xml". This METS file will include:

- [4.4.1.] The HathiTrust identifier from [4.1.] will be used as the OBJID attribute value and as the ID attribute.
- [4.4.2.] A PREMIS event representing the initial ingest into MDL Reflections as a TIFF.
- [4.4.3.] A PREMIS event representing the conversion to JPEG2000.
- [4.4.4.] A PREMIS event representing the contribution of the item to HT.
- [4.4.5.] A section describing the structure of the compound object and with pointers to the files involved. This includes pointers to the DC and OCR files described in [4.5.].
- [4.4.6.] A section describing each file, including its type, creation date, size, and checksum.

[4.5.] For each "page" of the compound object, two files will be created as needed:

- [4.5.1.] If that page has associated Dublin Core metadata, a metadata file will be created using the MDL identifier of that page followed by ".dc" (for example "umn123.dc") which will include an XML representation of the DC metadata for that page.
- [4.5.2.] If transcript or OCR data exists for that page, a text file will be created using the MDL identifier of that page followed by ".ocr" (for example "umn123.ocr") which will contain a UTF-8 text representation of the contents of that page.

[4.6.] Put the METS file and all associated image files files (for example, "umn123.jp2", "umn124.jp2", "umn123.dc", "umn124.dc", "umn123.ocr", "umn124.ocr", and "umn123-all.xml") into a single directory (for example, "umn123-all.package") and ZIP it using gzip into a single file (for example "umn123-all.package.zip"). Each zip file must be greater than 128KB in size. If any ZIPed file is smaller, then save the directory in question to bundle with other undersized objects into a single ZIP file.

[5.] Transfer

[5.1.] Files will be transferred via FAT32 formatted hard drives with USB interfaces.

[6.] Ingest

[6.2.] HT will notify Bill of the results of each ingest via email.

4.6. MDL-HT Specifications for MHS objects

This document seeks to specify the details required for the packaging of Minnesota Historical Society objects for the MDL-HT Image Prototype project. It is meant to be used in conjunction with the MDL-HT Image Ingest Prototype Guidelines which provide a definition of the interactions between the MDL and HathiTrust. This document will use the same numbering sequence, but only address items where more detail is needed than that supplied in the guidelines. Any "missing" numbers are likely found in the guidelines.

Background

The Minnesota Historical Society (MHS) content management system (CMS) contains over 15,000 TIFF and JPEG images. Only a subset of these, the "Collections Online", will be preserved in HathiTrust (HT) as part of this prototype effort. This should be roughly 9,000 images the CMS, called EMu, is already sharing with the public. Some of these images are part of compound objects, so the total set of catalog records is probably closer to 8,000. It is these catalog records that are to be transformed into Dublin Core (DC) records.

Descriptive metadata will come from the catalog side of the EMu system. It will be exported as XML data.

This mapping to Dublin Core (DC) forms only a part of the transformation necessary for theMDL-HT Digital Preservation Project. In addition to creating the DC records, those records will be embedded in METS files and packed together with binary images for shipment to HathiTrust. There is a separate "mapping" document that describes the path from EMu to DC in more detail.

Technical metadata is in EXIF, IPTC, and XMP data already embedded in the binary files. We can extract what we need for the HT XMP.

MHS Collections Online includes both simple continuous tone images with individual descriptions and more complex "compound objects" where a single description applies to a set of two or more images. We will be including both types of content in this process.

Accomplishing this will require attention to six stages of processing: **extracting** master images and metadata from the MHS EMu system, **reformatting** the images to suit the preservation archive, **packaging** these binaries and associated metadata as required for ingest, **transferring** these packages to HathiTrust, seeing that these packages are **ingested** by HathiTrust, and providing for **display** of these images at HathiTrust and retrieval of the masters via API calls.

When "we" is used in this document it refers to MDL, and more specifically, to the development team of Bill, Jason, and Eric. In practice it almost always means Bill Tantzen. Thanks, Bill!

[1.] Content

[1.2.] The masters supplied by MHS are both uncompressed TIFF images and compressedJPEG images. The TIFFs will need to be transformed into JPEG2000 (JP2) images for this project. This will make them, effectively, no longer the "master" images used by MDLReflections. This is acceptable for the prototype.

See “hot spot” issue {C} for further discussion. HT has confirmed that the JPEG images can be sent in JPEG format, in other words: compressed.

[1.3.] MHS generated MD5 checksums many of the TIFF and JPEG masters as part of its own integrity checking on the existing store. Confirm the accuracy of the MHS master by using the checksum from the “Multimedia” table of the record if it is present. Save this checksum for inclusion in a PREMIS fixity check event, see [4.4.] Since we must transcode to JP2 format and generate new XMP information, however, these checksums will only be of use in assuring that we received uncorrupted copies of this material from MHS. We will have to generate new checksums for both the JP2 and JPEG images.

[2.] Extract

[2.1.] Descriptive metadata available in an EMu-specific XML format from MHS. Technical metadata is found in IPTC, EXIF, and XMP data with each image. Some further technical metadata recorded the EMu multimedia system will be ignored.

[2.2.] MHS descriptive metadata can be retrieved in XML format via a manual export from EMu. While the EXIF data is not in XML format, it is quite accessible via Kakadu and other tools.

[2.3.] The “MHS CATIRN identifier” is available from the descriptive XML data as the “CatalogIrn” atom.

[2.4.] The TIFF and JPEG masters will be made available on a hard disk from MHS.

[2.4.] Note that this disk will contain other images that are not to be shared with HT or any other party. After the required masters have been extracted, the data on this disk should be destroyed.

[2.5.] The TIFF and JPEG masters are referred to by name in the “multimedia” table of the descriptive XML.

[2.6.] The “MHS identifier” will be the CatalogIrn for each MHS object.

[2.6.] The CatalogIrn is either a four or eight digit number, like “1234”.

[2.7.] Prefix the MHS identifier for the MHS object with “mhs.catirn.” to create the unique project identifier for each compound object.

[2.7.] Continuing our example from [2.6.] the project identifier for the MHS object would be something like “mhs.catirn.1234”.

[3.] Reformat

[3.1.] Convert continuous tone TIFF masters into JPEG2000 via Kakadu. Leave JPEG masters in JPEG format, but make sure to check XMP for these, which may require reformatting via ImageMagick or some other tool. Use the given filename for each of these, though the extension may have to change, for example “mh218s.jp2”.

[3.2.] Technical metadata will be created for each image and recorded in an XMP package that will be embedded in that same image via Kakadu or ImageMagick. Some details will be saved for inclusion in a PREMIS event record, see [4.4.].

[3.3.] A new MD5 checksum of the resulting JPEG2000 or JPEG image will be generated via unix command line md5 and saved for inclusion in PREMIS event record, see [4.4.].

[4.] Package

[4.1.] The "HathiTrust identifier" will be generated by adding the namespace "mdl." as a prefix to the project identifier from [2.6.].

[4.1.] Continuing our example from [2.6.] the HathiTrust identifier for the compound object would be "mdl.mhs.catirn.1234".

[4.2.] DC metadata will be prepared by mapping the descriptive XML from EMu and become the descriptive metadata for each MHS object. This will include:

- [4.2.1.] The title as a dc:title.
- [4.2.2.] The identifiers present as dc:identifier.
- [4.2.3.] The link will be one of many dc:identifier elements.

[4.2.3.] Note that HT will have to recognize that any such element that contains a URL should become an actionable link in the HT catalog. This could be limited to any identifier starting with the string "http://collections.mnhs.org".

- [4.2.4.] Other dc elements found in Reflections data include: description, date, source, format, subject, coverage, relation, publisher, and rights.

[4.2.] See the separate "mapping" document for details on the mapping from EMu's exported XML to the DC we want for this project.

[4.3.] We will include any technical metadata we get from local sources or from probes of the binary objects and their embedded metadata. This will be provided as embedded XMP metadata in each JPEG2000 or JPEG in [3.2.].

[4.4.] The METS file will be created using the project identifier of the MHS object as the file name, for example "mhs.catirn.1234.xml". This METS file will include:

- [4.4.1.] The HathiTrust identifier from [4.1.] will be used as the OBJID attribute value and as the ID attribute.
- [4.4.2.] A PREMIS event representing the fixity check performed on the original master file.
- [4.4.3.] A PREMIS event representing the conversion to JPEG2000 or modification of the JPEG.
- [4.4.4.] A PREMIS event representing the contribution of the item to HT.
- [4.4.5.] A section describing the structure of the compound object and with pointers to the files involved. This includes pointers to the DC.
- [4.4.6.] A section describing each file, including its type, creation date, size, and checksum.

[4.5.] Put the METS file and all associated image files files (for example, "3412.jp2", "3413.jp2", "3414.jpg", and "mhs.catirn.1234.xml") into a single directory (for example, "mhs.catirn.1234") and ZIP it using gzip into a single file (for example "mhs.catirn.1234.tar.gz"). Each zip file must be greater than 128KB in size. If any ZIPed file is smaller, then save the directory in question to bundle with other undersized objects into a single ZIP file.

[5.] Transfer

[5.1.] Files will be transferred via FAT32 formatted hard drives with USB interfaces.

[6.] Ingest

[6.2.] HT will notify Bill of the results of each ingest via email.

4.7. Sample MDL METS file for HT

```
<?xml version="1.0" encoding="UTF-8"?>
<METS:mets xmlns:METS="http://www.loc.gov/METS/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:PREMIS="info:lc/xmlns/premis-v2"
xmlns:tiff="http://ns.adobe.com/tiff/1.0" xmlns:MDL="http://www.mndigital.org/premis_extension"
xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/"
xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd
http://www.loc.gov/MARC21/slim http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd info:lc/xmlns/premis-v2
http://www.loc.gov/standards/premis/v2/premis-v2-0.xsd http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd" OBJID="reflections.umn79102-all">
  <METS:metsHdr CREATEDATE="2010-12-29T18:28:50-05:00" RECORDSTATUS="NEW">
    <METS:agent ROLE="CREATOR" TYPE="ORGANIZATION">
      <METS:name>Minnesota Digital Library</METS:name>
    </METS:agent>
  </METS:metsHdr>
  <METS:dmdSec ID="DMD1">
    <METS:mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="MDL metadata">
      <METS:xmlData>
        <oai_dc:dc>
          <dc:coverage>Buhl; Chisholm; Coleraine; Ely; Eveleth; Gilbert; Hibbing; Mountain Iron; Soudan; Tower;
Virginia</dc:coverage>
          <dc:coverage>Mesabi Iron Range; Vermilion Iron Range</dc:coverage>
          <dc:coverage>St. Louis</dc:coverage>
          <dc:coverage>Minnesota</dc:coverage>
          <dc:coverage>United States</dc:coverage>
          <dc:creator>Oliver Iron Mining Company; United States Steel Corporation</dc:creator>
          <dc:date>1928</dc:date>
          <dc:description>1928 mapbook featuring both open pit and underground mining operations on the Mesabi and Vermilion
Iron Ranges of Minnesota.</dc:description>
          <dc:format>Atlases</dc:format>
          <dc:identifier>1993.2736</dc:identifier>
          <dc:identifier>http://cdm15160.contentdm.oclc.org/u?/irrc,2983</dc:identifier>
          <dc:publisher>Iron Range Research Center, 1005 Discovery Drive, Chisholm, Minnesota 55719;
http://mndiscoverycenter.com/research-center</dc:publisher>
          <dc:rights>Use of this image is governed by U.S. and international copyright law. Please contact the Iron Range Research
Center, Chisholm, MN, for more information in regard to this image, online at
http://mndiscoverycenter.com/research-center/archive</dc:rights>
          <dc:source>Oliver Iron Mining Company; United States Steel Corporation</dc:source>
          <dc:source>49 x 65</dc:source>
          <dc:subject>Business and industry</dc:subject>
          <dc:subject>Iron mines and mining</dc:subject>
          <dc:subject>United States Steel; Iron Mining; Adams Mine; Alpena Mine; Arcturus Mine; Auburn Mine; Burt Mine; Can-
isteo Mine; Carson Lake Mine; Chisholm Mine; Clark Mine; Culver Mine; Day Mine; Deacon Mine; Duncan Mine; Ely Mine;
Fayal Mine; Fraser Mine; Glen Mine; Godfrey Mine; Hartley St. Clair Mine; Hill Mine; Holman Mine; Hull Rust Mine; Judd
Mine; Kerr Mine; Leonard Mine; Leonidas Mine; Lone Jack Mine; McEwan Mine; Minnewas Mine; Missabe Mt. Mine; Monroe
Mine; Morrison Mine; Morrison Mine; Mt. Iron Mine; Myers Mine; North Star Mine; Ohio Mine; Ordean Mine; Palmer Mine;
Philbin Mine; Pillsbury Mine; Pioneer Mine; Pool Mine; Prindle Mine; Rouchlaou Mine; Sauntry Mine; Seller Mine; Sharon
Mine; Shaw Moose Mine; Shiras Mine; Sibley Mine; Soudan Mine; Spruce Mine; Stephens Mine; Sulivan Mine; Sweeney Mine;
Walker Mine; Wellington Mine; Bovey</dc:subject>
          <dc:title>Oliver Iron Mining Company Mapbook</dc:title>
          <dc:source>Iron Range Research Center</dc:source>
        </oai_dc:dc>
      </METS:xmlData>
    </METS:mdWrap>
  </METS:dmdSec>
  <METS:dmdSec ID="DC00000001">
  <METS:mdWrap MIMETYPE="text/xml" MDTYPE="DC" LABEL="Dublin Core metadata for umn79102.tif">
   <METS:xmlData>
    <oai_dc:dc>
     <dc:description>bc797078d8c6188255fac1e98b58ef83</dc:description>
     <dc:identifier>http://cdm15160.contentdm.oclc.org/u?/irrc,2548</dc:identifier>
     <dc:publisher>Iron Range Research Center, 1005 Discovery Drive, Chisholm, Minnesota 55719; http://mndiscoverycenter.com/research-center</dc:publisher>
     <dc:rights>Use of this image is governed by U.S. and international copyright law. Please contact the Iron Range Research Center, Chisholm, MN, for more information in regard to this image, online at http://mndiscoverycenter.com/research-center/archive</dc:rights>
     <dc:title>Front cover</dc:title>
    </oai_dc:dc>
   </METS:xmlData>
  </METS:mdWrap>
 </METS:dmdSec>
 <!-- remainder of the 435 METS:dmdSec elements removed for brevity -->
 <METS:amdSec>
  <METS:digiprovMD ID="premis1">
   <METS:mdWrap MDTYPE="PREMIS">
    <METS:xmlData>
     <PREMIS:premis version="2.0">
      <PREMIS:object xsi:type="PREMIS:representation">
       <PREMIS:objectIdentifier>
        <PREMIS:objectIdentifierType>MDL</PREMIS:objectIdentifierType>
        <PREMIS:objectIdentifierValue>reflections.umn79102-all</PREMIS:objectIdentifierValue>
       </PREMIS:objectIdentifier>
       <PREMIS:preservationLevel>
        <PREMIS:preservationLevelValue>1</PREMIS:preservationLevelValue>
       </PREMIS:preservationLevel>
      </PREMIS:object>
      <PREMIS:event>
       <PREMIS:eventIdentifier>
        <PREMIS:eventIdentifierType>UUID</PREMIS:eventIdentifierType>
        <PREMIS:eventIdentifierValue>C62D4AC6-13AB-11E0-8A1D-C740821A552F</PREMIS:eventIdentifierValue>
       </PREMIS:eventIdentifier>
       <PREMIS:eventType>capture</PREMIS:eventType>
       <PREMIS:eventDateTime>2009-09-22T12:33:05-05:00</PREMIS:eventDateTime>
       <PREMIS:eventDetail>Initial capture of TIFF master</PREMIS:eventDetail>
       <PREMIS:linkingAgentIdentifier>
        <PREMIS:linkingAgentIdentifierType>tool</PREMIS:linkingAgentIdentifierType>
        <PREMIS:linkingAgentIdentifierValue>Phase One</PREMIS:linkingAgentIdentifierValue>
        <PREMIS:linkingAgentRole>scanner</PREMIS:linkingAgentRole>
       </PREMIS:linkingAgentIdentifier>
       <PREMIS:linkingAgentIdentifier>
        <PREMIS:linkingAgentIdentifierType>MARC21 Code</PREMIS:linkingAgentIdentifierType>
        <PREMIS:linkingAgentIdentifierValue>MnU</PREMIS:linkingAgentIdentifierValue>
        <PREMIS:linkingAgentRole>Executor</PREMIS:linkingAgentRole>
       </PREMIS:linkingAgentIdentifier>
      </PREMIS:event>
      <PREMIS:event>
       <PREMIS:eventIdentifier>
        <PREMIS:eventIdentifierType>UUID</PREMIS:eventIdentifierType>
        <PREMIS:eventIdentifierValue>C62D5264-13AB-11E0-8A1D-C740821A552F</PREMIS:eventIdentifierValue>
       </PREMIS:eventIdentifier>
       <PREMIS:eventType>image compression</PREMIS:eventType>
       <PREMIS:eventDateTime>2010-12-13T14:27:51-05:00</PREMIS:eventDateTime>
       <PREMIS:eventDetail>Convert TIFF master to compressed format</PREMIS:eventDetail>
       <PREMIS:linkingAgentIdentifier>
        <PREMIS:linkingAgentIdentifierType>MARC21 Code</PREMIS:linkingAgentIdentifierType>
        <PREMIS:linkingAgentIdentifierValue>MnU</PREMIS:linkingAgentIdentifierValue>
        <PREMIS:linkingAgentRole>Executor</PREMIS:linkingAgentRole>
       </PREMIS:linkingAgentIdentifier>
       <PREMIS:linkingAgentIdentifier>
        <PREMIS:linkingAgentIdentifierType>tool</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>kakadu/kdu_compress v6.4.1</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>software</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>tool</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>ImageMagick v6.6.3-1</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>software</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
</PREMIS:event>
<PREMIS:event>
  <PREMIS:eventIdentifier>
    <PREMIS:eventIdentifierType>UUID</PREMIS:eventIdentifierType>
    <PREMIS:eventIdentifierValue>C62D63C6-13AB-11E0-8A1D-C740821A552F</PREMIS:eventIdentifierValue>
  </PREMIS:eventIdentifier>
  <PREMIS:eventType>message digest calculation</PREMIS:eventType>
  <PREMIS:eventDateTime>2010-12-13T14:27:51-05:00</PREMIS:eventDateTime>
  <PREMIS:eventDetail>Calculation of page-level md5 checksums</PREMIS:eventDetail>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>MARC21 Code</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>MnU</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>Executor</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>tool</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>perl v5.10.0/Digest::MD5 v2.51</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>software</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
</PREMIS:event>
<PREMIS:event>
  <PREMIS:eventIdentifier>
    <PREMIS:eventIdentifierType>UUID</PREMIS:eventIdentifierType>
    <PREMIS:eventIdentifierValue>C62D6A92-13AB-11E0-8A1D-C740821A552F</PREMIS:eventIdentifierValue>
  </PREMIS:eventIdentifier>
  <PREMIS:eventType>source mets creation</PREMIS:eventType>
  <PREMIS:eventDateTime>2010-12-29T18:28:50-05:00</PREMIS:eventDateTime>
  <PREMIS:eventDetail>Creation of HathiTrust source METS file</PREMIS:eventDetail>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>MARC21 Code</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>MnU</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>Executor</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>tool</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>makemets_compound.pl v1.1</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>software</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
</PREMIS:event>
<PREMIS:event>
  <PREMIS:eventIdentifier>
    <PREMIS:eventIdentifierType>UUID</PREMIS:eventIdentifierType>
    <PREMIS:eventIdentifierValue>C62D738E-13AB-11E0-8A1D-C740821A552F</PREMIS:eventIdentifierValue>
  </PREMIS:eventIdentifier>
  <PREMIS:eventType>zip archive creation</PREMIS:eventType>
  <PREMIS:eventDateTime>2010-12-29T18:28:50-05:00</PREMIS:eventDateTime>
  <PREMIS:eventDetail>Creation of ZIP for HathiTrust</PREMIS:eventDetail>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>MARC21 Code</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>MnU</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>Executor</PREMIS:linkingAgentRole>
  </PREMIS:linkingAgentIdentifier>
  <PREMIS:linkingAgentIdentifier>
    <PREMIS:linkingAgentIdentifierType>tool</PREMIS:linkingAgentIdentifierType>
    <PREMIS:linkingAgentIdentifierValue>makemets_compound.pl v1.1</PREMIS:linkingAgentIdentifierValue>
    <PREMIS:linkingAgentRole>software</PREMIS:linkingAgentRole>
```

```
              </PREMIS:linkingAgentIdentifier>
            </PREMIS:event>
          </PREMIS:premis>
        </METS:xmlData>
      </METS:mdWrap>
    </METS:digiprovMD>
  </METS:amdSec>
  <METS:fileSec>
    <METS:fileGrp ID="FG1" USE="image">
      <METS:file ID="IMG00000001" MIMETYPE="image/jp2" SEQ="00000001" CREATED="2010-12-13T14:27:51-05:00" SIZE="58004409" CHECKSUM="0f7da5033cb39305fe8d63312b10328c" CHECKSUMTYPE="MD5" DMDID="DC00000001">
        <METS:FLocat LOCTYPE="OTHER" OTHERLOCTYPE="SYSTEM" xlink:href="umn79102.jp2"/>
      </METS:file>
      <!-- remainder of the 435 METS:file elements removed for brevity -->
    </METS:fileGrp>
  </METS:fileSec>
  <METS:structMap ID="SM1" TYPE="physical">
    <METS:div TYPE="volume">
      <METS:div ORDER="1" TYPE="page">
        <METS:fptr FILEID="IMG00000001"/>
        <METS:fptr FILEID="OCR00000001"/>
      </METS:div>
      <!-- remainder of the 435 METS:div elements removed for brevity -->
    </METS:div>
  </METS:structMap>
</METS:mets>
```