Minnesota Digital Library Preservation Options Report

3 August 2011

Prepared by Eric Celeste with Marian Rengel

Eric Celeste 1993 Lincoln Avenue Saint Paul, MN 55105 651-323-2009 efc@umn.edu http://eric.clst.org

Table of Contents

0.1. 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 ensuring long term access to 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.

At a January 2010 meeting, stakeholders agreed to move the discussion of preservation of digital resources from the hypothetical to the practical by initiating a pilot study of preservation options. These options should be collaborative, meet the needs of the primary partners (the Minnesota digital Library, the Minnesota Historical Society, the University of Minnesota Libraries, and Minitex) and be extensible to other partners (e.g., Minnesota Public Radio, Twin Cities Public Television, and county and local historical societies).

The 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 August 2010 with the final report, “MDL Digital Preservation Demonstration Project: Digital Image Preservation Needs” (http://www.mndigital.org/projects/preservation/needs.pdf). The August 2010 report positioned the MDL to take the project into its next phase – prototype and demonstration.

The MDL then conducted a pilot project with HathiTrust to demonstrate the technological and organizational potential of a scalable digital preservation program and service for cultural heritage stewardship across Minnesota. The current follow-up project was designed to provide MDL with experience with a number of other preservation alternatives so that it could make the best possible choice for a Minnesota preservation archive.

0.2. Executive Summary

During March and April 2011, MDL conducted three trials by sending the same set of 6,000 images to MetaArchive, OCLC, and UC3/Merritt. During this time, MDL also participated in demonstrations of Tessella's Safety Deposit Box and UC3's Merritt at the Minnesota Historical Society and had conversations with the leadership of Chronopolis and DAITSS. These experiences were used to complete a Preservation Options Matrix (http://www.mndigital.org/projects/preservation/matrix.html) so that all these services could be compared to the findings from the MDL-HathiTrust Image Prototype Project of late 2010.

This MDL Preservation Options report describes the findings of the explorations with these organizations and elaborates on some of their scores in the Preservation Options Matrix. This report and the matrix are companion documents that are best understood alongside each other.

UC3/Merritt and MetaArchive offer a fitting combination of features, simplicity, business model, and strategic alliance to the MDL. OCLC's Digital Archive service may be too limited for MDL needs, and though OCLC pricing was attractive for smaller amounts of data, it becomes quite expensive when collections grow substantially. Chronopolis and DAITSS are both in the midst of generational shifts in their platforms that may make them more attractive to MDL in years to come, but Chronopolis also presents a tenuous business model and DAITSS does not offer a hosted solution. Tessella, too, may become more interesting to the MDL, but as of now, for MDL needs, has neither the necessary pricing guidance or requirements that would justify Tessella's full feature set.

It has become evident that in the near term MDL could require 5 to 10TB for the preservation of digital objects. Some of the providers in this study are uncomfortable holding the only copy of digital masters for the MDL, though the MDL will probably not have the option of holding a whole set of masters locally due to the size of this collection. The MDL will need to carefully weigh the single point of failure this lack of local holdings will create for its preservation archiving service.

0.3. Process

In February 2011, the sponsors of the MDL preservation options project met with MDL staff and consultants to determine attributes by which to evaluate potential solutions and to assign relative weights of these attributes. They built the Preservation Options Matrix (http://www.mndigital.org/projects/preservation/matrix.html) and came up with initial scores for the HathiTrust prototype project that had been conducted during the preceding months. The "preservation options matrix," as it came to be called, included attributes along a variety of dimensions including fitness for purpose, access and preservation, economics, and strategy and positioning.

During March and April, MDL consultant and the author of this report, Eric Celeste, conducted in-depth explorations of MetaArchive, OCLC's Digital Archive, and UC3/Merritt through trial loads of a subset of data from the MDL's Minnesota Reflections collection, stored at the University of Minnesota, which had been generated during the MDL-HathiTrust prototype project. For each trial load, the MDL developed a workflow to move the data from Minnesota into the alternative systems, demonstrated that workflow by moving about 6,000 images into each system, and worked with the vendors of those systems to test retrieval from them.

In addition, the consultant attended a presentation by Tessella at the Minnesota History Center and conducted interviews with staff of Chronopolis and DAITSS to make sure they were scored in the preservation matrix as well.

The Preservation Options Matrix was revised to reflect these conversations and trials. This report was written in May and June 2011 to complete the project. The matrix and report are highly subjective and intended only to guide a conversation among the sponsors; they do not represent any decision by the MDL. Anyone outside of the MDL reading this report should note that all rankings and judgments rendered in this report are highly skewed to address MDL requirements and should not be interpreted as a claim about the relative merits of the systems being discussed.

0.4. Contact Information

There were many participants in this project, this page captures contact information for only some of them.

MDL http://www.mndigital.org	**Bill DeJohn**, Minitex, w-dejo@umn.edu **Bob Horton**, Minnesota Historical Society, robert.horton@mnhs.org **John Butler**, University of Minnesota, j-butl@umn.edu
Chronopolis http://chronopolis.sdsc.edu	**David Minor**, San Diego Supercomputer Center, minor@sdsc.edu
DAITSS http://daitss.fcla.edu	**Priscilla Caplan**, University of Florida, pcaplan@ufl.edu
HathiTrust http://www.hathitrust.org	**John Wilkin**, University of Michigan, jpwilkin@umich.edu **John Weise**, University of Michigan, jweise@umich.edu
MetaArchive http://www.metaarchive.org	**Katherine Skinner**, Educopia Institute, katherine.skinner@metaarchive.org **Matt Schultz**, Educopia Institute, matt.schultz@metaarchive.org
OCLC Digital Archive http://www.oclc.org/digitalarchive/	**Taylor Surface**, OCLC, surface@oclc.org **Ron Gardner**, OCLC, gardnerr@oclc.org
Tessella Safety Deposit Box http://www.digital-preservation.com	**Mark Evans**, Tessella, mark.evans@tessella.com
UC3/Merritt http://merritt.cdlib.org	**Perry Willett**, University of California, perry.willett@ucop.edu **Patricia Cruse**, University of California, patricia.cruse@ucop.edu

1. Conversations

The MDL could not mount trials of all the services it wanted to review, so some services were only assessed by means of analyzing descriptions on the Web and talking with staff to make sure the information in the preservation matrix was reasonably accurate and clearly represented the services. This section is arranged alphabetically by service or corporate name.

1.1. Chronopolis

Interestingly enough, the Chronopolis team was present on a phone call with the Merritt team. UC3 is considering using Chronopolis as a storage option for Merritt. However, Chronopolis also offers its own preservation service and the MDL wanted to understand that offering so arranged a separate conversation with the Chronopolis team.

Chronopolis provides a geographically distributed and replicated preservation environment that shares some attributes with MetaArchive. However, instead of a LOCKSS backbone, Chronopolis uses the Storage Resource Broker (SRB) and is moving to an integrated Rule Oriented Data System (iRODS) for future generations. In place of the voting mechanism found in LOCKSS, Chronopolis provides an active checksum validation service that checks every object every month. Copies of material are kept at three geographically dispersed locations.

Submissions to Chronopolis are made using the BagIt packaging specification. They have a concept of “full bags” or bags that are complete and whose content is not expected to change; and “holey bags” or bags that are filled over time as the content is developed. A holey bag requires that content be at least initially available for harvest from the Web, though unlike LOCKSS, the source Website does not need to be maintained after the preservation harvest and ingest is complete.

Manifests for ingest are provided in XML form. Filename issues are probably the most common stumbling block for properly formatted manifests. Logs from the ingest are available and include checksums, but no automated confirmations are sent. Instead subscriber staff call to confirm successful loads. Chronopolis maintains an audit record that is available to subscribers without Chronopolis staff intervention. Items may be downloaded one-by-one without intervention, but mass recovery of content does require staff from Chronopolis to intervene and help out.

In addition to storage costs, participation in Chronopolis will likely include a $1,500 annual maintenance fee by the time the MDL were to get involved, but this would also include 80 hours of support from the staff.

1.2. DAITSS

The DAITSS project of the Florida Center for Library Automation is actually in the midst of a good deal of transition. It is important to realize, first of all, that there is no hosted DAITSS solution available. Instead, DAITSS is a set of software that the FCLA has made available so others can implement their own preservation solutions.

DAITSS1 was built on a Java foundation and turned out to be quite difficult to install and maintain. In fact, no other institution ever successfully ran an instance of DAITSS1. This experience encouraged the FCLA team to completely re-implement DAITSS. DAITSS2 will be based on Ruby and be provided as a set of discrete tools that are more loosely coupled. Some of the components will be community supplied, including DROID, FITS, JHOVE, and JHOVE2. DAITSS2 will use a "bundler" to make installation simpler.

The DAITSS2 SIP will demand a METS descriptor and structural references that are probably somewhat similar to MDL work with HathiTrust. Once the SIP is validated, any other errors in processing are considered DAITSS' "fault" and not the users. DAITSS does track PREMIS events as well as internal operational events, but only the PREMIS events are available for dissemination.

The system will provide five basic services: ingest, disseminate, refresh (migrating items to new formats), withdraw, and peek (a kind of lite dissemination). Dissemination in DAITSS1 always required staff intervention and could take days, but in DAITSS2 it should be more automated and responsive.

While FCLA is prohibited by Florida law from providing services to any entity outside the state, there may be a third party that builds a hosted DAITSS2 service. The MDL will watch for that possibility.

1.3. Tessella

The Minnesota Historical Society was reviewing Tessella for its own projects and invited the MDL consultant to sit in on the presentation being made at the Minnesota History Center. Tessella is a commercial provider that arrived at preservation services by way of the pharmaceuticals industry. Based in Europe, Tessella is stronger among national libraries there than in the U.S. Tessella was a driving force in the development of the open source DROID (digital object record identification) tool for automated file format identification.

The Tessella system enforces a hierarchy where files make up manifestations which make up records. Files have fixity, identifiers, and technical metadata; manifestations have structural metadata and information about their environment; records include descriptive metadata, context, provenance, and other characteristics. Fixity is applied at the file level. This is functionally similar to that of HathiTrust and Merritt, though it sounded a bit more formalized in Tessella. For example, metadata is migrated by Tessella via XSLT, so proper schemas for metadata are very helpful.

The “Safety Deposit Box,” as they call this system, is now in its fourth generation. It is built on Java incorporating a number of open source tools like DROID and technical registries like PRONOM. While a commercial product, the code can be put in escrow for clients. The sales presentation touched on all the appropriate technical topics (even “microservices”) and presented a picture of a system that could do just about everything the MDL is considering. The MDL, however, needs to learn more about Tessella.

2. Trials

The MDL intended to run two trials during March and April 2011 but ended up conducting three trials. The MDL consultant sent data to MetaArchive, OCLC, and Merritt. Details follow about each experience, but first a few words about what all the trials had in common.

Jason Roy, University of Minnesota Libraries, selected a subset of Minnesota Reflections data to include in the trials. He included four collections: the Basilica of St. Mary, the Blue Earth County Historical Society, the Minnesota Streetcar Museum, and the Nicollet County Historical Society.

While this data was coming from Reflections, which runs on CONTENTdm, MDL personnel purposely chose not use the CONTENTdm-specific processes available from MetaArchive and OCLC to transfer the content. Instead, MDL personnel wanted to force themselves to use a process that did not assume CONTENTdm as a starting place. In fact, MDL TIFF master files are not held in CONTENTdm, so they had to retrieve the masters from University of Minnesota storage servers. While Minnesota Reflections' CONTENTdm data reported that these four collections amounted to 5,784 images, there were actually 6,838 master images in the trial set. This was primarily due to the fact that there were some items among the masters (back sides of postcards, sheet music, etc.) that were never included in Minnesota Reflections.

The metadata prepared for each item was essentially whatever could be retrieved from Reflections along with some information about the fixity checksums for each item. In rare cases where the MDL holds masters for objects not in Reflections, there was no descriptive information about the items at all. Even when descriptions were found, project staff left these in a format specific to CONTENTdm, but embedded in an XML wrapper. These trials did not require that the metadata be correct or complete, so staff just made sure they had something for each record, but did not worry about what it was.

The dataset for all three trials consisted of 6,838 TIFF images along with an equivalent number of XML files, for a total of 13,676 files. The total amount of data to be sent was approximately 335GB.

During Fall 2010, the MDL also built a prototype preservation archive with HathiTrust (see report at http://www.mndigital.org/projects/preservation/fullReport.pdf). A brief recap of the issues with that prototype follows in section 2.4 of this report.

2.1. MetaArchive

While the MDL consultant contacted OCLC and MetaArchive at the same time, the MetaArchive team was prepared to get a project going quickly, so they ended up being the first trial.

MetaArchive is a preservation archive built on the LOCKSS platform by the Educopia Institute. All of the technical requirements for MetaArchive derive from LOCKSS requirements and a good bit of the MetaArchives documentation is actually LOCKSS documentation. However, MetaArchive is a separate network of LOCKSS caches from a network that many will know focused on preserving electronic journals. MetaArchive is focused on preserving archival collections that are generally larger than ejournals. The "private network" being run by Educopia for MetaArchive is made up of very large LOCKSS caches with an impressive collective capacity. Certainly the 335GB the MDL was bringing to the trial posed no issues for MetaArchive.

The practice at MetaArchive is to divide collections into "archival units" of 10-20GB each. These archival units (or AUs) facilitate the voting, polling, and repair inherent in LOCKSS. The MDL agreed to divide its trial collection into 17 AUs. Since the LOCKSS methodology requires that archival units be accessible to a Web crawler (Heritrix), MDL project staff also had to create a Website where these AUs would be stored and accessible. Scripts were written to divide the files into archival units of close to 20GB apiece and to copy these into a directory accessible to the Web.

MetaArchive also requires that LOCKSS manifests be present for every archival unit. These manifests are essentially HTML files that contain a list of each item in the AU along with header text that assures the LOCKSS crawler that the owner of the data intends it be harvested for preservation. The scripts that created the AUs also generated these manifest files.

To instruct the LOCKSS engine in how to traverse the source Website, MetaArchive also requires that each participant create at least one LOCKSS "plugin" describing the structure of their Website. These plugins are XML descriptions of the site structure, and they can get quite complex when addressing Websites designed for normal human use, such as a journal's Website. Since MDL project staff had created a Website specifically for harvest by MetaArchive, the MDL LOCKSS plugin could be quite simple. All staff told LOCKSS to do was dive one layer deep into the MDL site and harvest everything it ran across. This plugin was built and tested using LOCKSS tools, then uploaded to MetaArchive's source code repository and associated with an MDL institutional record in MetaArchive's Conspectus of collections.

Given the way the MDL structured this trial, most of the programming was actually focused on moving the files into a carefully structured Website that MetaArchive could harvest. The creation of

the LOCKSS plugin only required a few hours once the Website was in place. This has huge implications for the upkeep of an ongoing MetaArchive-based preservation archive: all inclusion, deletion, and changes to an archive happens through the maintenance of the source Website. This means that management of these processes lies in MDL hands. For this trial, project staff did not have to create tools or procedures to maintain this source Website, but any long-term preservation effort would require these management tools. Just as the simple construction of the source site required significant attention, the maintenance of such a site would be a significant undertaking. The MDL could choose to only provide new archival units via the source Website, asking MetaArchive to configure MDL archival units so they would not be repetitively crawled. If MDL did this, the LOCKSS voting and polling would continue even though the source data disappeared from the Web. Whether or not the MDL chose to maintain the source masters Website after the initial harvest of each archival unit by MetaArchive, creating this source site would be an additional burden on MDL.

Once all the tools were in place, launching the actual ingest process consisted of logging into one of MetaArchive's test LOCKSS caches and telling it to use the MDL plugin to start a harvest. The tools and reports on the test cache allowed for close oversight of the actual process. All 335GB were transferred in about 27 hours, largely due to the fact that both the source site at the University of Minnesota and the test cache at Auburn University have access to Internet2 connections. The transfer of all 6,000+ images went off without a hitch.

After the initial transfer was complete, the test cache contained a duplicate of the MDL source site. Files could be inspected one by one, or a Web browser could be configured with proxy settings that allowed the cached version of the site to be traversed just as the original could be. These techniques could easily be leveraged to provide either single image retrieval or batch retrieval of whole portions of the collection. Arrangements could also be made for batches of the content to be shipped back to MDL via hard drive, though that was not tested.

The LOCKSS cache provides many reports on the status of each item and whole collections. Although individual fixity checksums are not made available, the nature of LOCKSS polling among caches provides a very complete accounting for the fixity of the items in the cache. The system also ensures that contents will be replicated across at least six geographically dispersed caches, making for very robust continuity.

While from a technical standpoint, using MetaArchive was a pleasure, there are some issues that may give MDL pause about the service. MDL would be required to run a 16TB LOCKSS cache. MetaArchive could build the source site for the MDL at additional cost, and managing a LOCKSS cache requires very little local staff attention, so the impact of these requirements could be limited. Running a LOCKSS cache also means that MDL would have to become a member of the LOCKSS alliance in addition to becoming a member of the MetaArchive Cooperative. Furthermore, since

MDL is itself a collaborative, its MetaArchive Cooperative membership would have to be a "Collaborative Membership." The raw numbers shared by MetaArchive on their Website add up to some rather untenable membership fees. MetaArchive staff has provided verbal assurance that in fact the MetaArchive membership for MDL would be capped at $6,000 per year with another $833 per year for the collaborative membership. LOCKSS alliance membership would cost an additional $1,000 per year.

2.2. OCLC Digital Archive

OCLC has a simple form to use to set up an account for the Digital Archive. The trickiest element is that the MDL had to supply an OCLC symbol for the sponsoring institution. Since MDL currently has no OCLC symbol of its own, for this project, the University of Minnesota Libraries allowed the MDL to use its OCLC symbol. For any longer-term project the MDL may want to use the Minitex symbol or acquire an independent "billing-only" symbol for MDL. Project staff also learned that OCLC now regularly offers free 60-day trials of its Digital Archive service for potential customers. The MDL's whole 6,000+ image transfer was conducted under the auspices of such a free trial. It took about a week for the paperwork to make its way through OCLC and the Digital Archive account to be available.

Since the MDL already had the archival units created for MetaArchive, project staff decided to use those same chunks of data to conduct the trial with OCLC. The requirements for transferring material to OCLC were the simplest of any system explored. All that was needed was a manifest of the files to send and the files themselves. The OCLC manifest format is a simple text file with minimal header information and a list of the files to be included. It was simple enough to create such a manifest for each archival units, which OCLC terms "volumes." Although OCLC recommends volumes that are no bigger than 10GB, this turned out to be only a suggestion meant to keep transfer times reasonable. OCLC actually had no problem accommodating the MDL's 20GB volumes.

OCLC does allow for transfers on disk, but charges a bit extra for this service. The MDL was more interested in OCLC's network transfer offerings in any case. The documented version of this network transfer is an MS Windows-only process that uses a special application from OCLC to set up an SFTP transfer of content to OCLC. By digging through that application and its associated batch files, MDL staff were able to determine that any SFTP client ought to be able to do the job and confirmed with OCLC technical staff that SFTP from an MDL server could work. In fact, OCLC even had unreleased documentation for such a procedure that they shared. Unfortunately, OCLC did not allow key-based SFTP authentication which would make simple scripting of SFTP impossible. This is a planned enhancement, though no specific date was given for availability. MDL staff found another way to script the transfers using Unix's "expect" tool and some shell scripting to avoid manually setting up and checking on the progress of each of the 17 volumes of this transfer. This left OCLC DA passwords exposed, but otherwise worked well.

The biggest surprise of the transfer was how long it took. While transfers for both MetaArchive and Merritt took on the order of 27 hours to move 335GB, the transfer to OCLC of the same data took about 74 hours, almost three times as long. This may be because OCLC does not benefit from an

Internet2 connection, though it may also be that the “pull” transfers of MetaArchive and Merritt are faster than the “push” transfer to OCLC because the Web crawlers used to pull the data use overlapping simultaneous transfers while SFTP only transferred a single file at a time.

Upon completion of each volume’s transfer, OCLC sends an email acknowledgement to confirm that all went well. This message contains a malformed URL, that once corrected points to an ingest report verifying the size and mime type of each object ingested. Unfortunately, this report does not convey a checksum for the item at OCLC nor does any report from the DA system provide such a value, making it more difficult to confirm that the objects did, in fact, arrive without incident. OCLC did claim they are making four copies, two in Dublin and two in other widely dispersed locations.

Recovering files from OCLC’s DA one-by-one is simple enough, but no batch recovery tools are provided and batch recovery by OCLC is only provided at an extra charge. Single file recovery is done by constructing a URI that includes the OCLC symbol, collection name, and server name supplied as part of the ingest manifest along with the name of the file itself. Anyone with this information can recover the file; no authentication or other verification need be done once someone has a properly constructed URI. This wide-open recovery interface might be problematic for the MDL.

At first glance, OCLC’s costs seem steep, but when one takes into account the fact that OCLC requires no annual membership or other recurring equipment costs, the pricing looks much more attractive.

2.3. UC3/Merritt

The MDL had initially intended to run two trials, but in April realized there might be time to add one more and received encouragement from HathiTrust and the Minnesota Historical Society to look at UC3's Merritt system. After a couple of conversations with UC3 staff, it became clear this was feasible and the MDL proceeded with a trial of Merritt.

MHS had participated in a test of Merritt, but had largely used the Web interface to submit items one by one. The MDL wanted to test the batch process available to submit large numbers of items at once. Project staff were able to use the same server that was used for MetaArchive and OCLC trials without moving the files at all. In fact, preparation of the files for submission was simpler for Merritt than for either of the other services run through this trial. MDL staff wrote a script to produce the manifests Merritt needed for submission; this required some extra effort because staff wanted to take advantage of Merritt's awareness of the structure of objects. In Merritt, an object can consist of more than one file. The MDL wanted to take advantage of this to at least put each TIFF image and its associated XML metadata into the same Merritt object. But beyond that, the 6,000+ images included 30 "complex objects," or objects that were really made up of more than one image, such as a farmer's almanac with many pages, or the front and back of a postcard. In these cases the MDL wanted to test Merritt's ability to hold all these associated images in a single Merritt object. This required some greater sophistication of scripts so that, instead of producing one manifest for each file, MDL staff produced one manifest for each object including all of its associated files.

Since the manifests are plain text documents, this was not terribly difficult to achieve; the challenges were mostly of the MDL's own making, due to the record keeping done by CONTENTdm and other MDL processes. However, one oddity of UC3 also contributed to making this a touch more difficult than it had to be. UC3 seems to have a bit of "not invented here" syndrome, to the extreme that they invented a new text file format. Instead of using comma or tab separated values, both longstanding formats with many tools to facilitate their creation and review, UC3 has invented a " | " (space-pipe-space) separated format that seems inspired by ISBD. Unfortunately, this is also a pain for humans to proofread and provides little benefit over tab or comma delimited formats. At least it is plain text and easy to generate from scripts.

Merritt also produces an identifier for each object it ingests. UC3 uses its EZID generator to produce these. MDL staff could produce these ahead of time if they wanted and include them in the manifests. Since the identifiers produced by Merritt appear on the ingest report along with the local identifiers provided in the manifests, it is also easy to produce a mapping of local-to-preservation identifiers after the fact.

Although checksums are not visible on the Merrit system for each file, they are used in fixity checking and upon ingest checksums supplied in the manifests are compared to what was actually ingested to determine if the ingest was successful. UC3 explained that that material is being duplicated at a geographically separate location as well as being stored on RAID arrays.

The MDL used a testing system at Merritt that could not ingest all 6,000+ images at once, so broke the load into three parts for ingest. The first two thirds were loaded in sequence, one right after the other. The final third was held until Merritt staff had a chance to clear the staging server on their end; then it, too, was loaded. To load content into Merritt, the MDL used a Web interface to upload a manifest of manifests (staff called this a “mom”). The “mom” pointed Merritt to the individual manifests on the MDL server for each object. There was no way to monitor the progress of the load, but after each job was completed Merritt emailed a full report (in both plain text and CSV format) which included the EZID identifiers for the objects ingested into Merritt. The loads progressed fairly quickly; all 335GB took about 30 hours.

While Merritt does not provide a public search interface, managers of collections can search their collections for identifiers, titles, and a few other indexed fields. Merritt presents the objects in such a way that individual files can be downloaded, as can zipped archives of the objects as a whole. This facility makes recovery of information from Merritt quite painless.

The cost structure of the Merritt system is quite similar to that of MetaArchive, in that the storage charges are pretty much cost-recovery (about $1/GB) and a substantial membership charge is likely (we were told $10,000/year, but in fact the rates have not been set yet so this is an estimate). However, unlike MetaArchive, Merritt is a fully hosted system so there would be no need to run local hardware to participate.

2.4. HathiTrust

The full report on the MDL's work with HathiTrust during the fall of 2010 is available at http://www.mndigital.org/projects/preservation/fullReport.pdf. For purposes of this report on options, this section recounts salient points that are useful for comparison to the systems in this review project. The MDL worked with HathiTrust to push HathiTrust beyond preservation of books. MDL work with them was an early effort by HathiTrust to ingest material that was not page images structured into volumes. As such, MDL staff spent much more time developing procedures and process with HathiTrust than with any other organization.

HathiTrust proved to be challenging to MDL for two major reasons: (1) HathiTrust could not ingest MDL masters in their native format and (2) HathiTrust required very sophisticated submission packages from the MDL.

Because HathiTrust's attention is as much on access and dissemination as it is on preservation and migration of content, they have very strict criteria for the binaries they will accept and curate. In the MDL's case, HathiTrust would not accept continuous tone TIFF images, which made up the majority of MDL masters. MDL staff had to transform these into JPEG2000 images before sending them to HathiTrust. In addition, even where the format of MDL masters was acceptable, HathiTrust required XMP metadata be inserted into the binaries, which meant that every master image MDL submitted had to be altered prior to submission. None of the other solutions MDL staff investigated required any alteration of the binaries, though some did validate formats or offer the promise of migration for some formats.

The images the MDL sent to HathiTrust had to be combined with a very exact XML description of the item, its constituent parts, and its preservation history. This file included components that had to adhere to both METS and PREMIS standards as well as HathiTrust practices. Some pieces of the preservation history, in particular, had to be negotiated very carefully because HathiTrust documentation is not terribly precise in its specification. This XML file, with the binary images and any associated transcription texts, were zipped together into a submission interface package for HathiTrust.

While HathiTrust could have provided the MDL with a network file transfer option, they were much more comfortable with using hard disk transfers for the scale of our material. The MDL sent HathiTrust nearly 10 times the data sent to any other vendor, so this was not an unreasonable arrangement. As a consequence, though, the MDL cannot compare the speed or quality of network transfers of the other systems reviewed to HathiTrust. HathiTrust did tell MDL staff that material is replicated in Indiana and stored on tape as well as on disk in Michigan.

3. Comparisons

These comparisons are based on the questions MDL staff asked to create the preservation matrix. Please review the full matrix (http://www.mndigital.org/projects/preservation/matrix.html) to see individual responses for each system. These are highly subjective scores. Even the questions MDL staff asked and scales the MDL designed were subjective and very much influenced by the needs of the MDL and should not be read to indicate any kind of global degree of fitness of any particular system. The charts are based on scores in the matrix.

HathiTrust is the only archive that provides public access as part of its mission and service, the only "light" archive. As such, it made much higher metadata demands upon ingests, and performed stricter validation tests to ensure its ability to provide continuous access to material. The other services are "darker," though they all provide some degree of backend access to staff from participating organizations. In other words, while there is not public access, there is access available to select staff. The ideal position on this spectrum of light to dark archives for the MDL, which currently has its own publicly available access interface, has yet to be determined.

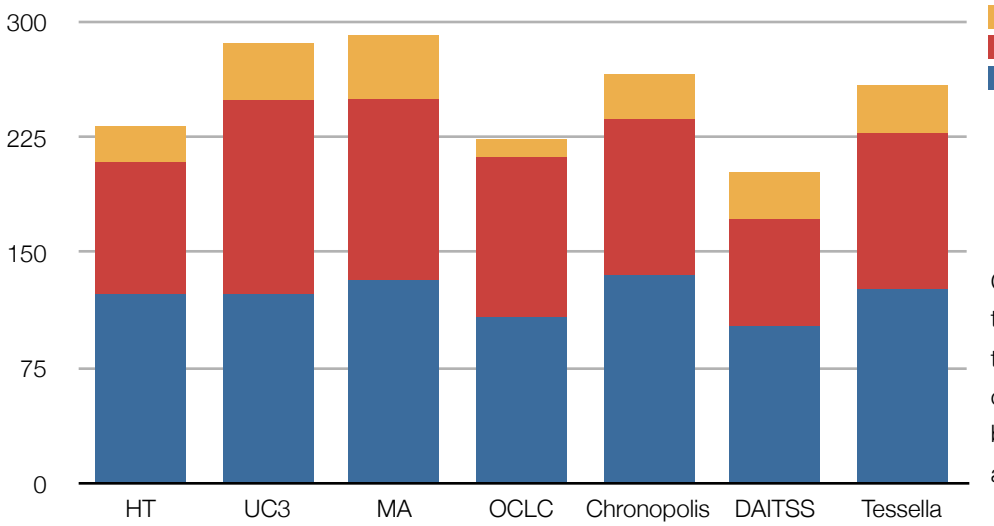

Chart 3. represents all the graded elements in the MDL preservation options matrix weighted by how important they are to MDL.

Chronopolis, DAITSS, and Tessella are excluded from much of the comparative conversation for different reasons. All three were only reviewed through conversations, but that is not the main reason that discussion of these systems will be limited. Tessella is a commercial system that, while

quite interesting, would require further exposure for a fair review. The MDL may want to follow up with Tessella further if the Minnesota Historical Society has a good experience with the company. DAITSS is not provided as a hosted solution, and is also undergoing a major transition from its first to its second generation of software; it did not seem well positioned for MDL needs. Chronopolis may be more viable for the MDL, and certainly is priced well, but again, the verbal review did not provide enough experience by which to judge them. In addition, Chronopolis, too, is going through a generational shift and a full trial would best await the launch of its revised system. While data for all three systems is provided in the matrix and charts, there will be little discussion in the comparisons.

3.1. Fitness for Purpose

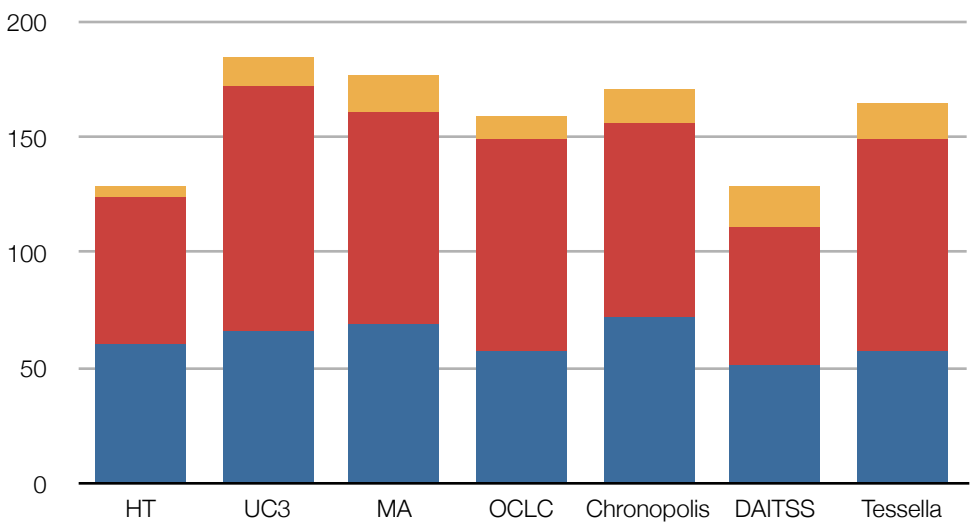

Chart 3.1. represents just the "fitness for purpose" elements in the MDL preservation options matrix weighted by how important they are to MDL.

One of the motivations for embarking on this review process was the sense that HathiTrust, while strategically well positioned, may or may not provide the optimal fit for the tasks the MDL has been considering to accomplish with a preservation archiving system. The MDL needed a comparative perspective involving other services to make an informed judgment.

The systems other than HathiTrust are mostly agnostic to the format of material they allow to be put in their archives; they can all accommodate the MDL's native masters. HathiTrust, however, makes promises of migration and validation that other services cannot provide; this is partly based on HathiTrust's stricter ingest requirements. Also, unlike HathiTrust, all of the other systems are essentially "dark" archives that do not allow public interaction with the material submitted.

All of the systems allow for essentially any metadata to be stored as files of data in their own right. MetaArchive and OCLC do not look inside the metadata files at all. UC3/Merritt allows for metadata to be associated with what it represents in a way that carries less overhead than HathiTrust.

Documentation and customer service vary a great deal. All three of the trial systems had adequate documentation to conduct the trials, with OCLC and MetaArchive providing the most helpful guidance. The task of preparing material for submission was simplest, with UC3/Merritt and OCLC vastly easier than working through the HathiTrust process.

MetaArchive deserves a special mention of its own, since the task of "preparing a SIP" was really one of "creating a Website." Since LOCKSS, the heart of MetaArchive, works by harvesting a

Website, the task of preparing material really was building this site to be the target of the harvest. Given the way LOCKSS works, this site would not necessarily need to be maintained for the preservation task to properly continue, but new archival units would still need to be presented through a source Website. This would require a greater staff commitment than the other systems.

While OCLC and UC3/Merritt seem geared more to the needs of users who would upload individual items, both are quite easy to use as batch solutions as well. UC3/Merritt probably deserves the honor of being cited as the simplest system to work with overall. While it does not provide inventory reports, the reports it provides during and after an ingest load were the most informative. HathiTrust deserves recognition for the format validation it enforces upon ingest; this, however, leads to some difficulty successfully importing masters that for one reason or another failed those tests.

3.2. Access and Ownership

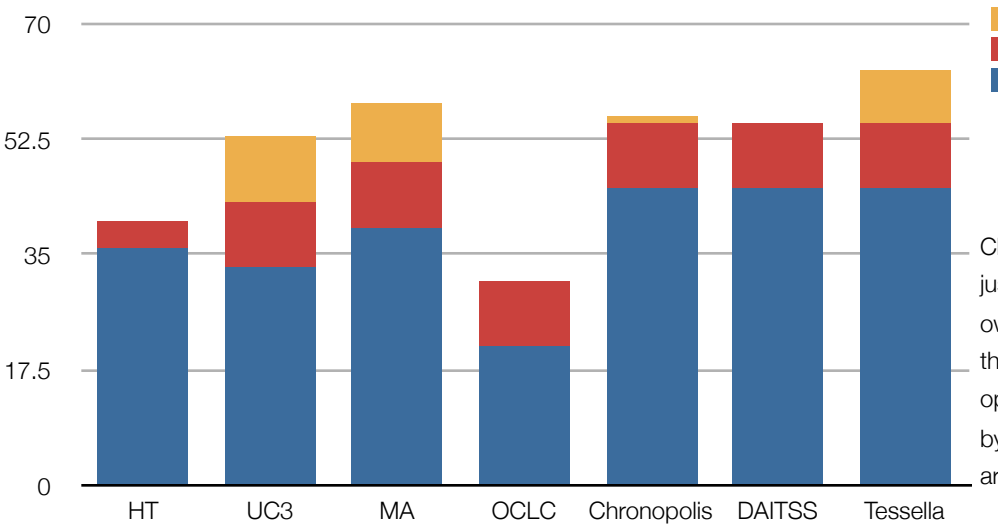

Chart 3.2. represents just the "access and ownership" elements in the MDL preservation options matrix weighted by how important they are to MDL.

Difficulty with HathiTrust arose from the fact that it is a "light" archive, with full public access to the material housed there. The preservation matrix does not rate dark over light, but all the other archives the MDL investigated are dark archives, without any intentional public access.

Only OCLC exposes the actual master images. The OCLC Digital Archive will serve up the master to anyone who has the right URI. Granted, this URI is rather obscure and requires some knowledge of both the Digital Archive and the methodology of the MDL, but such exposure is not trivial. Accidentally or through malice, URIs could be shared and masters retrieved without the knowledge of the MDL. OCLC is aware of this issue and has an enhancement request for the Digital Archive to fix this, but this enhancement has not been scheduled for development. OCLC expects to implement a more comprehensive "universal login" for OCLC services that will resolve this, but it does not sound like this will happen soon.

The fact that most of these services provide dark archives obviates the question of rights to some extent. Still, the service level agreements required for some of these services may require an expression of the rights the MDL has to the content being preserved. Before entering into a long-term relationship, these agreements should be reviewed very closely for terms that might compromise the MDL legally.

Removal of objects is more or less possible in each system. UC3/Merritt and OCLC/DA both require staff intervention. Using MetaArchive, one can tell LOCKSS to "ignore" old material, but it will not actually be wiped from the caches without active staff intervention at each cache. True deletion is not encouraged.

The UC3/Merritt system does an outstanding job of managing versions of objects that change. However, changes require the re-ingestion of a whole object (all its pages or associated files). Versions are tracked by LOCKSS for MetaArchive, and, after some additional development by MetaArchive, old versions will even be voted, polled, and repaired as are the current versions. OCLC/DA does not have any concept of versions. Each “volume” uploaded is its own universe without relationships to past volumes. HathiTrust also does not have explicit provisions for versioning, though some staff intervention is possible cases where replacing a given object justify the special treatment.

3.3. Strategy and Positioning

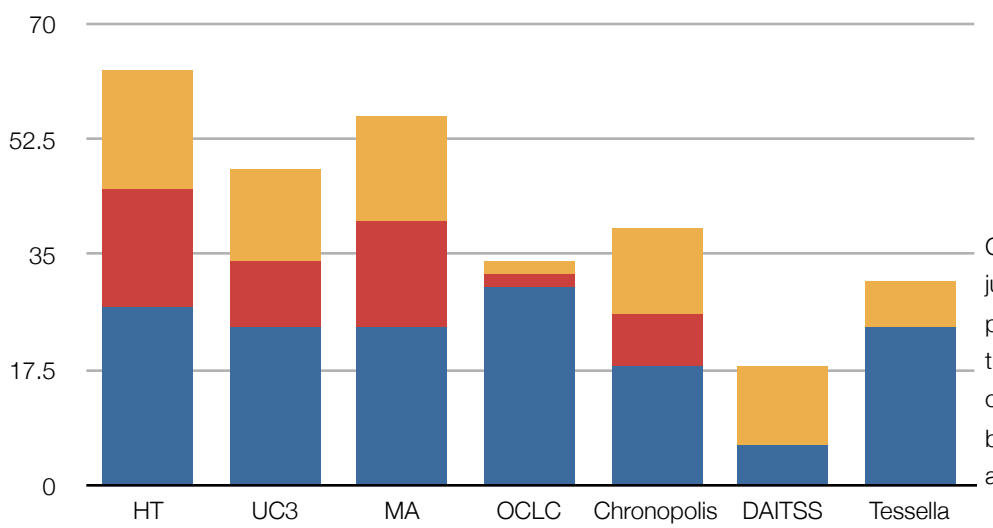

Chart 3.3. represents just the "strategy and positioning" elements in the MDL preservation options matrix weighted by how important they are to MDL.

HathiTrust shines as a strategic option, and this is probably why the MDL conducted its initial prototype project with them. Without knowing any of the technical issues, HathiTrust is an exemplary partnership, with great viability and public profile. OCLC and Tessella present themselves as commercial or commodity-based services, and as such, not offer a sense of partnership or strong lessons for the MDL to learn. DAITSS fails to meet the MDL's desire for a hosted solution and from being a very limited community at this time.

On the other hand, the commercial systems have clear business plans and seem likely to be stable ventures into the future, with perhaps a touch more confidence in OCLC than Tessella. Although HathiTrust depends a great deal upon the University of Michigan, Indiana, CIC, and other sustaining partners, those commitments seem quite robust and dependable. Merritt also depends heavily on state funding, but it is seeking ways to reduce the University of California share as it expands its customer base. MetaArchive is completing a transition from sponsored funding to membership funding and anticipates being able to continue operations into the foreseeable future at current and anticipated membership levels. Chronopolis seems the least clear about its future funding, though it has commitments through at least the next two years. We did not closely explore DAITSS funding.

Both UC3/Merritt and MetaArchive look very good from a strategy and positioning perspective. Like HathiTrust, they are building communities of users and inspiring some confidence that they will be around for a while. Both are reasonably innovative and using either would teach the MDL some valuable lessons. Merritt feels a bit more innovative and MetaArchive a bit more collaborative. Even

Tessella recognizes the value of a strong user community and is building a group on LinkedIn that shares information such as workflows, code, and reports.

HathiTrust is the only TRAC (Trusted Repositories Audit & Certification) certified option the MDL looked at. Chronopolis expects soon to achieve TRAC certification, perhaps in the summer of 2011. Merritt expects to conduct a self-audit, and MetaArchive has a self-audit publicly available (http://www.metaarchive.org/sites/default/files/MetaArchive_TRAC_Checklist.pdf). MetaArchive staff is also working to change the perception of multi-site archives like LOCKSS so that they can potentially qualify for TRAC certification. Since TRAC essentially assumes the whole OAIS model is being implemented by the provider, it may have an inappropriate scope for MetaArchive, which really only provides the "storage" box in the OAIS model. This is true of Merritt and OCLC as well, and OCLC acknowledges as much.

3.4. Economics

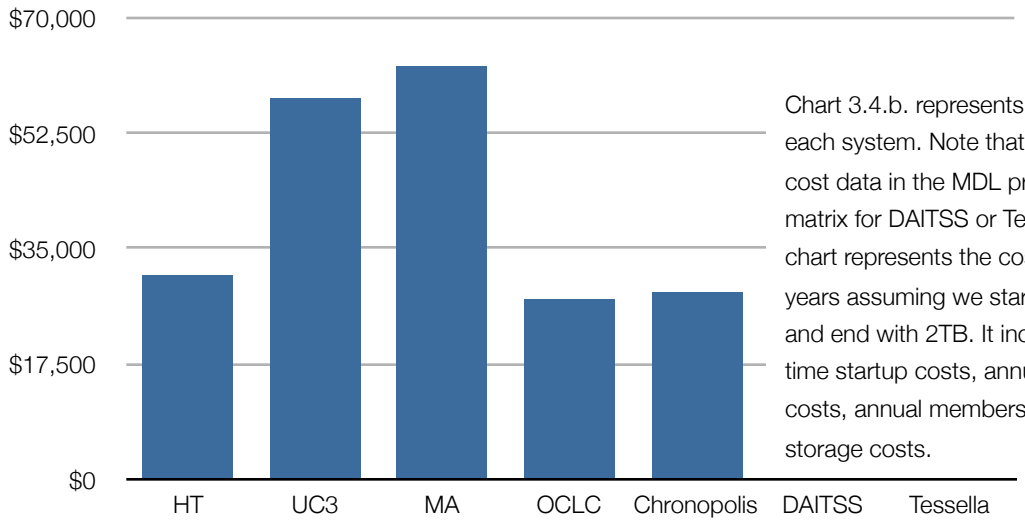

Chart 3.4.b. represents the cost of each system. Note that there is no cost data in the MDL preservation matrix for DAITSS or Tessella. This chart represents the cost over five years assuming we start with 1TB and end with 2TB. It includes one time startup costs, annual hardware costs, annual membership costs, and storage costs.

All the options have storage costs, of course, and those vary a great deal. However, the bigger differences can be found in the variety of memberships and recurring costs that are built into these choices. DAITSS2 has not been released, so no cost estimate could be made for running it. Tessella was not asked to make an estimate for MDL, so their costs are also not reflected.

The storage costs at HathiTrust end up being the highest, with UC3/Merritt and MetaArchive being the most affordable per TB. Those each appear about one-quarter the cost of HathiTrust and OCLC and one-half of the storage cost of Chronopolis. OCLC costs, however, drop as the amount of storage increases, coming to rest closer to Chronopolis pricing, although OCLC does have a hidden charge for processing a hard disk instead of online ingest. The higher overall costs of HathiTrust might be attributable to their additional public access and migration services which go beyond preservation-oriented storage services.

The most common additional recurring cost is a membership fee. Even HathiTrust has a fee of this kind, but MDL may not have to pay this fee since the MDL is administratively located at the University of Minnesota Twin Cities Libraries through its Minitex program and may be included in HathiTrust on the University of Minnesota partnership. UC3/Merritt has not yet established a fee, but they will probably have one similar to their current $10,000/year Web archiving fee for non-UC participants. Chronopolis also does not currently demand a fee, but they anticipate that a $1,500/year membership will be added soon.

MetaArchive is a special case. They not only have their own fee structure, but they require that members also join the LOCKSS Alliance. The LOCKSS Alliance does not have a fee for a

collaborative like the MDL, but MetaArchive has told MDL staff they would charge the MDL $1,000. Furthermore, though the MetaArchive Website lists a collaborative membership fee of $2,500 plus $100 per collaborative member, MetaArchvies has again told MDL staff the maximum fee for the MDL would be $6,833 per year. This is important, since the MDL currently has roughly 175 participating institutions which would result, per the Website, in a $20,000 annual cost just to join MetaArchive if the much lower maximum rate was not provided.

However, MetaArchive does have other one-time and recurring costs that are not present with the other hosted systems. To use MetaArchive, the MDL has to create a source Website of masters for LOCKSS to crawl as it acquires each archival unit. The MDL would also have to run a LOCKSS cache as part of its membership obligation. These costs are significant in their own right, adding an estimated $4,500 startup cost and $1,500 annual cost to the MetaArchive price tag. OCLC and Chronopolis also have one time setup fees, but they are much smaller.

To come up with a five-year cost for each system, the MDL consultant computed costs for a scenario where MDL starts with a 1TB storage requirement and adds to the preservation archive 0.25TB per year for the next four years. Given that scenario, MetaArchive presents the highest cost at $62,655 (due primarily to the combination of membership and hardware costs). OCLC actually presents the lowest costs ($27,250), followed by Chronopolis ($28,400), HathiTrust ($30,880), and UC3/Merritt ($57,800). Clearly, the relatively high annual fees overwhelm any storage savings MetaArchive and UC3/Merritt appear to offer at this scale.

However, the MDL's collections are actually relatively small and all the systems cited here could in fact handle quite a bit more data. If the MDL were to preserve 10 or 20 TB instead of 2 TB, the numbers would look substantially different as storage costs made up for membership fees and other flat charges. Just to keep this in mind, the chart below shows the annual cost of maintaining 2TB, 10TB, and 20TB of data in each system given current pricing:

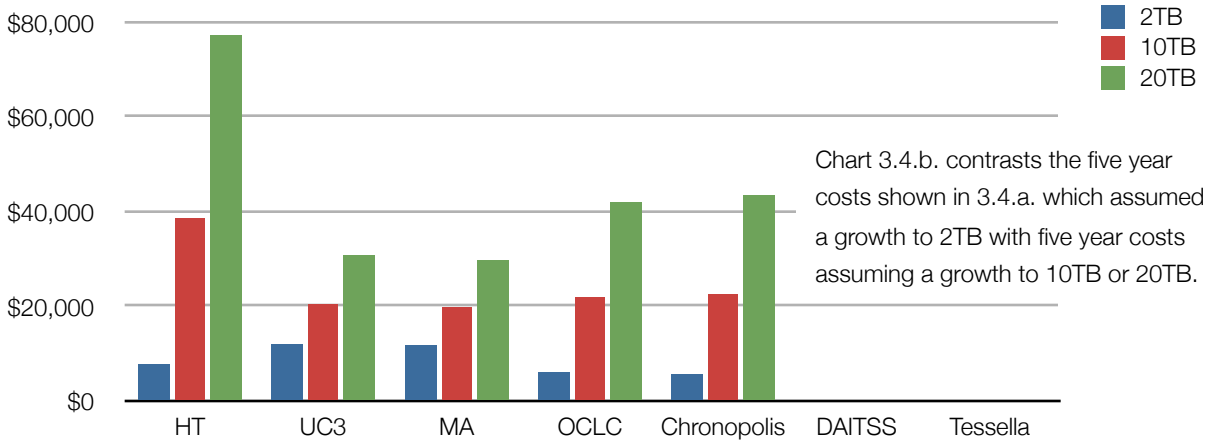

Chart 3.4.b. contrasts the five year costs shown in 3.4.a. which assumed a growth to 2TB with five year costs assuming a growth to 10TB or 20TB.

The most expensive systems at 2TB become the least expensive systems at 10TB and beyond. It is important to note that the MDL could easily pull together 5TB of data today, and 10 or more terabytes does not look at all excessive amount of data for planning purposes.

Some effort was also made to translate the experience of working with these systems into an estimate of staff expense required to manage the MDL's service based on each of these platforms. These estimates cannot be directly included in the analysis above since they are measured by "collection" instead of by terabyte. The staff effort is really encountered in setting up the foundation upon which the preservation service rests and then massaging each new collection so that it can take advantage of that foundation. The very rough estimate provided in the preservation matrix combines setup and system maintenance costs with the costs of preparing five collections.

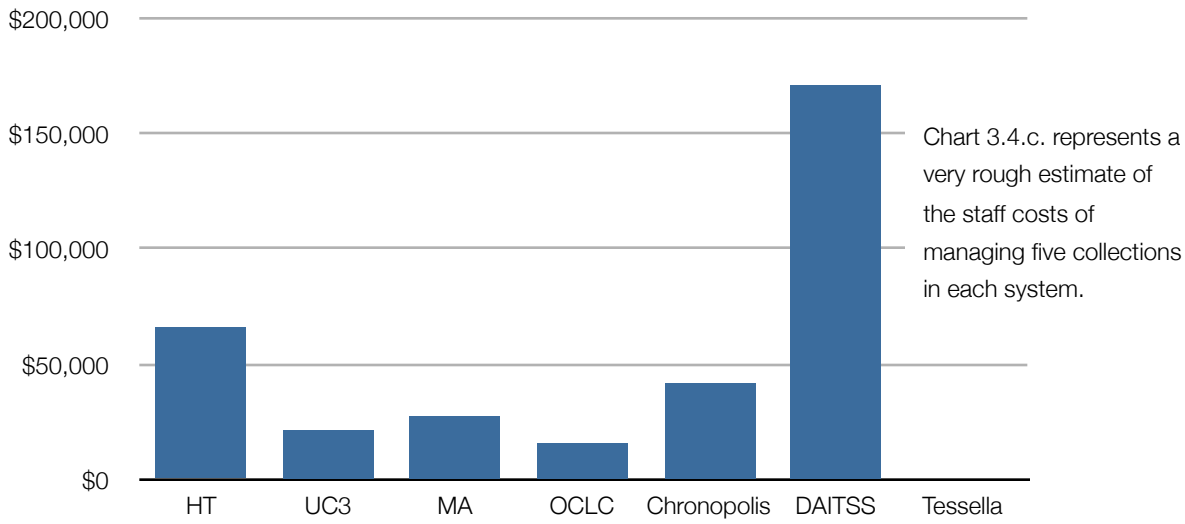

Chart 3.4.c. represents a very rough estimate of the staff costs of managing five collections in each system.

DAITSS, of course, turns out to be very expensive because it must be installed and maintained locally. HathiTrust is the next most expensive system, mostly because their metadata requirements are so much stricter. UC3/Merritt and MetaArchive each come in at about one third the staff expense of HathiTrust, and OCLC appears to have the lowest staff costs of the bunch since there is really so little prep needed to work with the OCLC Digital Archive.

4. Conclusions

The MDL began with an admittedly subjective device in developing the project preservation matrix. This report and the matrix are intended only to drive a conversation among the planners for the MDL preservation service. The conversation may reveal dimensions to these systems that are not yet clear.

The preservation matrix, conversations, and trials do make it clear that there are preservation options available to the MDL. Still, none is a notably better strategic fit and pricing throws a big wildcard into the mix.

UC3/Merritt and MetaArchive are very strong contenders for MDL attention, if the MDL is to pursue a solution largely focused on bit preservation. Merritt was easy to work with, informative in its output, very useful in its services. The MetaArchive staff was also a joy to work with and LOCKSS proved quite capable of handling the scale of data the MDL threw its way.

Chronopolis scores well in the matrix, but this score is based on documentation from the Web and a conversation with the Chronopolis team. The trials with UC3/Merritt, OCLC, and MetaArchive were very revealing, and a similar trial with Chronopolis may help illuminate its benefits and drawbacks for MDL. However, Chronopolis is in the midst of a transition to a next generation platform, has a relatively high dependence on sponsored funding with a short time horizon, and is a bit fussier than the MDL would ideally like in its ingest requirements (particularly Bags).

DAITSS feels inappropriate for MDL requirements at present. The MDL is, at this time, not interested in setting up its own preservation repository.

OCLC's Digital Archive was simple enough, perhaps a bit too simple. Even some fundamental information, such as checksums, was not accessible. While its pricing looks attractive for modest collections, it begins to get expensive for 10TB and above.

Tessella looks appealing, although its pricing is still completely unknown. However, the strength of Tessella, an interface that allows many organizations to collaborate on building a single preservation archive, ends up being unnecessary overhead for a project that will be as centrally run as the MDL preservation archive.

HathiTrust, the partner with whom the MDL built the prototype that preceded this preservation options project, still offers many unique features. It is the only light archive among the services studied, with greater emphasis on strictly managing ingest rules so that it can ensure a comprehensive validation and migration strategy in addition to safeguarding MDL bits.

Additionally, the MDL has heard whisperings of an offering from the Internet Archive that may be worth adding to this analysis and the MDL will likely investigate that possibility.

5. Preservation Options Matrix

The chart below is an abbreviated version of the preservation matrix prepared for this report. Please review the full matrix at Google Docs http://www.mndigital.org/projects/preservation/matrix.html. The version below shows only attributes that were "weighted" by the sponsors, which are also the only ones to appear in the charts earlier in this report. It also does not include the "economics" section of the matrix.

The weights represent how critical an attribute is to the MDL: items weighted 3 are vital to an MDL preservation archive, 2 is important, and 1 would be nice. The scores are on a scale of 0 to 5 with 5 being best for the MDL. Notes on the weights and the precise use of the scale for each attribute, as well as notes for specific responses for some of the questions, can be found in the full preservation matrix document at Google Docs. This data has almost no bearing outside the context of the MDL.

Fitness for Purpose	Wt	HT	UC3	MA	OCLC	Chron.	D	T
		60	92	85	75	81	65	85
Allows MDL to archive native formats.	2	0	5	5	5	5	5	5
Allows MDL to limit rights to use certain material.	3	3	5	5	5	5	5	5
Reliable bit preservation	3	5	5	5	5	5	3	5
Allows MDL to archive metadata in native format	1	0	5	5	5	4	4	4
Reliable metadata preservation	2	5	5	5	5	5	3	5
Metadata can be associated with object it describes	2	5	5	2	2	5	5	5
Internal monitoring present and accountable	3	4	4	4	2	5	5	4
Access to code driving the repository	1	4	4	5	0	4	5	3
Clear documentation	2	3	3	4	4	2	2	3
Customer service plan	2	3	3	4	5	4	0	5
Requires local copy of full archive	2	5	5	3	5	5	0	5
Level of effort in preparation of material or SIPs	2	0	5	3	5	3	2	3

Complexity of recovering from errors in the preparation of material or SIPs	2	0	5	4	5	3	2	3
Complexity of content retrieval	2	0	5	4	4	4	5	5
Readiness to accommodate MDL needs	2	4	4	5	3	3	2	3
Level of MDL staff required to manage process	1	1	4	3	2	3	5	3
Number of copies maintained in archive	3	4	3	5	4	5	0	0
Records management services available	1	0	0	0		0	0	1
How long is our contract, how predictable is the cost?	2	5	5	3	0	0	0	0
Batch upload available	3	4	5	4	3	4	4	5
Batch retrieval available	2	2	3	4	3	3	4	4
Inventory reports available (things like number of collections, number of objects, etc.)	1	0	0	3	3	4	4	5
Access and Ownership	Wt	HT	UC3	MA	OCLC	Chron.	D	T
		17	27	30	16	26	20	32
Access to masters restricted	3	4	5	5	2	5	5	5
Access to high resolution derivatives	3	5	5	5	1	5	5	5
Rights retention of sourcing institution	2	2	5	5	5	5	5	5
Material can be deleted from archive	3	3	1	3	4	5	5	5
Material can be updated and versioned	1	0	5	4	0	1		3
Ability to store new copies alongside old copies	1	0	5	5	0	0		5
Strategy and Positioning	Wt	HT	UC3	MA	OCLC	Chron.	D	T
		41	31	36	16	27	15	18
A partnership we can be proud of	1	4	4	4	1	3	0	3
A partner with great organizational viability and stability	3	4	3	3	5	1	2	3
Public audit available to review	2	5	2	3	1	4	0	0
An innovative partner	1	5	4	3	1	4	4	3
Available as a hosted system	3	5	5	5	5	5	0	5
We learn and improve ourselves by going this way	1	5	4	4	0	3	5	1
Are we a "shareholder" by means of participation	1	4	2	5	0	3	3	0
Fosters collaboration and cooperation	2	4	3	5	0	0	0	0