MDL Digital Preservation Demonstration Project: Digital Image Preservation Needs

D R A F T : 28 July 2010

Eric Celeste 1993 Lincoln Avenue Saint Paul, MN 55105-1455 651-323-2009 efc@umn.edu http://eric.clst.org

Table of Contents

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).”
Given that the MDL has already amassed a reasonably large set of image data on behalf of partners around the state, the first step in exploring such a common infrastructure will be to define digital preservation and access requirements for image content. This document defines the needs and bounds of a demonstration project that would preserve digital images already in the MDL’s care and some beyond its current reach.

The MDL intends to develop infrastructure for a wide variety of formats, but understands the value of narrowing the initial scope of this demonstration effort.

Executive Summary

After some discussions and investigation in late 2009 and early 2010 the Minnesota Digital Library, University of Minnesota Libraries, and Minnesota Historical Society plan to move ahead with a project to demonstrate the collaborative preservation of digital material. We have an opportunity to use HathiTrust platform to accomplish this. This demonstration project will be limited to 100,000 digital images from a variety of collections, including simple photographs, letters, journals, and possibly even a newspaper.

The MDL will develop the METS (Metadata Encoding and Transmission Standard) wrappers, identifiers, and workflow that will be used to package and ship image data to HathiTrust. Since HathiTrust does not currently accept the kind of content the MDL intends to ship, this will require a degree of leadership and effort on its part that will distinguish Minnesota in the digital preservation and HathiTrust communities.

To facilitate participation, this demonstration project will have to develop policies and promises around preservation and access that can be used as models for the preservation of further formats in the future. Finding a balance between the HathiTrust desire for open “light” archives and the need for careful moderation or even elimination of third-party access required by some Minnesota collection will be an especially important deliverable for this demonstration project.

The demonstration project will have to address the question of governance for a collaborative of diverse institutions, with quite disparate resources and capacities. In that context, the MDL also needs to consider the practical and political implications of presuming the use of ACHF funds for sustaining the effort.

After the needs identified in this document have been validated by a large group of stakeholders, a development effort will aim to complete this demonstration project by November 2010 so that the MDL can show the value of this collaboration to the Minnesota legislature in early 2011.

The MDL proposes that this demonstration move ahead with a partnership with the HathiTrust because this provides Minnesota a chance to lead in a nationally respected venue. The HathiTrust has accumulated a vast degree of respect in a short time, but it needs strong partners in order to grow in new directions. Minnesota can be such a partner if the MDL take the quick action.

Picking a Partner

While the Minnesota Digital Library (MDL) and Minnesota Historical Society (MHS) have done some investigation of a number of preservation options, the HathiTrust has become an early favorite for this demonstration project. HathiTrust was born out of the fact that libraries needed a way to absorb, preserve, and present the vast quantity of data being born of the Google Books project. HathiTrust has quickly become a respected player in the preservation world, with principled framework that national granting agencies value and sometimes now insist be built into applications. HathiTrust intends to move beyond the "book," other areas of digital preservation and access. They are seeking partners in this move.

The University of Minnesota Libraries have been a member of HathiTrust from its inception. The University, also a member of and a service provider to of MDL, can serve as a host for Minnesota participation in HathiTrust activities. A representative from HathiTrust presented to a January 2010 meeting of stakeholders in this project and assured us that HathiTrust was, indeed, interested in implementing preservation of digital images and would welcome Minnesota's leadership in this effort.

While long-term participation in HathiTrust would not be without cost, the costs appear to be quite reasonable when compared with alternatives like OCLC or managing a local preservation infrastructure. Working with HathiTrust also allows Minnesota to become part of a national infrastructure that would be very well suited to the mission of preservation. Taking the lead on the development of procedures and workflow for digital image preservation also affords Minnesota with a chance to demonstrate our capabilities and expertise on a national stage.
While the MDL could do more legwork to make sure that HathiTrust is the absolute best fit for its needs, expending further effort investigating alternatives could cost the MDL the present opportunity to work with HathiTrust. The MDL has determined that moving ahead with this demonstration project in collaboration with HathiTrust, while it has risks, provides the best opportunity to judge whether this course would be suitable beyond this demonstration project. Thus the rest of this document assumes the MDL will proceed with HathiTrust as its partner in this demonstration project. HathiTrust, for its part, recognizes that this demonstration will require deep collaboration.

Note that HathiTrust cost models are due to change in 2012 and OCLC is already lowering prices, so further review may be warranted after the demonstration. Currently HathiTrust costs are under $4/gb/year, though the new model will take into account resources that are shared or not shared. During this initial phase the HathiTrust would incorporate the content contributed by the Minnesota preservation collaborative at no direct cost as long as our plans make sense and meet one another's needs.

Needs Assessment

After a series of interviews with leaders within MDL, UMN Libraries, MHS, and HathiTrust the following set of needs were identified. These address what kind of content the MDL would include in this demonstration project, what workflow issues have to be considered, the limits of preservation and access, and governance issues. We acknowledge that our timeline looms large, our primary objective is to accomplish this demonstration by November 2010, in time to prepare plans for follow-on funding from the Minnesota legislature. To do this, our scope must be as narrow as possible.

Content

In order to maintain a narrow scope, the demonstration project would deal with content types in ascending orders of difficulty, addressing the simple images first, and newspapers last. The total image count for this demonstration would be no more than 100,000 individual image files, and might well only succeed at ingesting under half that number. While at this stage few of these images come from projects funded by Minnesota “legacy” dollars, this is only because those projects will not be ramping up in the timeframe of this demonstration. The demonstration project will be designed in such a way that it anticipates the participation and needs of legacy projects

Demonstration content would include the roughly 50,000 MDL Reflections images, a 20,000 image subset of the MHS collection management system, and the images from one newspaper prepared by the MHS for the National Digital Newspaper Program (NDNP). These collections include a wide range of image 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, to structurally complex documents with associated text like newspapers.

[T A B L E] [File Type : Count : Avg Size : Complexity : Text]

Even though the demonstration may not get to ingest of all the formats listed above, the project will have to give some considerations to the challenges all these formats present. This is particularly important since some material may present policy questions that HathiTrust should answer sooner rather than later, but those questions may be embedded in some of the more complex types.

Workflow

Each image to be preserved needs to be identified in a clear and unique way, described to facilitate its retrieval when needed, and shipped to the preservation archive. This constitutes the workflow that images must go through as part of the preservation process.

MDL Reflections already assigns unique identifiers to each image in the collection. However, for two reasons this identifier will not be sufficient. (1) Images from MDL will only make up part of the demonstration content and (2) there is some niggling concern that MDL identifiers may not be as unique as was intended. One of the tasks for the demonstration project would be to develop an identifier scheme, most likely a “namespace” for identifiers from collections around Minnesota, that can be used to pinpoint items within the preservation archive, HathiTrust.

Of course, an identifier is often not enough information to ensure retrieval of an image. If you know the identifier, then you are in good shape. But what if you lose track of the identifier? HathiTrust actually requires descriptive metadata also be submitted with contributed content. The demonstration project will have to define the METS packages that wrap both simple images and the more complex compound objects that are present in the content. In some cases this may be relatively straightforward, for example using the NDNP standards for newspapers, but in others there may be some creativity involved, such as the metadata wrappers for journals or multi-page letters in MDL Reflections. HathiTrust expects that Minnesota handle as much of the metadata creation as possible, including not only descriptive metadata, but also administrative, technical, and structural metadata in the submitted packages.

The whole package must also be subject to validation. Validation ensures that the digital object is well-formed and the metadata properly constructed.

These packages of metadata and content need to be transferred to the preservation archive, to HathiTrust, in a reliable and expeditious way. Since many of the digital images involved are quite substantial, this will either have to be managed over very high speed internet connections or by shipping hard disks back and forth. Either option will require logistics and planning, not to mention proper verification.

Consistently assigning these identifiers, wrapping content in appropriate metadata packages, validating, and shipping the results to the preservation archive will demand a degree of automation that does not yet exist within either the MDL or MHS operations. The demonstration project will have to assemble the toolkit and develop the code that provides these consistent results.

The demonstration must also address the concerns and practices of the HathiTrust. It may demonstrate technologically feasibility, while yet proving the effort unattractive to or unsustainable by the HathiTrust.

Preservation

At heart, the preservation archive must provide at least a few core services to be of value to Minnesota. It must preserve a bit-accurate binary of the contributed item and monitor that contribution for any corruption that can result from the nature of the storage medium or more intentional attacks or vandalism. However, the MDL seeks to store the logical data more than the structural information that might serve future accessibility. The structural content will be the minimum that serves for preservation.

As formats mature and fluctuate there may also arise the need to migrate objects of one setting format to another rising one. Since the master images stored in the preservation archive will be critical to the smooth operations local systems, the MDL will need some assurance that such migrations won't be carried out without its explicit approval or at least participation in the decision making process.

It must be understood by parties in Minnesota that the preservation archive being demonstrated in this project cannot be viewed as the only store of master images for any given collection. It is likely that some local storage of master images will continue, at least for images that don't yet have the requisite descriptive or structural metadata to build the submission package for the preservation archive. The preservation archive will also likely not provide the same speedy access to masters that can be had from local storage, so certain local processes may still require local masters.

Access

The point of preservation is access. It makes no sense to preserve something unless someone, somewhere, sometime will have access to it. With physical items, preservation often means limiting access and exposure today to ensure access by a small cadre of experts in the future. Digital preservation can often provide much greater access today, because this level of immediate access does not diminish the preservative quality of digital storage. In fact, the act of accessing the material in the preservation archive helps ensure that the material is still intact and useable. As a result, HathiTrust, for example, insists that material in its archive be accessible to the broadest audience allowable by law. This pragmatic approach has, to date, ruled out "closed" collections which institutions want limit based on criteria other than law.

In some cases, in-copyright books where the copyright holder has not provided requisite permissions, for example, HathiTrust does have to limit access to material in the archive. In this case, the copyright of most of the images MDL would preserve in this demonstration project are not held by those institutions that hold the physical objects, so asserting a copyright claim for inhibiting access would be questionable. Still, some of the member MDL institutions are very worried about the digital versions of objects in their collections escaping their hold and becoming available for free at high resolutions from internet sites other then the local historical society or MDL

Reflections. This demonstration project will have to call the question with all participating institutions: will they be willing to allow HathiTrust to present a copy of their digital images (perhaps with some sort of watermarking present) as part of the preservation process? Too much interest in opting out of this access arrangement might make these collections inappropriate for preservation at HathiTrust or any other preservation archive that insists “bright archives” are necessary to successful preservation.

Given the lack of actual copyright control over these images, the MDL also has to be prepared for objections from copyright holders should they determine that what MDL has done with these images violates their expectations. The preservation archive will have to be capable of restricting access to individual images which have drawn such an objection. The MDL will have to come to terms with the risk tolerance of HathiTrust as well.

Some digital images may well have much stricter regulations than copyright applicable. For example, the birth and death records managed by MHS must be treated in accordance with state statute which forbids certain access outside of state systems. Such requirements may demand that parts of the preservation archive be only “dimly lit” but also easily meet the “lawful access” standard of HathiTrust.

While access to relatively low-resolution derivative images may be necessary to a successful bright archive, the access to the master digital images MDL supplies to the preservation archive must be very carefully controlled. The application programming interfaces (APIs) for access to such masters may require that local systems retrieve them only by their identifiers, but those identifiers will likely be easy enough to recover from the preservation archive or assume from the local identifiers of participating institutions. These APIs must provide measures to prevent unauthorized delivery of the submitted masters to any third parties, while still allowing for efficient access to the masters from authorized users. Such authorization on the MDL side will be to a very limited set of users who will mediate access for other participants.

Note that the preservation mission of this project does not require sophisticated presentation of complex material, such as newspapers. These could be “dumbed down” to a simple “page-stream” for access and still fulfill the preservation mission at hand. Simplifying the challenge of complex formats in this way might help enormously in meeting the timeline for this demonstration. This project would emphasize the round trip objects make to the repository and back out in times of need, with the potential for greater access down the road.

Governance

As noted at various points in this document, Minnesota will have to make various policy decisions (can participants opt-out of public access via the preservation archive?) and develop solutions to many workflow issues (what will be included in the METS package for handwritten journals?). While

there is a great deal of trust among the partners here in Minnesota, there is also a need for a clear governance model so that all the participants understand their role and the ways they can apply leverage when decisions have to be made.

The demonstration project will have to address the question of governance for a collaborative of diverse institutions, with quite disparate resources and capacities. In that context, we need also to consider the practical and political implications of presuming the use of ACHF funds for sustaining the effort. The current MDL "management team" model is widely viewed as insufficient for a long term preservation collaborative. A model that clearly defines voting power and the bounds of decision making authority is required.

Even so, it is important to note that some very powerful organizations in the state, such as Minitex itself, are not separate full-standing 501(c)(3) organizations. This collaborative likely does not need such incorporation either, but it does require trust of its members and accountability to the state.

Though the MDL expects that the costs of this demonstration project will be costs of participation and not direct charge backs from HathiTrust, it does anticipate that there will be real direct costs once we get beyond the demonstration phase. The governance structure must facilitate the negotiation and distribution of these charges should the legislative initiative not cover these costs in full. It will likely have to do so through the funnel of the University of Minnesota since UMN occupies Minnesota's only seat on the HathiTrust board.

The demonstration project must also address ownership and use rights questions. HathiTrust does not allow material to be pulled from its archive once it is submitted. Furthermore, HathiTrust would like to be in a position to provide primary access to its holdings should the local source of access (such as MDL Reflections) be discontinued. What promises does the MDL make as part of building this service within Minnesota? These policies raise questions that require a common response from the MDL and other participants in the preservation collaborative. Clear governance is required for such a response.

Next Steps

This draft will be reviewed by a large group of stakeholders in July 2010. The revisions suggested will be incorporated into a final version of this document which will serve as a foundation for the next stage of the demonstration project. The MDL will then have to put names and target dates next to tasks.

The MDL will engage a preservation expert in August and September, to help us monitor this work and develop some evaluation criteria. This person may also be designated as a liaison to HathiTrust from this demonstration project. Working with HathiTrust will require establishing clear milestones, synchronizing our plans, and being prepared to put checks on our ambitions if they are outstripping either partner's ability to perform.

The MDL will then contract with a developer to work with HathiTrust staff to design the automated procedures that facilitate the demonstration project. The goal would be to have the demonstration completed by the end of November 2010. The MDL will seek to keep this development team very small given the short timeframe and the importance of being as efficient as possible.

Using what the MDL learns from the demonstration it will develop a plan for a full-scale implementation of the preservation collaborative using ACHF funds in FY12-13. This plan should include some use cases derived from the demonstration describing the effect of preservation or its lack. This plan is to be presented to the Minnesota legislature in January-February 2011. It will be presented as a start, with a defined scope, but aiming for broader inclusion.