

**Technical White Paper for MODeL- Draft 2a**  
**September 1, 2004 w/September 8 revisions**  
**Prepared by Amy Benson, NELINET, Inc.**

A working implementation of MODeL will require a means to store, manage, and provide access to digital collections created by cultural heritage institutions in Massachusetts. The purpose of this white paper is to begin to lay the groundwork for the technical basis of a system to support the goals of MODeL. This paper will discuss five components of the infrastructure needed for MODeL: the system to support and manage the digital collections, the metadata to describe the collections, digitization specifications, intellectual property issues, and education. Each component is discussed in more detail below.

**I. TECHNICAL INFRASTRUCTURE**

To be fully inclusive of all types of digital collections in Massachusetts, the system implemented by MODeL must be flexible enough to accommodate both digital collections that are held locally at each institution and the collections of institutions that are unable to store and manage their own digital collections. One means to bring together both the locally held digital collections and the collections stored in a centrally hosted repository, is the Open Archives Initiative Protocol for Metadata Harvesting (OAI PMH).

**OAI PMH**

The Open Archives Initiative (OAI) has developed a means of gathering metadata from multiple institutions (Data Providers) into a single database, which can then be searched by end users. An OAI Data Provider is an institution that chooses to make available, or expose the metadata for their digital collections. The OAI Service Provider uses an automated protocol to harvest, or gather, the metadata from each Data Provider. The harvested metadata is stored in a central database, or harvested repository. Under the OAI model, Data Providers supply \*only\* the metadata for their collections: the digital resources themselves remain at the owning institution. After gathering the metadata from participating institutions, the Service Provider offers access to the harvested metadata through a web-based search engine, and may optionally offer additional services such as digital rights management, or e-commerce. Technical requirements for serving as a Data Provider and a Service provider are available on the web at the following location: <http://www.oaforum.org/tutorial/english/page4.htm#section5>.

**Harvested Repository**

For the many institutions in Massachusetts that already have, or will create and manage their own digital collections, OAI PMH provides a mechanism to gather the metadata from individual collections, including the Centrally Hosted Repository, into a single Harvested Repository (HR) that would serve as the core of MODeL. The Harvested Repository would function something like OCLC's WorldCat database, which holds the contributed MARC records that describe resources held at institutions all over the world. Through WorldCat the collections of hundreds of institutions can be searched at one time. Information in the resulting records helps a user to locate the resource or item and access it. The HR would do the same for MODeL. It would consist of a database of all the

metadata records contributed by participating institutions in Massachusetts. The records would describe the digital collections and provide information that would allow a user to locate and access the desired resource.

### **Updating the Repository**

The OAI Harvester queries the system of each participating institution on a regular basis to identify new, modified, or deleted entries. In this way, the Harvested Repository is kept up to date. It is the responsibility of each Data Provider to maintain their own data, adding, modifying, and deleting items as appropriate.

### **Harvested Repository Components**

Implementation of a Harvested Repository would require a number of physical components, including

- Server equipment
- Database software
- OAI PMH protocol implementation
- Internet connectivity and bandwidth
- User interface
- Personnel
- Physical space

The technology and expertise needed to implement the core of the OAI harvester for MODeL are definitely within reach. Functional OAI harvesters exist and are currently available for use via the web (see the Resources section of this paper for one example). Implementing the HR for MODeL could be viewed as a first phase, to be followed by the development of the Centrally Hosted Repository, which will be a longer process.

### **Centrally Hosted Repository**

Not every cultural heritage institution in Massachusetts will be able to host its own digital collections and provide OAI-compliant access to its metadata. Based on information gathered through focus groups and input from attendees of the MODeL conference, one component of MODeL must be a Centrally Hosted Repository for the digital collections of institutions that do not have the resources to mount digital collections themselves but do have collections that would benefit Massachusetts residents and add to the value of MODeL. To bring the metadata for the collections in the CHR together with the metadata from institutions hosting their own collections, the hosting solution must be OAI compliant. Assuring that it is OAI compliant means that metadata from the CHR could be harvested and added to the Harvested Repository discussed above, providing unified access to all collections in the MODeL project. The governance and business plan for this and other aspects of MODeL are treated in a separate white paper.

#### **CHR Components**

- Database software to input, describe, and manage the digital collections
- Digital storage for contributed files
- User interface, including search options and documentation
- Bandwidth
- A preservation and/or migration strategy/mechanism

### **Potential Solutions**

Two approaches to the CHR should be considered. The first could be termed the “do it yourself” approach. This would involve setting up a facility to house both the necessary equipment and personnel to implement a homegrown digital content management solution. The second approach could be described as “off the shelf.” MODeL could contract with a vendor, or vendors, to provide either parts of the system needed to implement the CHR, or the entire project. Currently, the Boston Public Library is using its own staff and systems to mount its growing digital collections. The Northeast Massachusetts Regional Library System (NMRLS) has chosen to outsource its digital collections and has selected OCLC’s CONTENTdm Hosted Service for the purpose. A sub-committee will work to study both the practicality and the budget implications of each approach.

### **Available Off the Shelf Products**

- Integrated Library System Modules, e.g., ENCompass from Endeavor, DigiTool from Ex Libris (installed locally)
- CONTENTdm from OCLC (hosted or installed locally)
- Luna Insight (hosted or installed locally)

## **II. METADATA**

To provide appropriate access to the digital collections of Massachusetts institutions, the collections must be described, or cataloged. The metadata needs of the HR and the CHR are different. Requirements for each system are discussed below.

### **Harvested Repository**

Dublin Core (DC) is an international metadata standard that has been used as a means of description for countless digital collections worldwide. Simple or standard Dublin Core consists of a set of 15 elements, all of which are optional and repeatable. The Colorado Digitization Program (CDP), a well-known statewide digital collection, uses modified DC as its metadata lingua franca. It is recommended that MODeL look closely at CDP’s implementation and use of Dublin Core and consider it as a starting point for the development of its own project-specific metadata scheme. (A link to CDP’s DC documentation is provided in the Resources section of this document.) A sub-committee will be formed to develop usage guidelines.

### **Tools**

To incorporate metadata from existing standalone collections, tools are needed to translate existing metadata into a format determined by MODeL stakeholders. The OAI PMH requires the use of a shared metadata schema. Dublin Core (encoded in XML) is the default metadata schema for OAI harvesting, and it is recommended that MODeL use this default schema. A sub-committee will develop a MODeL-specific implementation of the Dublin Core, including means to identify owning institutions and handle other administrative tasks.

To facilitate the gathering of metadata from disparate collections, crosswalks will be developed that will translate metadata from one standard to another. Crosswalks already exist for translations between many metadata standards, such as MARC and DC. MODeL will require the services of a programmer to modify crosswalks to go from existing standards to the MODeL DC.

### **Centrally Hosted Repository**

The CHR will most likely require a standard metadata scheme for all collections added to the repository. Depending on the system selected to manage the hosted collection, a single metadata scheme may be a requirement, as some systems can handle only one type of metadata. Other systems may be able to handle multiple metadata schemes as long as they are encoded in XML. A sub-committee composed of stakeholders would determine the best metadata scheme to use, based on the technology of the system selected. A MODeL-specific implementation of Dublin Core may meet the metadata needs of the participants.

### **Metadata Creation: Possible Scenarios**

While metadata is best created by persons with knowledge of the collections, individual institutions may not have the staff or expertise necessary to do it themselves. MODeL should consider offering a range of services so that each institution will be able to participate at level that is comfortable for them. A range of possible scenarios includes

- Develop tools such as templates that allow metadata creation and entry from remote locations
- Provide a list of MODeL-aware vendors for institutions to hire for metadata creation
- Create a central metadata creation facility where trained MODeL staff will create metadata for participants' collections.

### **Tools**

#### **Crosswalks**

To integrate the metadata from the CHR collections with the metadata from the standalone collections into the Harvested Repository, a crosswalk may need to be developed that will translate the MODeL-specific metadata standard to standard Dublin Core.

#### **Automated Data Entry**

Some metadata required for the maintenance and preservation of digital files may be able to be automatically pulled from the digital files themselves. Where possible, tools that automate the creation of metadata should be employed or developed.

#### **Templates**

Because metadata is best created by persons with knowledge of the collections, MODeL will develop tools such as templates to allow staff at participating institutions to add metadata to the CHR from remote locations. However, some institutions may not have the staff or resources to create metadata themselves. MODeL may seek vendor partnerships to take advantage of bulk pricing for members who need assistance. Vendors would be able to make use of any templates created for remote

data entry. MODeL may also consider having catalogers available to assist institutions with the metadata creation for their collections. This may entail the creation of a mobile unit that could travel, or a centralized site where digital files are received and then described using tools developed for MODeL.

### **Controlled Vocabularies**

To optimize search results for the user, it is important that the metadata descriptions for the digital collections in MODeL be consistent and thorough enough to meet a variety of user needs. One means to ensure consistency of search results is to make use of controlled vocabularies at the time of metadata creation. Controlled vocabularies such as the Thesaurus of Graphic Terms from the Library of Congress may be used and supplemented to address the needs of MODeL participants. It is recommended that MODeL define controlled vocabularies as appropriate for collection description.

### **III. DIGITIZATION SPECIFICATIONS**

Minimum digitization specifications will help to guarantee a level of quality for the resources available through MODeL. Specifications developed for MODeL could provide guidance to institutions that are having materials digitized on their own. Much work has already been done on the topic of digitization specifications for digital collections and MODeL should be able to rely heavily on published industry standards. For example, it is generally agreed that whenever possible digital master files should be created. Digital master files are scans done at the size of the original item, and at a resolution of at least 600 dpi. The digital masters serve as an archive copy of the file from which derivative, or access copies can be made. The minimum resolution of 600 dpi ensures that functionality beyond web display, such as printing and zooming can be accommodated without having to re-scan. A link to the specifications of the New Jersey Digital Highway is provided in the Resources section below as a possible model for MODeL. NMRLS has also established digitization standards for its Digital Library Initiative. Informed by existing standards and practices, a sub-committee will determine minimum digitization specifications for different formats of materials.

### **Possible Scenarios**

As with the creation of metadata above, a range of digitization options exist. Institutions could scan their own items, or they could outsource the scanning to a vendor. It is also possible that MODeL could equip, staff, and maintain its own scanning lab or mobile unit that would handle digitization for participating institutions. It should be noted, however, that few individual institutions have the resources to establish and maintain a properly equipped digitization studio that would be capable of producing consistent, high-quality digital master files. Due to the costs involved, and the need to constantly keep up with new technologies, it is unlikely that creating a MODeL scanning lab would be practical or cost efficient. Vendors, on the other hand, are in the business of purchasing and maintaining the expensive equipment needed to properly digitize original materials. Their staff are also trained and experienced in using the scanning equipment. With the quantity of digitization that MODeL will engage in, it may be possible to enter into agreements with vendors to provide discounts on their services for MODeL participants. It is therefore

recommended that, where possible, MODeL use or recommend vendors for participants' digitization needs.

### **Quality Control**

Setting minimum digitization specifications for collections in MODeL means that a quality control mechanism must be put in place to monitor compliance, both for locally hosted collections and for materials entered into the Centrally Hosted Repository. Quality control of the digitized files must be done by trained individuals.

## **IV. INTELLECTUAL PROPERTY ISSUES**

Managing any digital collection requires attention to copyright issues. There are two major components to the issue of intellectual property rights and digital collections. The first is establishing ownership of the resources. The second is protecting the resources from inappropriate access and use.

### **Ownership and Access**

In a shared collection such as MODeL, it is important to contributing institutions that ownership of their digital resources be established. Metadata about each resource must indicate the owning institution and any rights, restrictions, or privileges associated with it. Clear use policies for MODeL resources must be established and stated prominently on the web site. A sub-committee will develop an appropriate use policy statement for MODeL.

### **Security**

Protecting digital assets is a challenge. Currently, several possibilities exist to assist with the protection of digital files made available to the public. These include:

- Copyright statements
- Banding or branding
- Digital watermarks
- Restriction of access to low-resolution or low quality copies
- Access restrictions (passwords, IP recognition, etc.)

It is the case however, that none of the options mentioned above is foolproof. Every effort should be made to protect the digital assets made available through MODeL. Technology in this area is evolving and developments should be tracked and applied to MODeL as appropriate.

## **V. EDUCATION**

For many Massachusetts institutions to participate, educational programs must be made available to train staff in a number of areas including digitization, the OAI PMH, metadata creation, system tools, and the like. A subcommittee will develop a curriculum, programs, and documentation.

## **STANDARDS**

It is appropriate in a technical white paper to address the issue of standards. Use of open standards in digital collection building serves two main purposes. Adherence to standards maximizes a collection's interoperability with other applications and collections. It also helps to ensure that as technology changes, mechanisms to migrate the data to new

systems will likely be available. For example, OAI is recommended for use as a harvesting mechanism because it uses an open architecture to access metadata. Dublin Core is an international metadata standard that is used heavily in digital collections worldwide. If changes occur to the Dublin Core, or if a new standard is adopted, it will be possible to translate the DC metadata into the new standard. It is impossible to predict the direction technology will take, but, by relying on industry standards, MODeL will position itself to maximize the longevity of its collections, tools, and systems. It is recommended that relevant standards be adopted wherever possible and appropriate, and monitored for changes.

## RESOURCES

### Open Archives Initiative (OAI)

OAI main page	<a href="http://www.openarchives.org/">http://www.openarchives.org/</a>
OAI for Beginners	<a href="http://www.oaforum.org/tutorial/">http://www.oaforum.org/tutorial/</a>
OAI Service Providers	<a href="http://www.openarchives.org/service/listproviders.html">http://www.openarchives.org/service/listproviders.html</a>
OAIster (example site)	<a href="http://oaister.umdl.umich.edu/o/oaister/">http://oaister.umdl.umich.edu/o/oaister/</a>

### Dublin Core

Dublin Core Metadata Initiative	<a href="http://dublincore.org/">http://dublincore.org/</a>
Dublin Core Element Set	<a href="http://dublincore.org/documents/dces/">http://dublincore.org/documents/dces/</a>
Dublin Core metadata generator (tool)	<a href="http://www.ukoln.ac.uk/metadata/dcdot/">http://www.ukoln.ac.uk/metadata/dcdot/</a>

### Colorado Digitization Program

<http://www.cdpheritage.org/>

### Digital Library of Georgia

<http://dlg.galileo.usg.edu/>

### MIT Libraries Metadata Mappings (Crosswalks)

<http://libraries.mit.edu/guides/subjects/metadata/mappings.html>

### New Jersey Digital Highway Digitization Specifications (one example)

<http://www.njdigitalhighway.org/documents/njdh-image-specs.pdf>

## NOTABLE SITES

### Columbia University Libraries

Digital Library Projects  
<http://www.columbia.edu/cu/lweb/projects/digital/>

### Cornell University Library

Digital Initiatives at the Library  
<http://campusgw.library.cornell.edu/about/digital.html>

Moving Theory into Practice: Digital Imaging Tutorial  
<http://www.library.cornell.edu/preservation/tutorial/>

**Georgia Institute of Technology**

Digital Collections

[http://www.library.gatech.edu/search\\_locate/digital\\_collections.html](http://www.library.gatech.edu/search_locate/digital_collections.html)

**University of Kansas**

Digital Initiatives Program

<http://kudiglib.ku.edu/>

**U.S. National Archives & Records Administration (NARA)**

<http://www.archives.gov/>

**Vanderbilt University**

Television News Archive

<http://tvnews.vanderbilt.edu/>