Beyond Dublin Core: Development of the Workflow Management System and Metadata Implementation at Rutgers, The State University of New Jersey

Mary Beth Weber, Head, Cataloging and Metadata Services, Technical and Automated Services, 47 Davidson Road, Rutgers University Libraries, Piscataway, NJ 08854
email: mbfecko@rci.rutgers.edu
phone: 732-445-0500
fax: 732-445-5888

Sharon Favaro, Head, Digital Projects and Gifts, Cataloging and Metadata Services, Technical and Automated Services, 47 Davidson Road, Rutgers University Libraries, Piscataway, NJ 08854
email: sbs49@rci.rutgers.edu
phone: 732-445-7990
fax: 732-445-5888

Abstract: In 2003, the Rutgers University Libraries (RUL) received a National Leadership Grant from the Institute of Museum and Library Services (IMLS) to develop the New Jersey Digital Highway (NJDH), a statewide repository and collaborative portal. It was created with the following project partners: the American Labor Museum/Botto House, the New Jersey State Library, the New Jersey Historical Society, the New Jersey State Archives, and smaller libraries, museums, archives, historical societies, public broadcasting organizations, and public schools. The grant partners brought different types of primary source materials to the project that they wished to preserve and make available. The goal was to provide a "one stop shop" for information about New Jersey's history, culture, heritage, and immigration experience through photographs, books, documents, periodicals, three-dimensional objects, audio, and video. This paper will discuss two components of digital curation at RUL: (1) the Workflow Management System (WMS), a metadata creation and management tool developed by RUL and how it feeds into the institutional repository, and (2) RUL's metadata schema and how it enables digital curation.

1. Overview

In 2003, the Rutgers University Libraries (RUL) received a National Leadership Grant from the Institute of Museum and Library Services (IMLS) to develop the New Jersey Digital Highway (NJDH), a statewide repository and collaborative portal. It was created with the following project partners: the American Labor Museum/Botto House, the New Jersey State Library, the New Jersey Historical Society, the New Jersey State Archives, and smaller libraries, museums, archives, historical societies, public broadcasting organizations, and public schools. The grant partners brought different types of primary source materials to the project that they wished to preserve and make available. The goal was to provide a "one stop shop" for information about New Jersey's history, culture, heritage, and immigration experience through photographs, books, documents, periodicals, three-dimensional objects, audio, and video. This paper will discuss two components of digital curation at RUL: (1) the Workflow Management System (WMS), a metadata creation and management tool developed by RUL and how it feeds into the institutional repository, and (2) RUL's metadata schema and how it enables digital curation.

2. Developing a Content Management System for Metadata

2.1 Repository Development- RUL worked to develop a repository to meet the needs of the NJDH and Rutgers University, as well as future collaborations. A decision made was not to purchase a commercial product for content management. The benefits of creating our own metadata creation tool far out weighed those provided by a commercial product. The top reasons for creating an in-house repository are: free access to source codes; the ability to guide developments, enhancements and adapt to changes; and to have a cross breed of metadata schema for interoperability with other systems and presentation platforms. Given that repositories are an emerging technology, a metadata creation utility was needed to grow and change as RUL's ideas, and those of our faculty and partners, changed with respect to the role of the repository. This led to the WMS, a Rutgers developed platform independent Web accessible system used to create metadata that serves as the front end to ingest digital objects into RU**core**, the Rutgers institutional repository. The system enables creation and editing of metadata, and creates a METS-XML wrapper for digital objects and associated metadata for export and ingest. The WMS uses an underlying metadata schema that draws on MODS, METS, FOXML, and PREMIS to ensure scalability for various projects, and interoperability with other systems. The schema maps to Dublin Core, which is mandatory for participation in the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) for sharing information between repositories. It also supports multiple strategies for entering objects and creating metadata, including the ability to batch import digital files, to add an entire digital object collection through batch import, to migrate metadata from a collection owner's database into the WMS, or to create metadata for each object.

2.2 Institutional Repository Software- The WMS feeds into Fedora, making resources available in NJDH and RU**core**. A number of products were evaluated, and their capabilities were extensively reviewed.

Fedora was selected as the software for RUL's institutional repository for several key reasons:

- The goal was to build a trusted repository architecture with a strong preservation focus.
- Fedora provides native support for METS, and RUL collects preservation, rights and provenance information as a critical part of a long-term preservation strategy.

- Fedora provided a flexible architecture. At that time, it was the only service oriented repository architecture available. This permits extension of Fedora with the WMS and integration into Sakai through the capability to seamlessly layer services on top of the core architecture.
- Fedora was open to collaborative development, and Rutgers became a Fedora development site. This enables us to benefit from shared collaboration and to extend our finite programming resources.

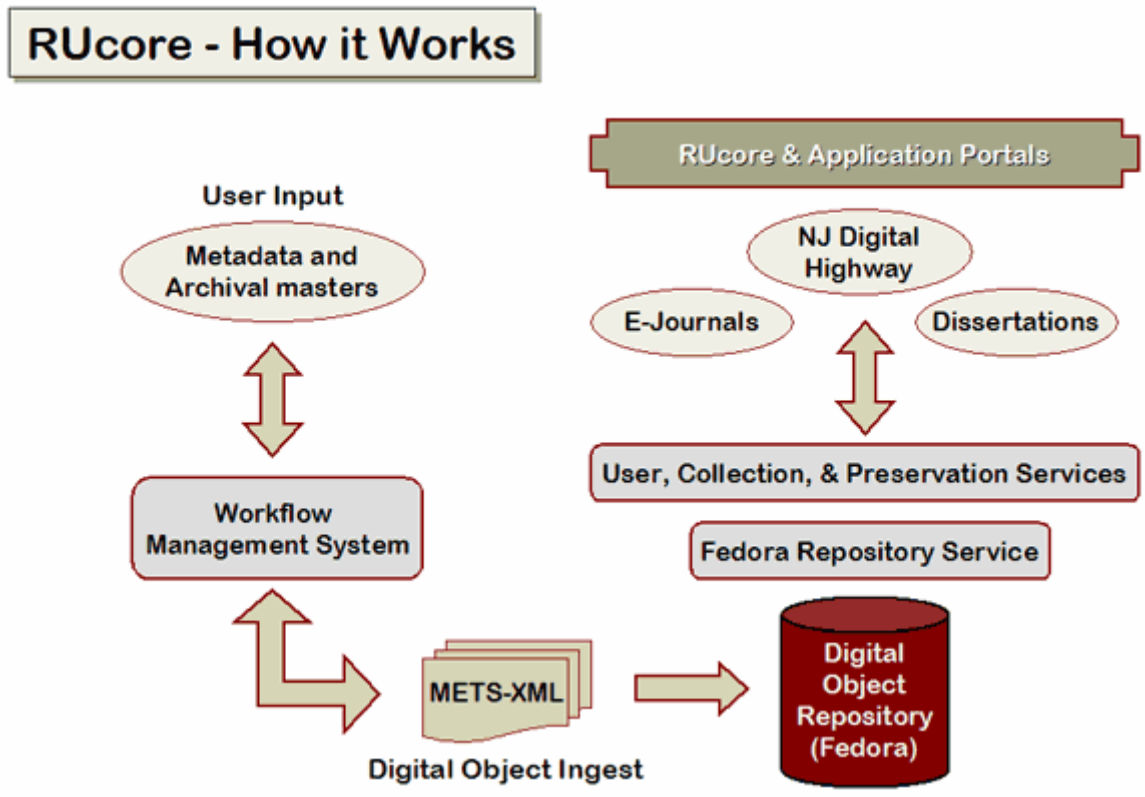

Figure 1: RUcore Operation[1]

3. WMS Hierarchy and Metadata

3.1 WMS Hierarchy- The hierarchy of information in the WMS is as follows: organization, collection, project, and user.

- Organization provides information about the institution or corporate body that owns the collection(s).
- Collection is the official name of the group of resources that have been designated by the owner, to which the resource belongs.
- Project is the work provided by an organized group of people chosen to work on a subset of a resource collection. User refers to all categories of individuals who work in the WMS. Anyone who uses the WMS must first be assigned a user account. This function is also used to assign the users to the appropriate role(s) for a given project.

3.2 WMS Metadata- RU**core** can support any schema, including MARC, for sharing records with IRIS, RUL's integrated library system (ILS) and Dublin Core, for sharing with other digital

library initiatives, such as the Networked Library of Digital Theses and Dissertations, an international open access portal to theses and dissertations (http://rucore.libraries.rutgers.edu/cyber/metadata.php).

As previously noted, the WMS uses an underlying metadata schema that draws on MODS, METS, FOXML, and PREMIS to ensure scalability for various projects, and maps to Dublin Core for interoperability with other systems through OAI-PMH. "The METS schema provides a flexible mechanism for encoding descriptive, administrative, and structural metadata for a digital library object, and for expressing the complex links between these various forms of metadata."[2] METS provides a means for exchange of digital objects between repositories, which is ideal for NJDH and RU**core**. In addition to the schema cited, the WMS also incorporates RULIB, a set of locally developed elements and vocabulary that ensures a flexible system that can handle the special needs of objects.

The WMS is composed of five types of metadata that are drawn from various schemas:

- **Descriptive metadata** provides information for users to discover and obtain access to information resources. MODS was chosen for RUL's descriptive metadata since it enables records to be mapped back to IRIS in MARC (Machine Readable Cataloging) format, and provides and retains standard bibliographic cataloging principles. "As an XML schema, the 'Metadata Object Description Schema' (MODS) is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records."[3] This capability will soon be available for Rutgers Electronic Theses and Dissertations (ETDs). It enables RUL to provide information through multiple presentation standards in a schema easily understood by each user community.
- **Source metadata** describes the provenance, condition and conservation of analog source materials, such as photographs, books, maps, audio, and video. It defines the nature of an analog source object and uses elements from the PREMIS (PREservation Metadata) schema.
- **Technical metadata** provides information about the digital master files that RU**core** will maintain for long-term preservation and access. It is required for all digital objects regardless of whether they were born digital or digitized from an analog source object. RUL uses the ANSI/NISO Standard Z39.87-2006 Data Dictionary – Technical Metadata for Digital Still Images for technical metadata. Metadata standards for sound and video are currently underway.
- **Rights metadata** identifies the rights holder(s) for each information resource and identifies the permissions for use that the rights holder has granted, including any restrictions. The rights metadata refers to digitized objects that are available via the Internet. RUL uses PREMIS for Rights metadata. Some of the challenges that libraries and archives face include integrating information collected about a resource including its provenance and condition, and copyright and permissions status. Since technology changes, access methods and technologies should be documented. Permissions and rights must also be documented to guarantee access and other types of usage. RUL has developed a simple rights schema based upon rights events that enables archivists or rights administrators to document every instance of research or permissions requests, rights transfer, copyright status, etc. The draft rights schema can be reviewed at the RU**core** website at: http://rucore.libraries.rutgers.edu/collab/ref/doc_mwg_rulib_rights_md_draft.pdf

- **DigiProv metadata**, or digital provenance metadata, provides a digital "audit trail" of any changes to the metadata. The WMS uses elements from PREMIS 1.0 for digital provenance metadata.

RU**core** uses an innovative "event-based" data model, within a METS data architecture, to capture information and ancillary documents for any facet of a resource's lifecycle, whether describing the intellectual content, the provenance, the digital lifecycle or the rights lifecycle.

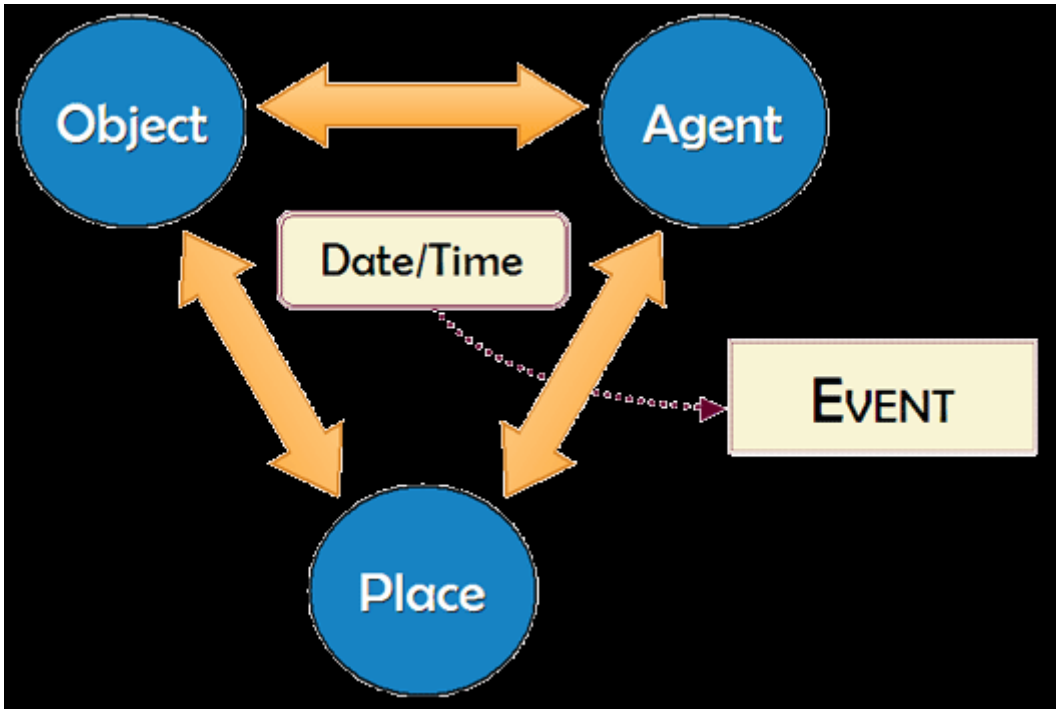

Figure 2: Rutgers Event Based Metadata Model - "Object" represents an information object, which could be a photograph, a three-dimensional piece of artwork, or a digital image. An "agent" is a person or organization responsible for creating, describing, managing, or using information about a particular object. "Place" is context independent, and may be place of creation, display, location, etc. When an object interacts with an agent or place or place at a specific point in time, an event occurs in the lifecycle of the object. Examples of events are provenance events (donation, acquisition), or preservation events (tape removal, digital restoration).[4]

4. Digital Workflow

4.1 Team Approach and Roles- Digital projects at RUL are handled using a team approach. The team consists of four roles: an overall Project Manager, a Collection Manager/Curator, a Metadata Manager, and Object Creators. In addition to the project team, an individual has been designated as the WMS Manager. This individual oversees all development of the WMS, including programming, specifications for new releases, coordination of testing for new releases, and metadata standards and applications.

The Project Manager is the primary liaison to the owner/curator of a given collection and also coordinates the digitizing of the objects.

The owner/curator of a collection is responsible for providing a description of the collection as well as descriptive information for each object in the collection (format, dates, restrictions on access, copyright information if applicable).

The Metadata Manager has primary responsibility for metadata design and creation, including working with the project manager to analyze incoming collections, managing metadata timetables and deliverables, periodically reviewing metadata with the collection owner/curator and project manager to insure that the information needs of the primary user group and preservation needs to the owner/curator are addressed, developing templates and controlled vocabularies, training the curator/owner and staff when metadata is created locally.[5] The current WMS staffing at RUL is comprised of four Catalog and Metadata Librarians who have full access to the all administrative settings and are responsible for setting up collection and project records, creating user accounts creating project specific templates, and designing project workflow.

When new collections are identified for NJDH or RU**core**, the Project Manager and Metadata Manager meet with the Collection Owner to establish workflow for scanning digital objects and metadata creation, evaluate the collection, and gather administrative information. The condition of the collection is evaluated. If a finding aid is available (digitally or in analog format), it provides valuable information about a collection, such as content, description, provenance, and supplies much of the information needed by the Metadata Manager. The Metadata Manager works with collection manager/curator to set up a finding aid, inventory or other resources that will assist with metadata creation when collections are unprocessed; restrictions on access or display of information are determined; and digitization standards are discussed and recorded.

4.2 Templates- Templates are used to streamline workflow and to automatically populate metadata records with required information specific to a given project. Templates created at the project level free the user from repeatedly supplying this information. This is particularly useful for individuals with little familiarity with the system as this limits elements and information that need to be provided.

As with any automated process, there are drawbacks to using project templates. The user/ Object Creator cannot change or modify a template or create another template. The current WMS release supports a single object architecture, and expanding this capability is under consideration for a future release.

Single templates may also be created. The benefit of using a single template is that it permits the Object Creator to edit or create new templates with specific common elements or different object architectures. The downside to this approach is that the information added by the Metadata Manager, such rights and source provenance, which are crucial project information, could be misapplied when an Object Creator creates a new template.

4.3 Batch Processing Functions-The WMS also provides the ability to add an entire digital object collection through batch import or to migrate digital files and metadata from the collection owner's database into the WMS. This feature may be used for any databases or spreadsheets that export records in text. Batch import is quite efficient for pre-existing databases, as was the case for a public library's historic photograph collection. The library created a database with core metadata elements that required some tweaking such as mapping their metadata elements

to the WMS proprietary metadata, and providing controlled vocabulary terms and authority control for names.

The batch import function was used for a project in which all the metadata was contained in an Excel spreadsheet. The Metadata Manager created a basic Excel sheet with required elements and also suggested additional elements that the project archivist was likely to use. The archivist did her work in the spreadsheet yet was able to view the WMS to view vocabulary and to add additional elements to the spreadsheet if needed.

Some of the challenges that accompany the batch import process include reconciling subtle inconsistencies in vocabulary, attention to details such as capitalization, and resolving the issue of combined architectures on one spreadsheet since the current WMS implementation can handle one object architecture per spreadsheet for batch import

4.4 WMS Output and Display- When metadata creation is complete, objects may be viewed in FOXML, METS, or HTML. The HTML view breaks the metadata down by sections and displays descriptive metadata, source metadata, technical metadata, rights metadata, a structure map, information on files associated with the object, and digiprov metadata. Each time a record is revised or changed, this is recorded in a new display of digiprov. An object may have two or more instances of digiprov metadata. After items are ingested into RU**core** or NJDH, objects are available in different presentation formats. Still images are available as DejaVu, PDF, or JPEG. Sound objects are available as MP3 files. Video resources are available as Quick Time or Flash Video. File upload assigns a persistent identifier based on organization information. RUL uses the Corporation for National Research Initiatives' Handle System to assign, manage, and resolve persistent identifiers for digital objects and other resources on the Internet.[6] Clicking on the handle, or persistent identifier, permits users to view records in MODS, Dublin Core, MARC, or a combined full record with all schemas.

5. Digital Curation at Rutgers University

5.1 Definition- Wikipedia defines "digital curation" in this way: "Digital curation encompasses all of the actions needed to maintain digitised and born-digital objects and data over their entire life-cycle and over time for current and future generations of users. Implicit in this definition are the processes of digital archiving and digital preservation but it also includes all the processes needed for good data creation and management, and the capacity to add value to data to generate new sources of information and knowledge."[7] This quote describes the role of the Metadata Managers and how they use the WMS to provide digital curation at RUL. Metadata creation adds value to materials by preserving resources, facilitates resource search and discovery through description and controlled vocabulary terms for names and subjects, and promotes use and re-use of materials that might previously have been available to a limited privileged audience due to value or condition, or might not have been well served by traditional MARC cataloging.

5.2 Metadata Manager's Role- Once a collection is received, it is assigned a Metadata Manager to review the collection with the collection curator to evaluate the finding aid, verify that each object is matched to the digital object number. Three events take place when a collection assessment is made:

- The Metadata Manager will review the WMS to make sure the metadata is suitable for the collection and evaluate if any additions or edits should be made (since this is emerging software, enhancements are made as needed).

- Combing the collection to determine controlled vocabulary and to establish authoritative forms for names (personal and corporate), and assessing the collection for common elements that can be used to form a template. The Metadata Managers set up the organization record which includes Institutional information, such as MARC organization code ID, address, and a handle ID.
- The Metadata Manager works with the Collection Curator to set up a collection record, which includes assigning a collection id, title, entering information (name, affiliation, email) Metadata Manager, Collection Owner/Curator, and Object Creators (this step is very important because when objects are ingested into Fedora, edited or deleted a notification email is sent to the Metadata Manager and Object Creator). The collection record also contains descriptive information including and may be linked to other collections through the hierarchical structure.

5.3 WMS Controlled Vocabulary Module- A recent addition to the WMS is a controlled vocabulary module that is restricted to the Metadata Managers, which enables a Metadata Manager to add vocabulary as needed without the intervention of a programmer. Previously, any changes to controlled vocabulary terms triggered a two-part process: updating a master spreadsheet and working with a programmer to accommodate changes and additions into the WMS. The controlled vocabulary module enables Metadata Managers to add vocabulary terms, such as a Library of Congress Subject Heading (LCSH) term or an Art and Architecture Thesaurus (AAT) genre term as needed. The controlled vocabulary lists consist of pull-downs of terms deemed most appropriate by the Metadata Working Group, a committee composed of the four Metadata Managers, the WMS Manager, and the Collection Project Manager. Since LCSH and AAT are quite rich, a complete pull-down list of all terms is not provided. The controlled vocabulary module is used in cases when it becomes necessary to add controlled vocabulary terms to the pull-down lists.

5.4 WMS Releases and Testing- Ongoing testing for new WMS releases occurs on a regular basis. Testing is intensive, and consists of daily three-hour sessions over a period of two or three weeks. The programmers work on enhancements between various WMS releases and have target dates for new functionalities geared to specific future releases. The Metadata Managers and Object Creators, who work in teams that handle specific categories of resources (electronic theses and dissertations, audio and video, periodicals), do all the hands-on testing. A great deal of time and effort are invested in testing to assure a robust system. Testers use an open source error reporting system (Bugzilla) to report problems to programmers, and to share information with other testers. The outstanding bugs are reviewed when testing for a release has been completed to determine what can be resolved, and what functionality should be pushed to a future release.

6. Conclusion

The metadata schema in the WMS draws from a variety of sources, making it possible to create rich descriptions for a variety of primary source materials from institutions that vary by size, type, and mission. In addition, the schema is scalable, which will permit it to evolve as the needs of RUL's repository and faculty grow and change.

References

1. Rutgers University Libraries. RU**core**- How it Works from "How Does RU**core** Work?" http://rucore.libraries.rutgers.edu/how/

2. Library of Congress. Metadata Object Description Schema (MODS) Official Web Site. http://www.loc.gov/standards/mods/

3. Library of Congress. METS: An Overview and Tutorial from the Metadata Encoding and Transmission Standard Web Site. http://www.loc.gov/standards/mets/METSOverview.v2.html

4. Rutgers University Libraries. RU**core** Metadata model from "How Does RU**core** Work?" http://rucore.libraries.rutgers.edu/how/

5. Corporation for National Research Initiatives. http://www.cnri.reston.va.us/

6. Favaro, Sharon. *Metadata Workflow for Digital Collections*. http://rucore.libraries.rutgers.edu/collab/ref/doc_mwg_metadata_workflow.pdf

7. *Wikipedia*, "Digital Curation." http://en.wikipedia.org/wiki/Digital_curation