Batch Loading in Metadata Creation: A Case Study the Rutgers University Libraries Experience

Rutgers University has made this article freely available. Please share how this access benefits you. Your story matters. [https://rucore.libraries.rutgers.edu/rutgers-lib/33944/story/]

This work is a SUBMITTED MANUSCRIPT UNDER REVIEW (SMUR)
This is the author's manuscript for a work which, at the time of deposit to RUcore, was under formal review managed by a socially recognized publishing entity. The manuscript may or may not have been subsequently published. Content and layout follow publisher's submission requirements.

Citation to this Version: Sun, L. "Batch Loading in Metadata Creation: A Case Study the Rutgers University Libraries Experience", Rutgers University Community Repository, 2011. DOI: http://dx.doi.org/doi:10.7282/T31J9830

Terms of Use: Copyright for scholarly resources published in RUcore is retained by the copyright holder. By virtue of its appearance in this open access medium, you are free to use this resource, with proper attribution, in educational and other non-commercial settings. Other uses, such as reproduction or republication, may require the permission of the copyright holder.

Article begins on next page

SOAR is a service of RUcore, the Rutgers University Community Repository. RUcore is developed and maintained by Rutgers University Libraries.
Abstract

Purpose – The purpose of this article is to describe a workflow of automated batch loading metadata from existing text to a database.

Methodology/Approach – It introduces a case for the experience of metadata creation at Rutgers University Libraries in a collaborative digital project with the Hoboken Public Library in New Jersey.

Findings – It is found that a well-designed workflow is crucial to the success of metadata batch loading. It is also found that the metadata manager needs to collaborate with people of different roles and work carefully with data reorganization and transferring.

Practical Implications – Metadata creation and management is an integrated component of any digital project. Our experience in metadata batch loading has practical significance that may be incorporated into the practice of other metadata projects. The workflow introduced in this article will provide valuable example for librarians and information professionals to consider or redesign their own digital efforts.

Originality – Based on our real exercise, this workflow has been proven to be unique and useful. It was, after the writing of this article, applied to a new collaborative digital project and once again fulfilled the requirements for another batch transferring process.

Keywords Digital library, Batch loading, Mapping, Evaluation, Editing

Paper type Case study

Introduction

Metadata creation has been extensively discussed in the literature of digitization (e.g., Hillmann and Westbrooks, 2004; Lopatin, 2006). Most studies have concentrated on introducing the process of creating metadata from the scratch for its challenges and solutions. On the other hand, the topic of utilizing existing data and transferring them into metadata elements has been relatively neglected, although like exercises may have been taken by some librarians and information professionals in practice. Its practical significance is obvious. Many collections, while being digitized and brought up online, have textual descriptions that can be potentially conveyed into metadata to describe the collection items. Even born-digital items may need reorganization or moving of their metadata for varying purposes. Reuse of obtainable textual descriptions for metadata will definitely save considerable resources in terms of the investment of time and energy of metadata managers.

The transferring of text to metadata can be facilitated by automated batch loading through computer applications. This process, however, requires tremendous human intervention to reorganize original text, map to appropriate metadata structure, monitor the loading process,
and control the quality of the transfer throughout. The procedure that a metadata manager has followed for handling a loading will determine how well the loading is taken and thus how well the quality of the generated metadata is. It is, therefore, helpful to retain an effective workflow that has been carefully designed and repeatedly tested, although workflows may vary from project to project.

Some of previous studies of metadata creation and management have talked about various situations of metadata mapping which are related to the topic of metadata batch loading. Yet, very few of them, if any, have systematically portrayed a picture of undertaking like process. This article describes the work of a metadata manager in the digital practices of the New Jersey Digital Highway project at Rutgers University Libraries (RUL). It presents the RUL’s experience in metadata creation by introducing how a workflow has been established and implemented in support of an automation process. Particularly, it introduces a collaborative project that demanded batch loading of metadata elements for a historical photographic collection into the RUL’s digital repository. The author hopes that the RUL’s experience will help people evaluate the efficiency and effectiveness of an established workflow in the automated creation of metadata and incorporate the experience into their own practices.

The New Jersey Digital Highway (NJDH)

NJDH (http://www.njdigitalhighway.org) is a digital library supported by a grant from the Institute for Museum and Library Services (IMLS) in 2003. It is a collaborative project that involves digitization efforts by many New Jersey cultural and educational institutions hosted at RUL. The principal grant partner organizations include Rutgers University Libraries, New Jersey State Library, New Jersey Division of Archives & Records Management, The New Jersey Historical Society, and American Labor Museum/Botto House National Landmark. The purpose of the project is to collect, preserve and disseminate cultural and historical heritages of the state in the form of online presentations. Through NJDH, users are able to access collections of many libraries, museums, archives, and historical societies.
The inaugural collection of NJDH was the “Changing Face of New Jersey – The Immigration Experience from Earliest Times to the Present” that has been presenting the immigration history of New Jersey. This initial collection has included audio and video oral histories, photographs, household and industrial implements, sheepskin deeds, diaries, annual reports of ethnic organizations, and the like. With an active participation of various organizations lately, the collection has been expanded; and more new collections have been added to the system. As of the summer of 2007, NJDH has been able to collect more than 10,000 digital objects.

For better operating the NJDH system, RUL has provided important technical components to fulfill the requirements for necessary hardware and software, based upon emerging national and international standards for digital artifacts of current as well as next generations. To make it a real collaboration to benefit all participants, RUL invested tremendous time and energy to improve the functions of this user-centered information repository so that a variety of digital formats can be preserved and accessed through the Website such as photographs, books, documents, periodicals, three-dimensional objects, audio and video, etc. NJDH has developed into a centralized system for better maintenance and convenient development. At the same time, this system has also individualized its customization for every participant institution so that all partners can create a dynamic online portal for their own collections. This individualization provides unique look and feel of a custom portal for each institution whereas the digital content is actually linked back to NJDH at RUL. This technical flexibility has attracted many more generous contributions of cultural agencies and the active involvements of New Jersey educators to develop lesson plans for the digital collections around the state.

RUL has also made great efforts for the optimization of its project management by developing a balanced infrastructure of the digital repository. Both librarians and staff members with diverse expertise and duties were assigned to work as teams on individual digital projects. A typical team would include a project manager, a metadata manager, and a digitization manager, among others, who could collaborate with each other efficiently and effectively inside the library. A team worked closely with a partner outside the library to make any digitization project working. Workflows were established, as well as improved upon changed working situations, to regulate the carrying out of digital projects at daily basis.

The Workflow Management System

Fedora, a general-purpose, open-source digital object repository system, was selected by the RUL as the framework for NJDH. RUL recognized Fedora as a flexible service-oriented architecture for managing and preserving in perpetuity digital contents. However, Fedora has its limitations in application, primarily its limitations in the flexibility of loading digital objects and in attaching data both to digital objects and related items. So, RUL developed in house a sophisticated software program to accomplish all these ends – the Workflow Management System (WMS).
WMS is a platform-independent Web accessible system for metadata creation and maintenance (Weber and Favaro, 2007). With its online forms connecting to the databases of NJDH, WMS is able to not only handle metadata input and management but also support the uploading of digital objects in varied file formats. WMS has been designed to meet the particular needs of our digital projects and those of our partners. An on-going refinement has been taken for the system over the years in order to provide more functions to accommodate a wider and more diverse need of digital developments and allow cross-domain interoperability. Having been able to allow RUL to make other contributions to the fields of inquiry, scholarship, and intellectual growth outside Rutgers, WMS is planned to develop into the open source software community to benefit other libraries and digital consortia. The system will be soon shared by the libraries at Northwestern University, Pennsylvania State University, Princeton University, and Virginia Tech, through collaborative projects for its further development. It will be released to be an open source product in a library cyberinfrastructure consortium led by RUL.

For the handling of metadata, WMS is able to support a full METS implementation that describes and preserves the digital objects. It adopts metadata schemas that draws on MODS, NISO, and PREMIS and maps them to Dublin Core and MARC. Five types of metadata are adopted by WMS for the digital collections – descriptive metadata, source metadata, technical metadata, rights metadata, and digital provenance metadata. MODS as descriptive metadata is suitable to be mapped into MARC format, while PREMIS as both source metadata and rights metadata can accommodate digital source elements and identify rights and permissions for individual items. With ready controlled vocabulary for LCSH subjects, AAT Genre and other standards, WMS allows metadata creators to make selections and add new terms.

Through the creation of METS-XML wrappers, WMS has been approved to be efficient in transferring metadata elements and ingesting digital files. At the same time, WMS also supports customized templates to create metadata specifics to the collection owners’ context and to export metadata in Dublin Core for sharing with other repositories. This very capability of exporting metadata makes batch loading from any external database or spreadsheet an automatic and easy job. WMS has been a useful tool for the metadata manager in the process of metadata creation.
and maintenance at RUL as well as with our partners. The batch loading work described below was taken with WMS.

**The batch-loading project**

*Project description*

One of NJDH’s partners is a small New Jersey public library that agreed to participate in the digital project for its collection of about 1,200 historical photographs. The City of Hoboken’s Public Library (HPL) has a photographic collection covering a time range from the mid-19th to the mid-20th Centuries. The collection includes over a thousand of black and white, mostly original, photographs, under the subjects of Hoboken buildings, railroads, ferries, and Hobokenites. These photographs were taken by professionals as well as amateur photographers for social activities, city landscape, figures, dressing, and other historical events in the city. It is the only known collection of this type in the country.

RUL took the responsibility of scanning all photographs in this collection and creating metadata for each photograph. According to an agreed procedure, RUL took around 75 photos from HPL each week and returned them to the library after digitization. It was very helpful that HPL maintained an established small database with MS Access for the description of this collection. The database contained the following information – the photo sequential number, subject, category, caption, date, and source of each photograph. The metadata manager at RUL found that the datasets could provide a rich source to the creation of metadata elements for the project, and the format of these datasets would be able to save considerable time if they were re-organized and imported in bulk to the NJDH’s database.

In order to take the advantage of the datasets, the metadata manager decided to batch-load them through an automated metadata importer on WMS. This automated batch importation required a different procedure than taking metadata into database by human labors. Therefore, it was necessary to develop a new workflow for the importation. The following steps represents a metadata development process developed in the implementation of this digital project:
Evaluation

To carry out the project, an evaluation was taken in two steps. The first step was to verify the existing datasets in the HPL’s MS Access database for accurately describing the collection. It was found that the datasets were inconsistently prepared. Terms were not used with controlled vocabulary so that expressions were made in an inconsistent way from record to record. For example, the names of the photographers were not used uniformly throughout the database. At the same time, inaccurate descriptions were also found. The inaccuracies included incorrect languages used to describe the collection and spelling typos in the datasets. We assumed that such fallacies were the result of hassled works by the original data compliers.

Such confusions would not be detected by a data-importing program automatically and might be normally accommodated in a metadata repository. As a result, data retrieval would become a confusing task and result in ineffective searches. The solutions could include a name authority list or more accurate descriptions. Yet, It was essential that the metadata manager carefully reviewed all information to make them consistent and accurate.

It was also necessary to examine the suitability of the datasets to the structure of RUL’s existing metadata schemas. The RUL’s digital library is built on Fedora, a free operating system that offers the best combination of stable and cutting-edge software and supports a variety of metadata schemas. The descriptive metadata adopts the MODS module, a schema developed by the Library of Congress for multimedia digital materials. The original datasets for the HPL’s photographic collection, however, were incompatible to the MODS structure. Extensive editing was required.

Editing

The purpose of editing the original datasets was to enhance data integrity and warrant data accessibility after the importing of the metadata and digital objects. The editing procedure would involve revising inaccuracies and inconsistencies and reorganizing original data elements. The reorganization might include mapping the metadata to the NJDH metadata schema and adding additional metadata elements. In general, the reorganization would have the following four operations:

1. Splitting. One field in the old datasets might have contained different information that could be split into two or more metadata elements. For example, data in the “caption” field could be broken into several pieces to fill the “title, name subject, business subject, and note” fields;
(2) **Merging.** Data from two or more fields in the old datasets might need to be combined into one metadata element. For example, one part of the data in the “caption” plus the data from the “subject” could make up a new “title” element;

(3) **Converting.** The old field might have an improper name that could be simply transferred into a new metadata field with a different name. For example, the old datasets mislabeled “category” and “subject”;

(4) **Transferring.** Both name and content of a given field could be transferred into a new metadata field without any modification. An example of the transferring was the “date” field. The following chart presents a sample of the relations between the original datasets and the new metadata elements:

<table>
<thead>
<tr>
<th>Original</th>
<th>Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>Shelf Number</td>
</tr>
<tr>
<td>Category</td>
<td>Subject</td>
</tr>
<tr>
<td>Subject</td>
<td>Title</td>
</tr>
<tr>
<td>Caption</td>
<td>Name Subject</td>
</tr>
<tr>
<td>Date</td>
<td>Business Subject</td>
</tr>
</tbody>
</table>

The edited metadata elements were merely a restructure of the original datasets. After the reorganization, the datasets would have contained not only data compatible to elements in the descriptive metadata module, but also data for the source metadata module. For example, the “source” field had data for such source information as “creator” and “provenance”, and the “caption” field recorded when and where a photo was published officially.

One more step then became necessary toward mapping these metadata elements to the WMS structure. In order to be compatible with WMS, it was required to add descriptive metadata, source metadata, technical metadata, and rights metadata. An example was adding the LCSH subjects to each of the photos.

**Testing and consulting with the owner**

The application programmers had previously developed online metadata templates for assistants to manually input data. However, these metadata input templates were not prepared for an automated loading of the metadata elements. To solve the problem, the metadata manager worked with the programmers to customize the existing tool to add an importing function that could perform such a batch loading of the external datasets. When successfully implemented, this automation would be able to save tedious human labor, reduce human error, and increase data accuracy.

With the increased capability of the WMS for batch-loading metadata into NJDH, the metadata manager selected several records to perform the initial testing. This was a necessary step to test the added WMS functions, through which some problems were indeed discovered and resolved. For example, the initial testing revealed that the application was very sensitive to
the “tab” and “enter” keys, which would potentially mess up the entire loading. Also, the WMS was too sensitive to diacritics that records were messed up if diacritics existed. The testing had triggered retrospective work on the metadata and the application. At the same time, this initial testing was designed to let the owner of the collection that recorded the original descriptions check the appropriateness of the metadata editing and mapping.

As soon as the initial testing was finished and necessary modifications were completed, the digital project manager and the metadata manager went back the HPL to ask them to help review the results and provide suggestions. All feedbacks were then brought back for the next route of examination and re-testing.

Re-editing

Based on suggestions from the HPL, a second editing process was scheduled. Upon the completion of the second editing, the metadata manager brought several samples of the metadata to the Metadata Working Group for review.

Additional issues raised in the succeeding process required further resolutions. One of such issues was the inconsistency in naming files, which was not detected during the first testing because of the small testing size. Several student assistants were involved in, under the supervision of the project manager, the digitization of the photos. Each digitized image was given a file name to uniquely identify it. Because batch loading is an automated process, the file names must be perfectly consistent, which unfortunately was not always the case in practice due to inevitable human errors. The solution was using a programming method to automatically generate file names and compare them to the ones created in the production process. Whenever a conflict was detected, the metadata manager examined it, and a final judgment was made accordingly.

Another big challenge with this project was that the undertaking of metadata loading had to be frequently adjusted with an on-going enhancement of WMS functions. This meant that we had to repeatedly test and evaluate the results while new production was at work. We spent a large amount of time carrying out testing with each new release of WMS, which could have been as many as forty-seven times a year.

Conclusion

This paper has briefly described the structure and workflow of the New Jersey Digital Highway project. Challenges and achievements that the metadata manager has encountered in the batch loading of metadata for a historical photographic collection of Hoboken Public Library are described in detail. The workflow that the metadata manager followed is introduced that contains the steps of evaluation, editing, testing, consulting with owner, examination and re-testing, and final batch loading. We have learned a lot from the practice. It is hoped that our experience on the batch loading can provide useful reference for our colleagues in their own digital projects.

It is interesting to note that after this paper was complete, RUL had another opportunity to practice the batch loading of metadata. This was another collaborative project between RUL and
Virginia Tech Library that worked on digitizing the archival materials in memorable of the April 16 shooting rampage at Virginia Tech in 2007. The procedure of automated metadata creation created for and exercised in the Hoboken photographic collection which is introduced in this article was applied again. The batch loading process had achieved the same success for this project as it did for the Hoboken project. This second practice has served as a perfect test bed for validating the established workflow and demonstrating its implications to other digital projects.

Acknowledgements

The author wants to thank Grace Agnew, Mary Beth Weber and Rhonda Marker at Rutgers University, Lina Podles at the Hoboken Public Library, and Sharon Favaro at the University of Maryland for providing great helps to me in carrying out the digital project of the Hoboken photographic collection. The project manager, Isaiah Beard, was part of the team and was appreciated.

References

