October 2007

Digital repository implementation: a toolbox for streamlined success

Michael J. Bennett
University of Connecticut - Storrs, michael.bennett@uconn.edu

Abstract:

Purpose – The purpose of this paper is to describe the tools and strategies that were employed by C/W MARS to successfully develop and implement the Digital Treasures digital repository.

Design/methodology/approach – This paper outlines the planning and subsequent technical issues that arise when implementing a digitization project on the scale of the large, multi-type, automated library network. Workflow solutions addressed include synchronous online metadata record submissions from multiple library sources and the delivery of collection-level use statistics to participating library administrators. The importance of standards-based descriptive metadata and the role of project collaboration are also discussed.

Findings – From the time of its initial planning, the Digital Treasures repository was fully implemented in six months. The discernable and statistically quantified online discovery and access of actual digital objects greatly assisted libraries unsure of their own staffing costs/benefits to join the repository.

Originality/value – This case study may serve as a possible example of initial planning, workflow and final implementation strategies for new repositories in both the general and library consortium environment.

Keywords – Digital repositories, Library networks, Data management.

Paper type – Case study

INTRODUCTION

Building digital repositories, from their initial conception to their first day of online public access, is a journey of real world logistical twists and turns. Though much has been written on issues of general planning, digital imaging and metadata standards, the documentation is by and large scattered throughout digitization journals, existing repository FAQ's, conference PowerPoint handouts or the depths of digitization discussion groups. Shaping the particulate detail that must be taken into account in order to build a successful and robust repository is a time consuming task of constant research and reading. In turn it is exactly at this nexus where the case study can often assist the new project manager in mapping out strategy and options that may best fit their own unique situations from the outset.

FIRST STEPS

The planning behind Digital Treasures, http://dlib.cwmars.org, began in the autumn of 2005 with a road trip that took the author and Central/Western Massachusetts Automated Resource Sharing's (C/W MARS) executive director, Joan Kuklinski, to a pair of existing repositories in eastern Massachusetts. The author's background to that point in time included library management, automation, digital photography, graduate digitization work with Greenstone digital library software in a small-scale pilot project, and recent attendance at the Northeast Document Conservation Center's "School for Scanning" three-day seminar in Boston during the spring of that year. What was gleaned from the road trip included important ideas on building a scanning lab from scratch

(possible physical space and equipment specs), metadata standards to follow (Dublin Core, not MARC), and suitable and affordable software packages (CONTENTdm).

C/W MARS is a large automated network of more than one-hundred forty member libraries. This fact alone presented unique issues of complexity beyond digitization's inherent intricacies. Though it was felt that the network possessed a clear overall vision of the repository it wished to build, implementation of such a unique project on the scale being considered had never been attempted there before. In turn, it was decided early on that it would be advantageous to the success of the venture if the network brought in partners to assist in planning, outreach and metadata training. Both the Central (CMRLS) and Western (WMRLS) Massachusetts Regional Library Systems were natural and willing candidates for this cause.

After a joint planning meeting with individual Regional systems' representatives, a series of three subsequent general informational meetings for interested libraries were scheduled and run. These were held throughout the central and western parts of the state and outlined both digitization's advantages and the repository's scope which was intentionally narrowed in order to make initial implementation easier to accomplish. Item limitations included only reflective material, preferably in the public domain that was no larger dimensionally than the network lab's 11'x17' scanner bed. Interested attendees were then asked to sign up for one of the Dublin Core group training sessions that CMRLS's technical services specialist, Dodie Gaudet, developed and both regional systems subsequently hosted.

It was at this juncture, while metadata training was taking place and libraries were considering possible materials for digitization, when the CONTENTdm software began to be tested at the network. It was also during this period when plans for the greater logistics of the entire project from digitization, to metadata capture, to the accurate and efficient melding of both among a varied and geographically dispersed membership were also formulated.

Hovering over the project were a number of salient facts. C/W MARS had made a large capital expenditure in hardware and software in an area that was novel to the network and its libraries. Half of the author's fulltime hours were now being devoted toward digitization and project management which meant that the network had also to hire and pay a new part time person to take over aspects of the author's old position. Member libraries admittedly new to digitization were as a result unsure of the cost/benefits with regard to their own expenses in staff and time towards item selection and metadata creation. This led to a subsequent push from the network's administration to implement quality initial digital collections as quickly as possible in order to assure those weighing possible involvement that discrete deliverables were achievable in a real and not simply speculative sense.

The top-down collaborative administrative approach that the network and regional liaisons adopted was found to be greatly appreciated by the libraries that had little experience in digitization up to that point. It also allowed the author, as project manager, to call meetings on short notice with the administrative group in order to efficiently

deliberate key issues as they arose and to keep the steps towards implementation progressing at a smooth pace.

NEW METHODS = NEW TOOLS

In order to save expenses, the decision was made that the author would attempt to setup and configure CONTENTdm without the optional fee-based site visit and training that the vendor offered. This was eventually accomplished through the vendor's documentation and email queries to the CONTENTdm help desk and listserv. Gradually a working knowledge of the software's global settings and basic php structure was cultivated.

It was found that this structure could be easily manipulated in Dreamweaver where such objects as the global header and footer for the digital library's pages might be conceived and controlled. In turn, a unique brand and logo for the global header were created through Adobe Photoshop in addition to the site's homepage and "about" page body text that would describe to end users the nature of the project and also the metadata and imaging standards that the repository would be following.

Level I CONTENTdm subscriptions come with eight acquisition station licenses. This client-based software allows scanned images and their corresponding descriptive metadata to be put together and spell-checked before being sent to the repository server for final review, indexing and web display. This licensing scheme presented an interesting challenge. Just how was a network and regional system of this project's size going to handle library metadata input from so many sources with just eight licenses?

Part of the answer was already given by the fact that the network wished to maintain consistent digitization and administrative metadata standards and practices in-house within its own central site scanning lab. Yet, it was strongly felt that the libraries, already faced with the task of delivering their items to the lab, should also not have to spend their time keying in metadata at a dedicated lab terminal that had the acquisition station software on it. As a result this suggested a system design of remote online metadata entry by libraries before any scanning took place. This setup would also help ensure that once real items got to the lab that there would be a matching piece of metadata electronically in place. Online metadata entry separate from the acquisition station software structure and licensing would in addition allow multiple libraries to do such unique descriptive work synchronously without confusion, when and where they could fit it into their schedules.

In December the author began efforts on developing an online Dublin Core form in Macromedia Dreamweaver. The idea was for a form that would be password protected and hosted on the repository server. Fields would include not only simple descriptive Dublin Core elements but also Barcode, Library/Institution, Cataloger, and Email fields. take in Figure 1, “Online Dublin Core Metadata Entry Form”

The only problem at this point was trying to figure out how to make the form work. This required scripting of some sort. Unfortunately writing such code from scratch was a skill the author did not possess.

Luckily an inexpensive piece of web development software was discovered called Forms to Go which greatly helped automate the scripting process. After a number of

tests and tweaks, a php file was finally created that took a test online form submission and dumped all of the Dublin Core element data into an email to the author's inbox. As part of the script, the catalogers or metadata specialists would also receive copies of these emails as confirmation of their submissions with the item barcode, cataloger's name and institution entries placed into the email's subject line. The subject line information would be particularly important to the author in the lab in order to sort simultaneous submissions from differing member libraries into project folders and also to match the right piece of metadata to the right physical item to be scanned. A sample subject line would appear like so: "DC Record for 38119004103069 from Jane Doe at Jones Library." take in Figure 2 "Scripting E-mail Delivery of Dublin Core Form Submissions through Forms to Go"

After a library's batch of items had been scanned and jpeg derivate images were created in Photoshop and queued into the lab's acquisition station software, the strategy was to copy and paste each image's corresponding metadata from the author's Microsoft Outlook Web Access email account into the acquisition station as well. Subsequent to a batch of such images and descriptive records being put together, metadata would then be spell-checked and the soon-to-be digital objects would finally be uploaded through the network's LAN to the repository server. There these new objects would first have their metadata automatically indexed by CONTENTdm's server software, and then they would finally enter either a new or pre-existing Digital Treasures online collection.

This system worked flawlessly until the second library began to make its metadata submissions. What the library found was that for particular entries, they would receive

no confirmations. The author, in turn, would also not receive email copies of these problematic records. This led the library to make a number of frustrating duplicate submissions before they called the network for help.

What was discovered with the assistance of C/W MARS' systems staff was that the network's Microsoft Exchange email server, through which these Dublin Core emails were flowing, was interpreting certain pieces of metadata as spam and was blocking the emails through the server's heuristic spam engine. Once adjustments were made to these spam settings, the emails once again moved normally.

According to vendor documentation CONTENTdm's acquisition station software has the ability to ingest character delimited metadata text files and parse these out into coherent records. It is the author's hope to someday create a script that will automatically create such individual text files for each submission and that these files may then be dumped into the acquisition station software automatically. This would alleviate the need to copy and paste metadata from email text and thus save a great amount of lab time that currently goes toward this tedious work.

Prior to the CMRLS Dublin Core workshops, Dodie Gaudet researched a number of existing repositories, surveyed metadata standards and developed both training and general descriptive metadata standards documentation. The author too was looking at metadata, except from the administrative end of the spectrum where the recording of digital capture and derivative image creation history is important in terms of future refreshment and migration cycles. Through a Microsoft Access table and form, an administrative metadata database was created to record such information. Using PDF file

creator freeware known as PDF reDirect, PDF versions of both the network's descriptive and administrative metadata standards as well as the scanning lab's imaging standards were produced from Word document originals. These would be open for public view from Digital Treasures' "about" link once the site went live.

Take in Figure 3 "PDF reDirect"

By late February 2006, two libraries (one from the central and one from western Massachusetts region) had both submitted twenty Dublin Core records and their corresponding items to the scanning lab. With all systems in place, digitization commenced, and the first two CONTENTdm collections were created. On March 1st, the author notified C/W MARS' systems and networking department that all looked good, requested that the server's port be taken out from behind a firewall block, and that the two new digital collections be served up to the Internet. It had been a six month journey from that initial planning road trip to this point in time. The repository was finally live.

As additional libraries began to express interest in the project and were walked through the logistics of the Dublin Core form and scanning arrangements, attention was turned towards implementing a method of retrieving end user statistics. CONTENTdm offers a reports module based upon MySQL database structure that can provide collection-level statistics. Unfortunately C/W MARS resources did not exist at the time that would have allowed the set up of a test or mirror server to the live repository. With limited MySQL expertise among network staff, the thought of installing and starting such a database on the live server was not something the author wished to follow through on.

In turn exploration began into what kinds of meaningful statistics one could pull from the repository through a number of web statistics programs. Among the demo versions tested, 123LogAnalyzer proved to be the easiest to run from a month's worth of zipped server log files. With the repository's Windows Server 2003 IIS set to W3C Extended log file format, rich statistics could be generated through the log analyzer software. The next questions became, exactly which reports from the broad spectrum offered would the libraries most likely use, and also how might these figures be best presented to the libraries on a monthly basis?

In the end it was decided that three basic monthly reports for each participating library would be created. Correctly configured filtering options in 123LogAnalyzer allowed for the production of collection-level figures for the number of general hits and visitors, top referring domains, and top searches (by search engine, phrase, keyword). These three reports, all written in HTML by the software, were then saved to the repository server. In Dreamweaver a "reports" page was authored that listed each library with its own JavaScript-powered drop down menu for selection of these monthly figures. The reports page was then password protected and placed as a hyperlink from the "About" section of the Digital Treasures site.

take in Figure 4, "Digital Treasures Usage Reports Page"

Today Digital Treasures hosts twenty-four library collections and more than four hundred digital objects. In five months of existence, the site has had 564,288 hits and 28,592 visitors. These figures have been driven in large part by the exemplary Dublin Core records that have been created by the network's various library staff members, some

of whom are reference or special collection librarians with no prior cataloging or metadata experience beyond Dodie Gaudet's superb training and documentation. Through their diligent efforts and CONTENTdm's support of the Open Archives Initiative, the repository metadata continues to seamlessly be harvested by such portals as Oaister (http://oaister.umdl.umich.edu/o/oaister/) and accurately crawled and indexed by search engines, Google, Yahoo and others.