Aggregating Distributed Digital Collections in the Mountain West Digital Library with the CONTENTdm™ Multi-Site Server

AUTHORS
Kenning Arlitsch
Associate Librarian
Head of Information Technology
J. Willard Marriott Library
University of Utah
295 S. 1500 East, Rm. 463
Salt Lake City, UT 84112 USA
Phone: (801) 585-3721
Fax: (801) 585-5549
Email: kenning.arlitsch@library.utah.edu

Jeff Jonsson
Digitization Systems Analyst
J. Willard Marriott Library
University of Utah
295 S. 1500 East, Rm. 461
Salt Lake City, UT 84112 USA
Phone: (801) 587-3678
Fax: (801) 585-5549
Email: jeff.jonsson@library.utah.edu

Technical paper

KEYWORDS: digital libraries, digital asset management software, Mountain West Digital Library, CONTENTdm, aggregation, metadata harvesting

ABSTRACT

PURPOSE:
This paper describes the creation of the Mountain West Digital Library (MWDL), a cooperative regional program distributed throughout Utah and Nevada. Its metadata are aggregated at a single website. Six digitization centers at the largest universities in both states digitize their own collections and support partner institutions in their geographic regions. Each center runs a CONTENTdm server, and an aggregating server at the University of Utah harvests metadata into a single searchable index. Local control and identity of collections are hallmarks of the MWDL
.
DESIGN/METHODOLOGY APPROACH
This paper describes the technical structure of the MWDL, focusing on the Multi-Site Server aggregating software from DiMeMa Inc.

FINDINGS

The MWDL was the first cooperative digital project to implement this software, and inspired the same infrastructure for a nine-state project known as the Western Waters Digital Library. In 2005 the MWDL will also become the first in the nation to aggregate distributed digital newspaper collections.

PRACTICAL IMPLICATIONS
Describes the creation and aggregation of a distributed regional digital library with digital asset management software that is already in use at over 200 institutions in the United States. The solutions implemented in the MWDL offer a mechanism for large and small institutions to work together in a cost-effective manner.

VALUE
Examines the benefits and problems associated with creating a regional digital library.

BACKGROUND

The Mountain West Digital Library[1] was established in early 2002 after a pilot project between the University of Utah's Marriott Library and the Utah State Historical Society (USHS) proved the concept of remote digitization support using CONTENTdm software. After scanning a collection of glass plate negatives for the USHS, the Library created collection space on its CONTENTdm server and gave the USHS client-side CONTENTdm Acquisition Station software. The USHS imported the digital images into the Acquisition Station, added metadata, and then uploaded the images to the Library's CONTENTdm server. The USHS built its own website, with dynamic links into the CONTENTdm server to search and browse the images in the collection. To a user visiting the digital Shipler Commercial Photographers Collection[2] it appears as if they never leave the USHS server, though all images are actually served from the Library at the University of Utah.

The pilot project worked so well that the initial 400-image contract between the Library and the USHS turned into a 10,000-plus image contract. The Library charged a fee for every glass plate scanned, generating enough revenue to purchase two high-end flatbed scanners and hire part-time technicians to do the scanning work. The USHS paid for the initial work with a grant from the Library Services and Technology Act (LSTA) and the larger contract with a grant from the National Endowment for the Humanities (NEH). The success of this initial project laid the ground work for the idea of a state-wide digital library.

University of Utah discussions with DiMeMa Inc., the creators of CONTENTdm, revealed that the company was developing code that was designed to harvest metadata from distributed CONTENTdm servers to form a searchable index with links to the images on the remote servers. This architecture could then support the searching of a larger "virtual" set of collections state wide while still maintaining local initiative and control. The parameters were in place for the next stage.

THE UTAH DIGITAL LANDSCAPE IN 2001

In 2001 the largest academic institutions in Utah were just beginning to develop digital imaging projects. Brigham Young University was most advanced and was leading a collaborative effort with the University of Utah and several other institutions to digitize the Overland Trails

Collection[3] for the American Memory Project. Southern Utah University was in the midst of creating a multimedia digital collection known as Voices of the Colorado Plateau.[4] And Utah State University had begun to digitize some photographs and a collection of Jack London book covers. None of the institutions owned strongly established digital asset management software.

UTAH ACADEMIC LIBRARY CONSORTIUM

The Utah Academic Library Consortium (UALC)[5] comprises the libraries of the fourteen institutions of higher learning in Utah, as well as the State Library and the Utah Technical Centers. Nevada academic libraries are official members of the UALC. UALC has enjoyed a remarkable cooperative spirit, and its achievements include the purchase of a state-wide integrated library system, a federated state-wide catalog search[6], coordinated collection development, a shared database-purchase agreement, and a state-wide document delivery service known as Utah Article Delivery (UTAD)[7].

MWDL PROPOSAL

Following the success of the pilot project between the University of Utah and the State Historical Society the UALC Digitization Committee submitted a proposal for state-wide digitization to the UALC Council. The proposal outlined a plan to establish regional digitization centers at the following universities in Utah: Utah State University (USU); University of Utah (UU); Brigham Young University (BYU); and Southern Utah University (SUU). By participating in the project each center agreed to create and contribute their own digitized collections as well as support partners who wished to create digital collections, but who lacked the resources to establish their own digitization infrastructure. Partner institutions were generally understood to be public libraries, museums, historical societies, other colleges and universities, and other cultural heritage institutions. Partners were expected to pay a digitization center to scan their materials, thereby helping to offset the cost of scanning equipment, personnel, and server space. Suggested pricing ranged from $.50 for 35mm slides to $4.00 for print photographs, and $6.00 each for large-format glass plate negatives.

The proposal included the following goals:

- Build a digital collection reflecting the history of the region
- Include resources from regional cultural and educational institutions
- Offer local control and low-cost digitization to partners
- Standardize metadata for interoperability
- Make content accessible to all Internet users

The proposal requested $100,000 from UALC Council, mainly for purchase of hardware and software. Recipients of the proposed funds pledged to provide additional hardware as needed and to maintain and upgrade software licenses. Since the University of Utah and BYU had already established digitization centers most of the proposed funding was aimed at USU and SUU.

CONTENTdm software for USU and SUU:

- $10,000 x 2 = $20,000 (Level 2 licenses)

Flatbed scanners and computers for USU and SUU:

- $20,000 x 2 = $40,000:

Multi-Site Server hardware and license:

- $10,000

Part-time hourly scanning technicians:

- $7,000 x 4 = $28,000

Training:

- $2,000

Full funding was awarded by unanimous vote of the Council in January 2002, and by May all four centers were up and running. In September the first aggregation of collections occurred and a MWDL website was built for searching and browsing. The two largest universities in Nevada – University of Nevada-Las Vegas (UNLV) and University of Nevada-Reno (UNR) – joined the MWDL in 2003.

TAKE IN Slide 1
CAPTION: Mountain West Digital Library website

LIBRARY SERVICES AND TECHNOLOGY ACT

Partner institutions in the MWDL rely on grants and other funding sources to digitize their collections at one of the regional centers. Much of this funding has come from the Library Services and Technology Act (LSTA), monies from the Institute of Museum and Library Services (IMLS) which the fifty states distribute through locally administered grant programs. The LSTA Council in the state of Utah recognized early on that the MWDL could be effective in helping to maximize grant funds awarded for digitization projects. Council resolved to require participation in the MWDL for all digitization proposals to assure that digital collections were created according to national standards, and that they would become part of the state-wide collection. The cooperation of LSTA Council has helped tremendously to make the MWDL a success.

TAKE IN Figure 2
CAPTION: Structure of the MWDL

TECHNICAL STRUCTURE OF THE MWDL

Figure 2 shows the technical structure of the Mountain West Digital Library as it was first envisioned for the state of Utah. The ovals represent the regional digitization centers, each running a CONTENTdm server on the platform of its choice (Linux, UNIX Solaris, or Windows). Each center supports multiple partners by providing scanning services, Acquisition Station software, training, and support. The Multi-Site Server, running at the University of Utah harvests metadata from each of the centers, and users search the aggregated index through the MWDL website.

TAKE IN FIgure 3
CAPTION: Digitization workflow for MWDL partner institutions

MULTI-SITE SERVER

The CONTENTdm Multi-Site Server (MSS) quickly and efficiently harvests the index information from multiple CONTENTdm Servers. It presents a unified, aggregated search capability of all harvested collections, displaying a thumbnail and index data for each resulting record. Once a user clicks a thumbnail, he/she is instantly connected with the item using its unique URL from the host server.

CONTENTdm is an XML database, storing its index data in a monolithic tagged description file, with an offset file used to locate records within the larger file. The images and thumbnails are stored in a separate directory within the collection directory, which are exposed to the World Wide Web as uniquely addressable items. DiMeMa Inc. chose to build their software this way in order to maximize both extensibility and flexibility. The Multi-Site Server is an example of the flexibility of this back-end database software.

The Multi-Site Server was delivered by DiMeMa Inc. in September 2002. It was installed on a Windows server[8] at the University of Utah and harvesting began. Regional centers identified collections on their CONTENTdm servers for harvesting, providing the server address and the names of the collection directories. Currently the MSS uses command line configuration utilities to add collections for harvesting, and to execute the harvesting itself.

Because of its design, the MSS can run on a small, low-cost server, and still deliver good performance. The MSS harvests the indexed metadata files from the remote CONTENTdm servers using wget.[9] It downloads only the current index and offset files from the remote sites and stores them locally on the MSS, where it offers the indexes for searching at a central location. This lightweight data harvesting method is, in effect, invisible to the remote sites. The collection index files use little disk space, and take only seconds to download from the remote servers at LAN speeds. The MWDL database currently hosts 290,000 records, and the database files for the whole repository only use 2.5 GB of disk space.

TAKE IN Figure 4
CAPTION: The main search interface to the Multi-Site Server, as customized for the MWDL

A search at the MWDL website queries the MSS indexes, producing a local result. The search results page is generated by the MSS, and since the thumbnail images are not stored on the MSS itself, they are pulled, in real-time, from the remote CONTENTdm servers. A click on a thumbnail will instantly redirect the browser to the item on the remote server. If a remote CONTENTdm server is unresponsive at the point of search, the images from that server simply become unavailable, but the results from other servers are unimpaired.

TAKE IN Figure 5
CAPTION: CONTENTdm Multi-Site Server harvesting model

The Multi-Site Server includes a subset of the Custom Queries and Results wizard from the CONTENTdm software. This wizard generates the HTML code needed to create search boxes,

drop-down lists, or predefined queries for collection websites. Varying results views can also be specified in the query, including the following options:

1. Thumbnail images only
2. Titles only
3. Bibliographic view (up to five metadata fields may be displayed for each hit)
4. Grid view (up to five metadata fields, including thumbnail image, may be displayed for each hit)

TAKE IN Figure 6
CAPTION: The search results screen, showing the thumbnail images view (NOTE: apparent duplicates in results 11-16 are actually color and B&W images in the Karl Bodmer collection)

The MSS provides only the initial search results template. The regional center where the larger display image resides provides the local display template, allowing customization by collection and institution. The Web front-end for the MSS can be built to suit the needs of the project.

TAKE IN Figure 7
CAPTION: Search result being displayed from local site, using customized template. (NOTE: browser buttons have been cropped to show full image)

It is important to note that the Multi-Site Server is only one alternative for aggregating searches across CONTENTdm databases. CONTENTdm software is fully compatible with the Open Archives Initiative (OAI)[10], and in technical terms, registering the individual servers of the MWDL with one of the existing major OAI harvesters instead of with the MSS is certainly an option.[11] But a regional aggregation that focused on materials found in the MWDL was most desirable for this project, so registering with an existing OAI harvester to offer searching that included materials from other projects was not considered a good choice. (Future plans do include providing MWDL data to an OAI harvester.)

Local development of an OAI harvester was also a possibility, but it is the out-of-the-box readiness of the CONTENTdm MSS that makes it so attractive. The MSS includes a powerful user search interface, plus the customizable, predefined search and results tools described above, that make it easy to build links, drop-down lists, and search boxes for a website.

Yet another option would have been to purchase or build a federated search system based on Z39.50 queries. CONTENTdm is not natively Z39.50-compliant, but in 2003 the University of Utah developed open source add-on software called ZContent.[12] ZContent is a Perl module developed to convert Z39.50 queries to a CONTENTdm query. It returns results in MARC or XML format.

These options, while decreasing the software costs associated with the MSS, would have created the need for in-house development and maintenance of a front-end application. The Multi-Site Server was the most cost-effective method for achieving the goals of the Mountain West Digital Library.

The Multi-Site Server software installs on a Windows server in just a few minutes. Setting up collections for harvesting takes minimal time using a command-line interface. The MSS for the MWDL was largely pre-configured by DiMeMa as a test case, but a similar project, the Western Waters Digital Library (WWDL),[13] also utilizes the MSS. It took under 3 hours to completely install the MSS software on an existing Windows 2000 server and define and harvest ten collections from the first four institutions participating in the WWDL.

METADATA

The MWDL follows Dublin Core metadata standards as further defined by the Western States Dublin Core Metadata Standards document.[14] Utah representatives participated in the development of the document, which was led by Liz Bishoff as part of the IMLS-funded Western Trails project.[15] The document generally follows Dublin Core definitions and specifies field mappings to default Dublin Core whenever field labels are changed for local use. These mappings are crucial for the MWDL because the MSS searches across multiple CONTENTdm collections using default Dublin Core fields.

Participation in aggregated digital projects naturally requires additional care in adherence to metadata standards. Usage reflecting local needs or biases can become serious problems when those metadata are harvested into larger projects and do not match the usages at other institutions. The most important metadata fields for aggregations where the purpose is mainly to provide a single search interface are those that most users will want to search.

CURRENT CONDITIONS

By early 2005 the MWDL aggregated collection had reached nearly three hundred thousand digital objects of every format. Photographs, documents, books, maps, art prints, audio, and video clips are all represented. The MWDL supports partners in the form of colleges, public libraries, historical societies, and museums ranging from Logan, in northern Utah to Las Vegas, and to Reno, Nevada.

In March of 2004 a second generation beta version of the Multi-Site Server was installed to match the new developments in version 3.6 of CONTENTdm[16]. Updated results and display templates, a new custom queries and results wizard, and MSS Open Archives Initiative data provider abilities were all part of the new version.

WHY THE MWDL WORKS

Cooperation and a respect for the differences in participants' missions, institutional cultures, and funding structures are required to make digital consortia work. (Bishoff, 2004) The MWDL model has been successful for several reasons related to the architecture of the project that help address these differences:

Local control

Partners who contribute materials to the MWDL retain control and ownership of their original materials and the digital versions. Partners of the University of Utah are granted administrative rights to their collections on the center's CONTENTdm server, allowing them to edit metadata as

needed. They are also trained to apply metadata prior to initial upload, thereby retaining control of its application and contributing to the digitization process.

Identity

The concept of retaining identity cannot be overestimated. The flexible nature of the MWDL means the aggregated collections also allow individual creativity and design without duplicating server/storage resources. Each partner is free to construct its own website portals into its collections, and results and display templates allow that website design and identity to be retained even after a user has left the site and entered the CONTENTdm server. Some of the best examples of sites that have retained their identity are the Utah State Historical Society[17] and the Topaz Museum.[18] These institutions house their own websites and have built search and browse queries into the University of Utah's CONTENTdm server using the Custom Queries and Results wizard.

Low cost

The scanning fees charged by the centers include server space, training, and the Acquisition Station software. The cost savings to partners in terms of hardware, software, and personnel is enormous. Long-term storage of high-resolution files is also available for an additional fee if requested.

Digitization Uniformity

Because the digitization is done at the centers it occurs in a uniform manner, using high-end equipment and standards. Adoption of the Western States Dublin Core Metadata Standards, as developed by the Western Trails project, has also insured uniformity and interoperability across collections. Beyond the required metadata fields, all institutions are free to implement whatever additional fields they feel are necessary for their collections.

Revenue for regional centers

The scanning fee generates a revenue stream and helps the MWDL to be somewhat self-sustaining. Revenue at the University of Utah has funded new scanners, servers, and part-time scanning technicians.

CONCERNS

Creation of the MWDL has not been without problems. In any cooperative project there are concerns about willingness, agreement, and equal participation. Following are some specific concerns experienced in the MWDL:

Funding and staffing

Aside from the initial investment of the UALC, and the LSTA grants to fund individual projects, the MWDL has thus far received no additional funding. There is no staff devoted exclusively to the project; participation has been voluntary and in addition to regular jobs. This has impacted the speed of development of the project.

Participation

Not all regional centers have participated equally. Some have supported no partners to date, for reasons that have included personnel issues and initiative.

Notification

The CONTENTdm MSS currently has no ability to discover new collections automatically, and there is no way to harvest all collections from a server without knowing their directory names. Centers are therefore required to notify the University of Utah that new digital projects have come online.

Training and support

Partners who are supported by a regional center are dependent on training and continued support from that center. The CONTENTdm Acquisition Station software is not difficult to use, and an afternoon of training is usually sufficient to get started, but basic computer skills are a requirement. The diversity of institutions in the MWDL naturally means there is a diversity of technical expertise, and remote support can be challenging.

Data Transfers

CONTENTdm offers a Full Resolution Manager tool, allowing collection administrators to keep track of archival digital files, whether they choose to store them online or offline. University of Utah has encouraged its partners to store their archival files on the University's server, which requires a data transfer via FTP. Delivering high-resolution scanned images to partners for import into the Acquisition Station, and then uploading them back via FTP is not always feasible due to variations in Internet reliability and the large size of archival files. After experimenting with CD and DVD we have decided that the most reliable method for the high-resolution images is to mail external hard drives.

OAI Policy Issue

In 2004 we convinced DiMeMa Inc. (against their better judgment) to build OAI-compliance into the MSS. We thought that it would greatly simplify things if we could provide data to OAI harvesters at the aggregated level, but of course the danger is that redundant files from the MWDL could find their way onto OAI harvesters if individual institutions were also registering their CONTENTdm servers. In late 2004 the UALC Digitization Committee agreed to leave OAI data provider responsibilities with the participating institutions, and we have thus left disabled the OAI capability of the MSS.

OTHER PROJECTS MODELED AFTER THE MWDL

Western Waters Digital Library

The Western Waters Digital Library (WWDL), as mentioned earlier, is an IMLS-funded project to develop a digital library of water information resources for the western states. Led by the Greater Western Library Alliance (GWLA), a consortium of thirty research libraries in the Midwest and West, the WWDL is modeled after the infrastructure of the Mountain West Digital Library.

Twelve of the thirty GWLA institutions are actively participating in the two-year project, which focuses on four major river basins: Platte, Rio Grande, Colorado, and Columbia. As in the MWDL each institution runs a CONTENTdm server on the platform of their choice, and a Multi-Site Server installed at the University of Utah harvests metadata from the participants.

The grant began in November 2003, and by February 2004 the new WWDL Multi-Site Server had harvested existing digital collections at four of the participant libraries, quickly bringing the WWDL online with nearly 30,000 digital objects.

Utah Digital Newspapers

Over the past two years the University of Utah has developed an extensive newspaper digitization program, again employing CONTENTdm as the digital asset management tool. (Herbert and Arlitsch, 2003) There are already more than 200,000 pages posted, and with current grants from IMLS and LSTA we expect to reach more than 400,000 pages by late 2005.

As a partner to the University of Utah in the IMLS grant, BYU is separately digitizing 40,000 pages of the Deseret News and posting them on their own CONTENTdm server. During the early part of 2005 aggregation of the distributed collections with the MSS will begin. We believe this will be the first aggregation of a distributed digital newspaper collection in the nation.

Because of the zoning of individual articles the newspaper digitization process creates a much greater number of files than the actual pages counted in the project. The 200,000 newspaper pages on the University of Utah's CONTENTdm server translate to over two million individual files. For this reason we have been leery of simply aggregating them into the existing MWDL; the sheer number of files would overwhelm most searches with newspaper results. The aggregation plan calls for the installation of a separate MSS, which will aggregate only newspaper collections. Users of the MWDL will be able to select whether to include newspapers in the general search, or whether to search them separately. And the aggregated search will also be available from the Utah Digital Newspapers website.

CONCLUSION

There is always more work that can be done. The MWDL website can be improved to offer guided research rather than simply search and browse functions. Establishing a dedicated staff to manage the project, and securing a direct source of funding to support the MWDL would help enormously.

In its two years of existence the MWDL has grown to nearly three hundred thousand digital objects, and continued growth is expected. We have established a viable technical infrastructure that will serve us well into the foreseeable future, and the success and cooperative nature of the MWDL positions us well for future grant funding. It is the shared vision, cooperation, and enthusiasm of the members of the UALC that has made this project possible, and as long as that spirit endures the MWDL will continue to prosper.

NOTES

[1] Mountain West Digital Library website (2002). Retrieved June 20, 2004 from http://mwdl.org

[2] Shipler Commercial Photographers Collection, (n.d.). Retrieved June 13, 2004 from the Utah State Historical Society website: http://history.utah.gov/Photos/C275/
[3] Trails to Utah and the Pacific: Diaries and Letters, 1846-1869. *American Memory Project.* Retrieved: June 17, 2004 from http://memory.loc.gov/ammem/award99/upbhtml/overhome.html
[4] Voices of Colorado Plateau, 2002. Retrieved: June 17, 2004 from http://archive.li.suu.edu/voices/
[5] Utah Academic Library Consortium (n.d.). Retrieved: June 17, 2004 from http://www.ualc.net
[6] Utah's Catalog (2003). Retrieved June 20, 2004 from http://www.lib.utah.edu/kvk/
[7] Utah Article Delivery (UTAD (n.d.). Retrieved: June 17, 2004 from http://www.lib.utah.edu/ualc/jour.html
[8] The Multi-Site Server, like the CONTENTdm software, runs on three different platforms: Linux; Solaris; and Windows
[9] GNU wget - GNU Project - Free Software Foundation (FSF) (n.d.). Retrieved: June 17, 2004 *http*://www.gnu.org/software/wget/wget.html
[10] Open Archives Initiative (n.d.). Retrieved: June 17, 2004 http://www.openarchives.org/
[11] Michigan's OAISTER and the OAI harvester at UIUC are two examples of major existing OAI harvesters
[12] ZContent (n.d.). Retrieved: June 17, 2004 http://www.lib.utah.edu/digital/ZContent.html
[13] Western Waters Digital Library, 2004. Retrieved June 17, 2004 - http://www.westernwater.org
[14] Western States Dublin Core Metadata Best Practices document, 2003. Retrieved: June 17, 2004 http://www.cdpheritage.org/westerntrails/wt_bpmetadata.html
[15] Western Trails Project, 2004. Retrieved: June 17, 2004 - http://cdpheritage.org/westerntrails/index.html
[16] CONTENTdm software from DiMeMa Inc. Retrieved: July 29, 2004 from http://contentdm.com
[17] Utah State Historical Society (2004). Retrieved: June 16, 2004, from http://history.utah.gov/Photos/C275/
[18] Topaz Museum (2002). Retrieved June 20, 2004, from http://www.topazmuseum.org/archive.html

REFERENCES

Bishoff, L. (15 January 2004) "The Collaboration Imperative." *Library Journal*, Vol. 129, No.1, p. 34.

Herbert, J. and Arlitsch, K. (November 2003) "digitalnewspapers.org: the Utah Digital Newspapers Program." *The Serials Librarian*, Vol. 47, No. 1 and 2.