First Monday, Volume 13 Number 5 - 5 May 2008

First Monday

Metadata provision and standards development at the Collaborative Digitization Program (CDP): A History

by Christopher Cronin

Abstract

What began in 1998 as the Colorado Digitization Project is now known as the Collaborative Digitization Program (CDP). The CDP's *Heritage West* database represents not only the primary product of the organization, but also one of the oldest continuously operating collaborative repositories of cultural heritage metadata in the country. As a basis for the author's forthcoming quantitative and qualitative analysis of Dublin Core metadata in *Heritage West*, the following article offers a history of how the CDP has, over time, organized and managed the metadata provision for its digitization projects.

Contents

Introduction
The collaborative approach
Development of the *Heritage West* database
Development of the *CDP Dublin Core Metadata Best Practices*
Subject analysis and authority control in *Heritage West*
Future directions for metadata at the CDP
Conclusion and further research

Introduction

The Collaborative Digitization Program (CDP), which began in 1998 as the Colorado Digitization Project, has since experienced several name changes and iterations in staffing, funding, and scope. The most recent change came in April 2007, when the CDP merged with the Bibliographical Center for Research (BCR). While the presence of statewide and multi–state digitization consortia have increased dramatically over the past decade, the CDP was one of the first. When it started, the initial goal of the CDP was "to increase access to the special collections and unique resources of … cultural heritage institutions through digitization" [1]. This mission has expanded over time to also "provide resources and training to create digital surrogates of primary source collections" [2]. To accomplish this more robust direction, the CDP's infrastructure has grown to include working groups that have developed and maintained best practices for metadata, digital imaging, digital audio, and collection development. The working groups, comprised of individuals from institutions participating in CDP projects, represent long–term collaborations across four types of cultural heritage domains: archives, historical societies, libraries, and museums. Their work signifies some of the first such efforts within the cultural heritage community — a fact not lost on the founders of the CDP who expressly developed the organization to be a model for collaboration that could be adopted elsewhere (Bishoff, 2000).

The purpose of this paper is to provide a historical context to the development of the CDP's metadata standards and the *Heritage West* database. Such a context will serve to inform the author's quantitative and qualitative analysis of the Dublin Core metadata presently in the database, as well as how the different cultural heritage domains approached metadata provision for CDP projects. This is the first analysis of the metadata in *Heritage West*, which ultimately aims to assess the success of the metadata strategy applied by the CDP to date: "the adoption of the standards by institutions undertaking digitization projects will be the true test of whether the standards are appropriate, meet the needs of the institutions, and allow us to realize the goal of increased access" [3].

 ++++++++++

The collaborative approach

At the time the CDP was conceived, well–established collaborative programs existed among Colorado's public, special and academic libraries (Allen and Bishoff, 1999). However, in order to build a digital library that represented the breadth of Colorado's resources, the organization needed to expand beyond library relationships to include the archives, historical societies and museums that also held significant and unique portions of the region's special collections and primary source materials. As Bishoff (2000) notes, "these collections are widely dispersed, with limited access, sometimes poorly organized and generally underutilized" [4]. One significant advantage of collaboration was the ability to aggregate access to similar types of collections at disparate institutions that otherwise would not be associated (Bailey–Hainer and Urban, 2004). Bishoff (2004) also outlines some of the economic benefits that followed when resources were pooled to develop one set of standards, one infrastructure, one database platform, one training program, and the like.

While collaborative environments among libraries and even archives had already been established for other efforts, collaborations that included historical societies and museums were limited in the late 1990s. Nationally, there was "little evidence of work across institutions of different types at the implementation stage" of digitization [5]. Allen and Bishoff (2002) further document how collaboration among institutions within the museum domain alone has historically been limited, particularly because their primary mission is the stewardship of collections and not the sharing of resources [6]. Collaboration through the CDP between different cultural heritage domains would, for example, offer a first–time opportunity for an archive to associate its collection of materials on early gold mining efforts with a historical society's similar collection somewhere else in the state or region. For the CDP to be successful in its mission to provide access to a comprehensive set of primary source materials, all four cultural heritage domains would need to be involved from the beginning in order to assess, select and/or develop the standards and best practices that would be adopted by all participants in the projects.

Some differences between the domains became apparent when the CDP gathered initial data from prospective participants. One example that specifically affects metadata was the concept of titles. The CDP found that "many museums and historical societies do not include titles for their three –dimensional artifacts, rather relying on extensive description of the physical object for retrieval. In contrast many in the library and archival community frequently make up titles for items without a title" [7]. Analysis of the *Heritage West* database is currently underway to determine if such a distinction between the approaches of these domains truly manifested itself in the projects, or whether the collaborative planning for metadata resulted in a unified application of the CDP's metadata best practices, which mandated the provision of a title for all resources described in the database.

Other cultural and fiscal differences inherent in the domains themselves also made a collaborative approach necessary: "The word 'metadata' is foreign to many in the historical society world, and with a funding environment that is always stretched thin, many small museums and libraries are tempted to purchase the least expensive scanner possible, and set a volunteer to the task. It is our view that with a statewide effort to make equipment, training, and software infrastructure available, that lack of knowledge can and should change" [8]. In terms of metadata, the CDP used OCLC as an analogy. If large and small libraries could contribute to the shared OCLC metadata catalog, so, too, could cultural heritage institutions of all types create a database of reliable metadata that links to digital surrogates of Colorado resources — particularly if those domains were engaged in establishing a set of common best practices together.

Collaboration, though, did not come without its challenges. Bailey–Hainer and Urban (2004) cite the difference between building a "collaborative" versus merely a "cooperative" relationship between domains as being one of the more time–consuming aspects of creating the CDP, particularly when it came to communicating the impact of local practices in a shared environment. In summarizing the types of cross–domain collaboration that already existed in Colorado at the time, Allen (2000) notes that while most welcomed the idea of partnerships, there were nevertheless "fundamental assumptions about competition for collections and visitors that must be overcome" [9]. For example, the sometimes competitive business models employed by museums were perhaps in contrast to the freely–available resource sharing models valued by libraries (Bishoff, 2004). And while the same person may physically visit both a museum and a library, their reasons for doing so can differ greatly; conversely, in an online environment, "users desire increased access to the intellectual and cultural materials in a flexible manner, without concern for who owns the resource" [10]. Allen (2000) also highlights that museums often use their catalogs strictly for inventory purposes, never making them accessible to the public as a resource discovery tool.

Some of the institutions that were considering involvement in the first CDP project did not have an Internet connection at the time and were explicitly concerned that having collections available online would result in decreased visits, and, therefore, decreased revenues (Lutz, 2000). As a part of the CDP's initial grants, two research projects were conducted (1) to assess the impact that online access to digital collections has on museum attendance, and (2) to compare user preferences for accessing online collections that are organized according to the museum/exhibition approach versus the library catalog/database approach (Loomis, *et al.*, 2003; Fry, *et al*., n.d.). Interestingly, both studies concluded that the museum community's 'exhibit' approach for organizing online collections was more likely to prompt an on–location visit than the catalog approach taken by libraries.

For all of these reasons, it was evident that a unified training initiative would need to be offered by the CDP if it was going to produce a sustainable program. Kriegsman (2002) outlines how the CDP approached teaching metadata to such a diverse audience, with diverse terminologies and concepts of what it means to "catalog." Allen and Bishoff (1999) write: "One very clear problem is that while leading research institutions or very large public libraries or museums may be informed about new descriptive practices such as the Dublin Core, Colorado libraries, archives, and museums are not engaged in those practices and need to be better informed about the options" [11]. Similar training would also be necessary for such things as digital imaging and copyright issues, which at the time were even newer to most institutions than cataloging. Five scanning centers were established throughout the state (Colorado Springs, Fort Collins, Durango, Grand Junction, and Denver), with a full–time trainer working directly with participants (Allen, 2000). As the geographic distribution of participating institutions evolved, so did the locations for scanning centers; they are currently housed in Alamosa, Colorado Springs, Denver, and Grand Junction.

++++++++++

Development of the *Heritage West* database

Since its inception, the CDP envisioned that resources digitized as a part of its projects would be managed and delivered by the participating institutions themselves, not by a centralized digital asset management system at the CDP. Instead, institutions would contribute to a union catalog of metadata records that linked to the digitized content at their respective institutions. Today, this union catalog, *Heritage West*, contains metadata records from institutions participating in CDP projects. While each of the different domains participating in these projects had varied histories, traditions, practices, and standards for resource description, none had developed significant experience in the digitization of primary source materials or the cataloging of electronic resources. A review of the metadata standards being used by participating institutions at the time, as well as the standards emerging in the national and international metadata arenas, led the CDP to choose the Dublin Core Metadata Initiative (DCMI) standard for the union catalog. The entire review process has been documented by individuals involved in these initial founding stages of the CDP (Allen, 2000; Bishoff, 2000; Bishoff and Garrison, 2000; Garrison, 2001; Bailey–Hainer and Urban, 2004).

In order to understand the present state of the metadata in *Heritage West*, it is important to put the database into a historical context. In the late 1990s, the members of the CDP understood that "in a distributed networked environment standards are key to success" [12]. However, in looking at fifteen digitization projects already underway in the United States in 1998, the CDP found that "different cultural heritage institutions have different standards and different levels of adoption" [13]. Because standards for metadata provision were just then emerging and were neither commonly shared across domains, nor similarly adopted by participants, the CDP had to develop its own best practices for the metadata to be used by CDP participants within the union catalog itself. While the CDP considered several metadata schemes and standards—including Dublin Core, Encoded Archival Description (EAD), Government Information Locator Service (GILS), Visual Resources Association (VRA), and MAchine–Readable Cataloging (MARC) — they faced the additional problem that there was limited software available to support emerging metadata standards (Bishoff, 2000).

Considering the broader metadata environment at the time, the CDP's Metadata Working Group made decisions about the approach they would adopt by following several key goals and objectives — all of which affected, in their own way, the development of what is now *Heritage West*. Bishoff (2000) writes:

> How do we realize the goal of improved access in a distributed networked environment approach? How do we deal with the diverse set of standards, diverse communities, diverse clientele, diverse missions, and diverse knowledge base? What approach can we take that will realize the goal of increased access, while allowing for local flexibility and autonomy?
>
> After several months of exploration, the CDP metadata working group developed a set of assumptions upon which to make recommendations. These included: the CDP could not mandate one metadata standard; rather we had to build on the standards already adopted by the particular community, offering a variety of standards. The CDP could not rely on the Web search engines to provide access at the desired level. The CDP wanted to offer searching across print and digital collections. [14]

Ultimately, the CDP selected OCLC's SiteSearch WebZ software to provide the public interface of the catalog, with an SQL database storing the metadata records themselves. The DCMI's element set was selected as the common metadata scheme, largely because of the relative manageability of the number of possible element types, the intuitive simplicity of element labels that Dublin Core would provide in the public display, and also for the semantic likeness of Dublin Core elements to those in other metadata schemes: "Rather than adopting a specific communication form such as MARC or EAD, the working group developed a minimum set of elements that must be included in a cataloging or metadata record based on the fifteen Dublin Core elements" [15].

A Web–based data entry system, *DC Builder*, was also developed by the CDP using Cold Fusion to allow institutions to enter new records and edit existing ones (Bailey–Hainer and Urban, 2004). The *DC Builder* interface has experienced several iterations in its design and functionality, mostly to ease the task of data entry by linking to the best practices documentation, or to employ drop –down menus for commonly–entered data (*e.g.*, qualifiers for citing controlled vocabularies, like "LCSH," or specifying an encoding scheme applied to data in a field, like "W3C–DTF," for the format of a date that conforms to that standard). The most recent version of *DC Builder* eliminates the ability to enter dates as free text by requiring W3C–DTF input of the YYYY and MM and DD portions of a date in separate input boxes. This effort did not, however, prevent non –standard date formatting to enter *Heritage West* when records were batchloaded. Qualitative analysis of the database will determine the extent to which a standard date format has been used and not used.

By not mandating a single metadata standard for collections, the CDP sought to facilitate the process of loading metadata from other local databases by crosswalking that metadata into a common delivery format for the union catalog. Garrison (2001) explains that "participants contributing records in MARC or another format provide the CDP with a profile so that the fields coming into the CDP database can be mapped to the CDP Dublin Core record format" [16]. Since the profile (*i.e.*, the semantic mapping from the source metadata scheme to Dublin Core) was initiated by the participating institution, the success of this crosswalking process depended on the

participant having a thorough understanding of the relationship between their source metadata scheme and the CDP's Dublin Core element set. Consistent usage and application of the source scheme was required in order for the mapping to succeed and remain valid across all records in that institution's record set. Successful crosswalking further required that the technical staff at the CDP be able to identify errors during the process.

 ++++++++++

Development of the *CDP Dublin Core Metadata Best Practices*

Like all databases with metadata that spans several years and conforms (or does not conform) to evolving standards, *Heritage West* contains data that is not entirely consistent with current DCMI or even CDP recommended practice. The CDP recognized early on, largely from lessons already learned in the library community, that long–term inconsistency was inevitable:

> Adopting the Dublin Core framework at this early stage is risky; however it is likely to be the best option for integrating resources using a variety of international best practices/standards. Adopting Dublin Core in 2000 is like adopting MARC in 1970. Early adopters of MARC recognized that there would be changes to MARC, that the systems would have to be available to support it, etc. We are facing a similar issue in 2000 with Dublin Core … . [S] oftware supporting both the creation and use of Dublin Core based records is slow to develop and implementation is unsettled due to the evolving nature of the standard. The advantage of adopting Dublin Core is that many specialized communities, archives, libraries, and museums are creating Dublin Core based derivatives for their communities. [17]

Necessarily, the *CDP Dublin Core Metadata Best Practices* (CDPDCMBP) document has evolved over time to reflect changes in the DCMI, the needs of the institutions participating in CDP projects, and feedback received from users of the document. While early CDP projects largely involved digitizing still images, subsequent grant–funded projects focused on textual materials and digital audio. As the nature and scope of these projects grew, additions were made to the metadata best practices to accommodate the different kinds of resources being digitized and their associated formats, and to provide guidance on how to describe those objects.

The first iteration of the CDP's metadata best practices, published in 1999, was entitled *General Guidelines for Descriptive Metadata Creation & Entry*. The document was intended to make resource description accessible to a wide range of "catalogers" within the participating institutions, with each institution having different levels of experience (from novice to expert) in metadata provision. For instance, smaller or more geographically remote institutions often used volunteers to catalog the resources they were contributing to *Heritage West*. For this reason, clarity and accessibility of language in the document was critical. Moreover, these guidelines have functioned as a CDP–specific interpretation of the DCMI's Dublin Core element set and provide expanded and more robust explanations of the scheme for use by participating institutions:

> The intent of the *CDP Dublin Core Metadata Best Practices (CDPDCMBP)* is to provide guidelines for creating metadata records for digitized cultural heritage resources that are either born digital or have been reformatted from an existing physical resource, such as photographs, text, audio, video, three–dimensional artifacts, etc. … .These guidelines have been created to address the needs of a diverse audience of cultural heritage institutions composed of museums, libraries, historical societies, archives, etc. This document seeks to accommodate different backgrounds and metadata skill levels of those charged

> with creating metadata records, including catalogers, curators, archivists, librarians, Web site developers, database administrators, volunteers, authors, editors, or anyone interested in creating digital libraries of cultural heritage materials. [18]

One of the first and most controversial issues that the Metadata Working Group faced while developing the best practices was the DCMI's 'One–to–One Principle,' which states that "Dublin Core metadata describes one manifestation or version of a resource, rather than assuming that manifestations stand for one another" [19]. Essentially, this principle means that to be in true compliance with the DCMI standard, there must be a one-to-one relationship between the resource and the metadata record. However, the CDP Metadata Working Group recognized that at many of the participating institutions, separate cataloging of the digital reproductions would be financially prohibitive (Bishoff and Garrison, 2000; Bishoff and Meagher, 2004). Rather than lose their partnerships altogether, the Metadata Working Group developed the best practices to allow for as much repurposing of existing metadata as possible; separate CDP-specific elements were drafted to accommodate technical and administrative metadata for the digital surrogates.

The first version of the best practices contained eighteen elements. Ten of these elements were designated as mandatory, selected along the same guidelines for the Program for Cooperative Cataloging's concept of core bibliographic records (Bishoff and Garrison, 2000). The mandatory elements were: *Title*, *Creator* (if applicable), *Subject*, *Description*, *Identifier*, *Date Digital*, *Date Original* (if applicable), *Format Use*, *Format Creation*, and *Contributing Institution*. The optional, but recommended, elements were: *Contributor*, *Publisher*, *Relation*, *Type*, *Source*, *Language*, *Coverage*, and *Rights*. Three of these elements — *Date Original*, *Format Creation*, and *Contributing Institution* — were developed by the Metadata Working Group itself and do not form a part of the DCMI's element set. The first two, *Date Original* and *Format Creation,* directly address the CDP's compromise with the One–to–One Principle. The proposed qualitative analysis of the metadata records in *Heritage West* will attempt to assess (1) the extent to which institutions and domain types have adhered to the One–to–One Principle; (2) whether the CDP's splitting of the *Format* and *Date* elements has had an affect on the quality of records that describe multiple or complex digital objects (*e.g.*, analog original and digital surrogate; analog and digital master and digital access files; analog and digital and related resources, like transcripts); and, (3) what the implications are of not following the Principle on records harvested as unqualified Dublin Core by Open Archives Initiative (OAI) service providers.

Within the context of cultural heritage materials, the CDP had to address the fact that participating institutions would need to qualify searches based on the date of the original object, but that "using the *Source* field for this information would negate the possibility of qualifying searches by date" [20]. The separation of these data into two date elements — *Date Original* and *Date Digital* — allowed for such additional search functionality and resource discovery. Analysis of these two elements in *Heritage West*, along with the *Source* element, will provide a better understanding of whether institutions took advantage of these separate elements to provide access to these two types of dates in their descriptive metadata.

Likewise, the DCMI *Format* element, which is intended to only record information about how to display or operate (*i.e.*, use) the resource, was split into two elements: *Format Use* (which followed DCMI's *Format*) and the CDP–specific *Format Creation* element (later renamed *Digitization Specifications*). The latter was conceived by the CDP to record technical metadata about the creation of the digital resource in order to support its long–term preservation. The *Contributing Institution* element was created to accommodate "collaborative projects where records from multiple institutions are combined in a shared database" [21], and whose institutions may not always be semantically appropriate for, or provided in, the *Rights Management* or *Publisher* elements. This element was considered especially important during the first CDP grant when institutions from one domain were asked to collaborate on their project with an institution from another domain. For example, in a library/museum collaboration both would contribute content, but one may be doing the scanning while the other may be providing the metadata (Allen and Bishoff, 2002). Efforts to implement the *Contributing Institution* element, however, have not yet been fully realized, and the *DC Builder* interface has not been configured to accept data entry for this element. Currently, only the institution that entered the metadata record is assigned to this element, which does not necessarily explain the nuances of partnerships between multiple institutions. In addition, this CDP–specific element is being mapped to the *Publisher* element in some OAI service providers like the University of Illinois' *OAISTER* database.

Having grown to include several partners from outside of Colorado, the first version of the metadata best practices was replaced in January 2003 by the *Western States Dublin Core Metadata Best Practices.* This document was revised by a Metadata Working Group with

representatives from Colorado, Kansas, Nebraska, New Mexico, Minnesota, and Utah (Bailey –Hainer and Urban, 2004). In an effort to contextualize the best practices, members of the Metadata Working Group published case studies on how their institutions provided the metadata for their CDP projects and, specifically, how the best practices were used in that process (Cronin, *et al.*, 2003; Meagher, 2003; Meagher, 2002). By 2005, the CDP's metadata best practices document was being widely cited as an exemplar for the implementation and practical application of Dublin Core, and was being used by many digitization projects outside of the CDP.

To reflect the growing usage by a national and international audience, the document was renamed the *CDP Dublin Core Metadata Best Practices* in May 2005, hence removing the connotation that it could only be employed throughout the 'western states.' The current version was released in September 2006 and reflects the most recent grant–funded CDP project — 'A Sound Model: Collaborative Infrastructure for Digital Audio' — by incorporating guidance on metadata provision for digitized and born–digital audio recordings.

Throughout its history, the CDP has encouraged the use of qualified Dublin Core whenever applicable, particularly for those data elements that require context to enhance resource discovery (*e.g.*, the various types of relationships that can be expressed in the *Relation* element) or for those elements that document the use of encoding schemes. While the *Heritage West* database was configured as a data provider for Open Archives Initiative (OAI) service providers that use unqualified Dublin Core, CDP participants were still encouraged to supply the more granular qualified Dublin Core. As with the DCMI element set, the *CDPDCMBP* separates qualifiers into two categories: refinements and schemes. Element refinements usually make the meaning of the data value more granular (*e.g.*, Title.Alternative; Description.Abstract; Coverage.Spatial). Schemes are used to cite or identify encoding schemes, classification schemes, controlled vocabularies, and the like, which define the syntax and structure for the data value itself (*e.g.*, ISO–8601 for dates, IMT for the format of the resource, URI for identifiers, DDC for call numbers, LCSH for subject headings).

 ++++++++++

Subject analysis and authority control in *Heritage West*

The use of controlled vocabularies across the institutions in *Heritage West* is as varied as the institutions themselves. One of the goals of the analysis of the database is to determine the completeness of the data by tracking how often institutions cited the use of specific controlled vocabularies and what those vocabularies are. However, it must be noted that *Heritage West* was primarily developed for resource discovery, linking users to the potentially more robust catalog of the holding institution. It is possible that some institutions simply chose not to crosswalk citations to controlled vocabularies and encoding schemes because that data was being managed at the local level in the source metadata records. But it is also a fact that many institutions used *Heritage West* as their primary or only metadata repository. If information that should have been recorded in the scheme qualifier of Dublin Core records is not present in *Heritage West*, it is likely not recorded anywhere at all.

The CDP's union catalog does not have an authority control system nor is there access in *Heritage West* to the reference structures of taxonomies. As a result, for those institutions that used an established taxonomy for their subject analysis, only the authorized headings will appear in the records. As Garrison (2001) notes, some institutions that participate in CDP projects have no access to controlled vocabularies at all, and others have developed their own local vocabularies.

One of the Metadata Working Group's first methods of addressing the authority control issue was to develop a list of Colorado–specific terms (both topical and geographic) from the Prospector database, the union catalog of the Colorado Alliance of Research Libraries. Participants could search this list directly from the CDP Web site as they created their metadata records. Another list of Colorado–specific personal and corporate author names was created to provide access to authorized forms of names that could be applied to the *Creator*, *Contributor*, or *Subject* elements. This list was created by extracting author headings from MARC records in Prospector that described resources published in Colorado or by known Colorado publishers [22]. Similar lists for Utah and Kansas were developed as a part of a subsequent grant.

It is unclear, however, how often or how effectively the terms lists have been used by participants. To use these lists, participants are required to 'cut–and–paste' from the list into the record in *DC Builder* or in a local database; the records themselves, therefore, do not cite the source of the headings. Because Prospector is a union catalog of libraries' holdings, the vast majority of the topical headings would likely derive from the Library of Congress Subject Headings (LCSH), and the authorized forms of personal and corporate names from the Library of Congress Name Authority File (LCNAF). More specialized vocabularies are not likely to be found in the list created by the CDP. It should also be noted that the lists of Colorado–specific terms have not been maintained or updated since they were first created, so the lists do not reflect any subsequent changes to the authorized forms of LCSH or LCNAF headings.

Garrison (2001) documents one of the CDP's proposed methods of harmonizing some of the different approaches to subject access in *Heritage West* byassigning Dewey Decimal Classification (DDC) to records. This project did not move far beyond the theoretical phase, however, and DDC numbers appear in only a small fraction of the total records (more than likely entered or crosswalked by institutions already using DDC for their cataloging). Efforts to apply automated classification to electronic resources have not developed sufficiently for useful application in the CDP environment, nor has developing a CDP-specific classification tool become a priority for the Metadata Working Group.

 ++++++++++

Future directions for metadata at the CDP

Since its first project in 1999, the CDP has expanded its foundation in Dublin Core to include many other projects. For example, the Rocky Mountain Online Archive (RMOA) collaborative, which uses the Encoded Archival Description (EAD) format to encode finding aids contributed to a centralized database by twenty regional institutions in New Mexico, Colorado, and Wyoming (http://rmoa.unm.edu/). At the same time, some of the content described in these finding aids is being digitized by the participating institutions, with object–specific Dublin Core records providing access through *Heritage West*. Some of this content was created through the CDP's 'Sound Model' grant for the digitization of audio recordings.

The 'Sound Model' grant provided one of the first opportunities to actively explore new ways of providing access to compound digital objects. For instance, if an institution wanted to digitize a videorecording of an oral history interview, but also made derivative digital audio files of the interview, as well as text files of the transcripts, how would the various types of metadata (descriptive, administrative, structural, behavioral, etc.) for all of these formats be handled in the relatively non–hierarchical Dublin Core record structure? And what happens when the entire recording is spread over several consecutive digital files? If the DCMI's 'One–to–One Principle' seemed complicated for metadata used to describe the analog original versus its digital surrogate, it became even more complicated when applied to digitized audio.

In 2005–06, the CDP Metadata Working Group formed the Task Force on the Metadata Encoding & Transmission Standard (METS), which was charged to "explore the impact of emerging metadata standards on current CDP workflows and best practices" [23]. The task force determined that the current data–entry application used within the CDP, *DC Builder*, "is not currently able to support the hierarchical representation of data required in a METS environment. A new infrastructure will be needed at the CDP in order to accommodate METS" [24]. At the time of that study, the Colorado State Library, which supports the CDP's SiteSearch platform, was investigating moving to a new open source program; it was hoped that in the process they would acquire a METS –enabled architecture.

The task force's investigation of METS ultimately highlighted the need to revise the *CDP Dublin Core Metadata Best Practices* to include more guidance on the description of non–image formats and compound digital objects in Dublin Core, as well as the need to investigate a new infrastructure that better supports automatic generation of metadata. The task force concluded that the *CDPDCMBP* could still work within a METS environment, with Dublin Core records functioning as the descriptive metadata component within the METS wrapper. The CDP–specific *Date Digital* and *Digitization Specifications* elements, as well as the *Format* element, could be mapped to the administrative component of METS as technical and digital provenance metadata. Metadata from the *Source* element could be mapped to the METS source administrative metadata

section, and the *Rights Management* element could be mapped to the METS rights management administrative metadata section.

In explaining some of the earlier lessons learned by the CDP, Bishoff and Meagher (2004) cite the development of a metadata mentorship program as one enhancement that the CDP could explore for its projects and through which participating institutions could have additional one–on–one assistance with their metadata provision. Such an approach was adopted in 2006 with the 'Sound Model' grant, wherein institutions submitted their records to the project's metadata editor for review and feedback [25]. The summary findings that came out of this process were outlined in the final report for the 'Sound Model' grant [26]:

- Some of the most incomplete records came from libraries and library districts, and some of the most robust records came from historical societies and museums.
- In general, institutions that had their metadata reviewed at the initial stages of metadata provision produced high quality records; the Metadata Editor recommended a similar approach be taken in future projects, especially for institutions new to CDP and Dublin Core.
- When outsourcing digitization, the vendor should be required to provide technical metadata that conforms not only to the *CDPDCMBP*, but also to format–specific technical metadata standards and established encoding schemes.
- The process of crosswalking an institution's existing metadata into Dublin Core needs to be refined to include a crosswalking template that will provide all institutions with a common and consistent understanding of what needs to be present in their source metadata prior to the crosswalking and loading of records into *DC Builder*.
- Metadata training needs to occur as close to the time of cataloging as possible so participants can immediately put the training into practice.
- Subject analysis on digital audio proved challenging in situations where there was no textual transcript for the recording; the subject headings in many records are therefore uncontrolled and inconsistent, which will negatively affect the CDP's ability to perform pre–coordinated searches on common themes across records in *Heritage West*.
- *DC Builder* needs to be reconfigured to allow for global updating functions, both for an institution to globally change metadata in its own record set (URLs, etc.), but also as a larger database maintenance capability for the CDP's *DC Builder* administrators.
- The public brief record display in *Heritage West* currently only displays the first populated instance of the *Identifier* element in the Dublin Core record. While this might make sense in an environment where the 'One–to–One Principle' is never violated, this simply is not the reality within *Heritage West*. Many of these records contain multiple URLs to different formats of the same content, as well as related resources.
- While *DC Builder* was initially configured to associate each record with a specific CDP grant project, a project association was not provided for the most recent 'Sound Model' grant. This decision may affect the ability to refine and sort search results by project, should that be desired in the future.

++++++++++

Conclusion and further research

The CDP, now nearly a decade old, has provided local, regional, and national leadership on the planning and implementation of standardized yet collaborative approaches to digitizing cultural heritage resources. The major product of the CDP, beyond its best practices, is the *Heritage West* database. The author is currently engaged in the first comprehensive assessment of the metadata in that database. The goal of this research is to quantitatively and qualitatively assess how 46 cultural heritage institutions have employed Dublin Core metadata in their records. The research aims to provide an institutional, record and element–level analysis of the metadata, as well as to compare the metadata at the macro level of the type of cultural heritage domain to which institutions participating in CDP projects belong (archive, historical society, library, and museum). It is hoped that the historical context of metadata provision and standards development provided by this paper will allow for a better understanding of the state of the data in *Heritage West*.

End of article

About the author

Christopher Cronin is Assistant Professor and Head of Digital Resources Cataloging at the University of Colorado at Boulder. He has been on the CDP Metadata Standards Working Group since 2002, and is also presently serving on the Alliance of Colorado Research Libraries' Metadata Standards Working Group and the ALA ALCTS CCS Subject Analysis Committee. More information can be found at http://libnet.colorado.edu/facultyprofiles/public/profile.cfm?id=67.

Acknowledgements

The author would like to thank Liz Bishoff for her long–standing contributions to the cultural heritage community, and for being a mentor to so many in our profession.

Notes

1. Bishoff, 2000, abstract.

2. CDP, *About* page, at http://www.bcr.org/cdp/about/index.html.

3. Bishoff, 2000, para. 22.

4. Bishoff, 2000, para. 3.

5. Bishoff and Garrison, 2000, p. 3.

6. Allen and Bishoff, 2002, pp. 53–54.

7. Bishoff, 2000, para. 16.

8. Allen, 2000, para. 8.

9. Allen, 2000, para. 9.

10. Bishoff and Garrison, 2000, p. 3.

11. Allen and Bishoff, 1999, p. 33.

12. Bishoff, 2000, para. 7.

13. Bishoff, 2000, para. 9.

14. Bishoff, 2000, para. 12–13.

15. Bishoff and Garrison, 2000, p. 6.

16. Garrison, 2001, para. 8.

17. Bishoff and Garrison, 2000, p. 6.

18. CDP, 2006, p. 2.

19. Hillmann, 2005, para. 11.

20. Bishoff and Garrison, 2000, p. 8.

21. CDP, 2006, p. 8.

22. Conversation with Liz Bishoff, former Executive Director of the CDP, 29 November 2007.

23. CDP, 2007, p. 41.

24. CDP, 2007, p. 47.

25. The author served as the metadata editor for the Sound Model grant.

26. CDP, 2007, pp. 132–133.

References

Nancy Allen and Liz Bishoff, 1999. "The Colorado Digitization Project," *Colorado Libraries*, volume 25, number 1 (Spring), pp. 32–35.

Nancy Allen, 2000. "Collaboration through the Colorado Digitization Program," *First Monday*, volume 5, number 6 (June) at http://firstmonday.org/issues/issue5_6/allen/, accessed 8 March 2007.

Nancy Allen and Liz Bishoff, 2002. "Collaborative digitization: Libraries and museums working together," *Advances in Librarianship*, volume 26. San Diego: Academic Press, pp. 43–81.

Brenda Bailey–Hainer and Richard Urban, 2004. "The Colorado Digitization Program: A collaboration success story," *Library Hi Tech*, volume 22, number 3, pp. 254–262.http://dx.doi.org/10.1108/07378830410560044

Liz Bishoff, 2000. "Interoperability and standards in a museum/library collaborative: The Colorado Digitization Project," *First Monday*, volume 5, number 6 (June), at http://firstmonday.org/issues/issue5_6/bishoff/, accessed 8 March 2007.

Liz Bishoff and William A. Garrison, 2000. "Metadata, cataloging, digitization and retrieval: Who's doing what to whom: The Colorado Digitization Experience," In: *Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium*. Washington, D.C.: Library of Congress, Cataloging and Distribution Service (November), pp. 2–16, at http://lcweb.loc.gov/catdir/bibcontrol/bishoff_paper.html, accessed 15 September 2007.

Liz Bishoff, 2004. "The Collaboration Imperative," *Library Journal*, volume 129, number 1 (January), pp. 34–35.

Liz Bishoff and Elizabeth S. Meagher, 2004. "Building Heritage Colorado: The Colorado Digitization Experience," In: T.R. Bruce and D.I. Hillmann (editors). *Metadata in Practice*. Chicago: American Library Association, pp. 17–36.

Collaborative Digitization Program (CDP). "About CDP," at http://www.bcr.org/cdp/about/index.html, accessed 24 April 2008.

Collaborative Digitization Program (CDP). *Heritage West*, at http://www.bcr.org/cdp/search.html, accessed 24 April 2008.

Collaborative Digitization Program (CDP) Metadata Standards Working Group, 2006. *CDP Dublin Core Metadata Best Practices* (September, Version 2.1.1), at http://www.bcr.org/cdp/best/dublin-core-bp.pdf, accessed 24 April 2008.

Collaborative Digitization Program (CDP), 2007. *Sound Model: Collaborative Infrastructure for Digital Audio Final Report, Appendix D: Metadata* (March), at http://www.bcr.org/cdp/projects/soundmodel/docs/AppendixD.pdf, accessed 24 April 2008.

Christopher Cronin, Anna M. Ferris, and Marcelyn H. D'Avis, 2003. "Music to our eyes: Providing metadata for digitized sheet music using MARC and Dublin Core," at http://www.bcr.org/cdp/digitaltb/metadata/cs_cubm_music.pdf, accessed 24 April 2008.

Thomas K. Fry, Keith Curry Lance, Marti A. Cox, and Tammi Moe, n.d. "A comparison of Web–based library catalogs and museum exhibits and their impacts on actual visits: A focus group evaluation for the Colorado Digitization Project," at http://www.bcr.org/cdp/best/reports/cdp_report_lrs.pdf, accessed 24 April 2008.

William A. Garrison, 2001. "Retrieval issues for the Colorado Digitization Project's Heritage Database," *D–Lib Magazine*, volume 7, number 10 (October), at http://www.dlib.org/dlib/october01/garrison/10garrison.html, accessed 10 August 2007.

Diane Hillmann, 2005. "Using Dublin Core," (7 November), at http://dublincore.org/documents/2005/11/07/usageguide/, accessed 25 November 2007.

Sue Kriegsman, 2002. "Catalog training for people who are not catalogers: The Colorado Digitization Experience," *Cataloging and Classification Quarterly*, volume 34, number 3, pp. 367–374.http://dx.doi.org/10.1300/J104v34n03_08

Ross J. Loomis, Steven M. Elias, and Marcella Wells, 2003. "Website availability and visitor motivation: An evaluation study for the Colorado Digitization Project," at http://www.bcr.org/cdp/best/reports/loomis_report.pdf, accessed 24 April 2008.

Paula C. Lutz, 2000. "Colorado Digitization Project: Balancing technology and community connections," *Colorado Libraries*, volume 26, number 4 (Winter), pp. 13–14.

Elizabeth Meagher, 2003. "Access as a metadata bridge," at http://www.bcr.org/cdp/digitaltb/metadata/cs_du_access.pdf, accessed 24 April 2008.

Elizabeth Meagher, 2002. "Digital archives: Images in Colorado history," *Colorado Libraries*, volume 28, number 2 (Summer), pp. 25–28.

Editorial history

Paper received 7 February 2008; accepted 2 April 2008.

Metadata provision and standards development at the Collaborative Digitization Program (CDP): A history
by Christopher Cronin
First Monday, Volume 13 Number 5 - 5 May 2008
http://journals.uic.edu/ojs/index.php/fm/rt/printerFriendly/2085/1957