Final Draft Manuscript of Riley, Jenn, John Chapman, Sarah L. Shreeves, Laura Akerman, and William Landis. 2008. Promoting shareability: Metadata activities of the DLF Aquifer initiative. *Journal of Library Metadata* 8, no. 3: 221-247.

Promoting shareability: Metadata activities of the DLF Aquifer initiative

Jenn Riley, John Chapman, Sarah Shreeves, Laura Akerman, and William Landis

Abstract

This article describes the progress of the Digital Library Federation's Aquifer Metadata Working Group and demonstrates a model for the construction, application, and testing of collaboratively-developed best practices for sharing metadata in the digital library environment. We set the metadata aggregation context in which the Aquifer initiative began, describe the development of a set of Implementation Guidelines for Shareable MODS Records and their supporting documentation and tools, and discuss how this work has contributed to the understanding of what features metadata describing primary source and humanities-based resources needs in order to support scholarly use. We end with a summary of future efforts for the Aquifer initiative, and how its lessons can be applied in other metadata harvesting environments.

Keywords: DLF Aquifer, shareable metadata, MODS, OAI PMH, metadata harvesting

Authors

Jenn Riley, Metadata Librarian, Indiana University Digital Library Program. 1320 E. 10th St., Wells Library W501, Bloomington, IN 47405. jenlrile@indiana.edu

John Chapman, Metadata Librarian, University of Minnesota. 309 19th Avenue South, Minneapolis, MN 55455. chapm157@umn.edu

Sarah Shreeves, Coordinator, Illinois Digital Environment for Access to Learning and Scholarship. University of Illinois at Urbana-Champaign, 1301 W. Springfield Avenue, Urbana, IL 61801. sshreeve@illinois.edu.

Laura Akerman, Technology and Metadata Librarian, Emory University. Room 128 Woodruff Library, Emory University, Atlanta, GA 30322. liblna@emory.edu

William Landis, Head of Arrangement and Description, Manuscripts and Archives, Yale University Library. P.O. Box 208240, Yale University Library, New Haven, CT 06520. bill.landis@yale.edu

Acknowledgement

The Aquifer American Social History Online Project gratefully acknowledges The Andrew W. Mellon Foundation for its support.

Introduction

The Digital Library Federation (DLF) is "a consortium of libraries and related agencies that are pioneering the use of electronic information technologies to extend collections and services."[1] DLF member institutions are among those at the forefront of digital library development. Since 2003, the DLF has experimented with various methods to organize member institutions to work together for the purpose of building frameworks for a large-scale distributed, open digital library. (Kott, 2005) The current instantiation of these plans appears as the DLF Aquifer initiative, a collaborative project of fourteen DLF member institutions. The Aquifer initiative is designed to go beyond simply delivering aggregated collections and their metadata online, and strives to build services for the scholarly use of library collections and tools to enable cultural heritage institutions to better develop and implement those services. The Aquifer label is intended to serve as a metaphor for pooling resources together and "piping" them out in various ways to meet specific needs. Aquifer defines its goals in this way:

> Our purpose is to promote effective use of distributed digital library content for teaching, learning, and research in the area of American culture and life. We support scholarly discovery and access by:
>
> - Developing schemas, protocols and communities of practice to make digital content available to scholars and students where they do their work

- Developing the best possible systems for finding, identifying and using digital resources in context by:
    - promoting digital library best practices and
    - developing tools and services to improve digital resource access and use[2]

The activities of the DLF Aquifer initiative rely heavily on participation in "Working Groups," collections of individuals from Aquifer participant institutions with expertise in various areas. Three Working Groups are currently active: Metadata, Technology/Architecture, and Services. A Collections Working Group existed early in the project, but has been disbanded, as much of the collections-focused work of Aquifer is complete. These working groups are charged with advising Aquifer project activities in their areas of focus, leading information-gathering activities, and helping to develop standards and best practices within the scope of the Aquifer initiative.

To make significant contributions to the state of the art in digital library aggregations, however, Aquifer as an initiative needed dedicated staffing to develop a practical testbed for tool development and experimentation with emerging best practices. DLF therefore sought and obtained funding from the Andrew W. Mellon Foundation for a project entitled *DLF Aquifer Development for Interoperability Across Scholarly Repositories: American Social History Online*. ("DLF Aquifer Development for Interoperability," n.d.) This targeted project, referred to internally as ASHO, aims to test frameworks for scholarly use of distributed digital collections by building a portal of materials related to American social history[3] and pushing these same collections into "the flow" (Dempsey, 2005) of scholarly users through a series of local implementation

scenarios. Four local implementations are being developed in which the effectiveness of ASHO content will be assessed with targeted user groups; these include: a commercial search service, a course management system, a citation management tool, and a federated search tool. Mellon funding for ASHO allowed the Aquifer initiative to hire a small development staff to implement and provide a testbed for development ideas grown out of the Working Groups.

This article provides an overview of the progress of the Metadata Working Group throughout the Aquifer initiative and demonstrates a model for the construction, application, and testing of collaboratively developed best practices for sharing metadata in the digital library environment. We set the metadata aggregation context in which the Aquifer initiative began, describe the development of a set of Implementation Guidelines for Shareable MODS Records and their supporting documentation and tools, and discuss how this work has contributed to the understanding of what features metadata describing primary source and humanities-based resources needs in order to support scholarly use. We end with a summary of future efforts for the Aquifer initiative, and how its lessons can be applied in other metadata harvesting environments.

Precursors to Aquifer

While metadata aggregation has long been practiced in library circles, most notably in union catalogs, it has received renewed attention with the development of the Open Archives Initiative Protocol for Metadata Harvesting (OAI PMH) [4]. Several institutions participating in the Aquifer initiative, including Emory University, the University of Illinois at Urbana-Champaign, the University of Michigan, and the

California Digital Library[5], were among the leaders in implementing OAI PMH after its initial public beta release in 2001 and tested a variety of approaches to implementing OAI PMH enabled aggregations., The most common challenge reported from these early metadata aggregation experiments with OAI PMH is the difficulty in providing services much more advanced than keyword searching on top of aggregated unqualified Dublin Core records (Shreeves, Kaczmarek, & Cole, 2003; Halbert, 2003; Hagedorn, 2003; Shreeves et al., 2005). The National Science Digital Library has reported similar issues when harvesting metadata from small repositories, despite significant investments in training and documentation (Arms et al., 2003; Lagoze et al., 2006).

To help address these challenges and improve the robustness of services that can be provided by OAI PMH-based metadata aggregations, the DLF and the National Science Digital Library (NSDL) convened a working group in the summer of 2004 to develop a set of best practices for both data provider implementations as well as for interoperable (or "shareable") metadata, bringing together individuals with practical experience in both the aggregator and metadata provider sides of the OAI PMH landscape. The working group focused on developing recommendations that would apply in any workflow using OAI PMH, regardless of software or metadata format involved. The DLF/NSDL working group concluded that more robust and machine-understandable metadata was key to improving the state of the art in metadata aggregations. This group's final report, the *DLF/NSDL Best Practices for OAI PMH Data Provider Implementations and Shareable Metadata*, states as a recommendation that "…in addition to unqualified Dublin Core, repositories expose the richest possible metadata formats available for all items in the repository" (DLF/NSDL, 2007, p. 37). The *Best Practices* cite as the

rationale for this recommendation unqualified Dublin Core's lack of semantic complexity as well as its inability to communicate encoding syntaxes and controlled vocabularies used for specific metadata elements.[6] Indeed, the architects of the OAI PMH purposefully left the protocol open to accommodate other metadata formats, assuming that communities would develop their own standard practices that go beyond the baseline unqualified DC requirement. (Lagoze and Van de Sompel, 2003). Along these lines, the *Best Practices* add that:

> The choice of additional metadata formats should be based on the robustness of description desired for the resources in question; the commonly used metadata schema in the community in which the resources will be primarily used; and, if applicable, the needs of a service provider by whom a repository specifically wishes to be harvested. (DLF/NSDL Working Group, 2007, p. 37)

Aquifer as a next generation aggregator

It was in this environment that the DLF Aquifer project began. As an initiative dedicated to scholarly discovery and access, Aquifer aimed to provide "next-generation" services for scholars. Early Aquifer planning work determined that the quickest way to build services designed to test evolving best practices for distributed digital library operations was through the use of OAI PMH, although this decision was made with explicit recognition that this was not the only way that distributed digital libraries would be built in the future. Members of the Aquifer Metadata Working Group had a wide range of experience with metadata interoperability issues, often through OAI PMH, as metadata aggregators, data providers, and MARC catalogers.[7] This group immediately recognized

the potential for achieving the imagined advanced functionality through conformance to the DLF/NSDL *Best Practices* and through the promotion of "shareable" metadata optimized for aggregation (Shreeves et al., 2006). This vision was assisted by the fact that several of the Aquifer Metadata Working Group Members served on the DLF/NSDL working group, and some have been active in further advancing the notion of shareable metadata following the DLF/NSDL work.

The Metadata Working Group believed that Aquifer, as a collaboration between research libraries at the forefront of digital library development, represented a prime opportunity to develop the community-specific metadata practices imagined by the OAI PMH designers. We were therefore comfortable making a decision at a June 2005 meeting that participants in Aquifer would be expected to share MODS records via OAI PMH for at least the first stage of the project testing metadata harvesting workflows. This was not an easy decision, and many present acknowledged that their own institutions would not be able to provide MODS records at the start because of software or resource limitations. However, participants felt that Aquifer should be building on already established best practices and experience. There was also a general acknowledgement that DLF member institutions—given their status in the digital library community—were among those best positioned to provide MODS records via OAI PMH.

Requiring MODS records would allow the Aquifer initiative to start with rich, semantically complex records and would avoid the already well-documented challenges of working with unqualified Dublin Core. Qualified Dublin Core and MARCXML were briefly considered as well. There was a general consensus that MODS was more viable than MARCXML because it could be more easily understood outside of the library

community, and that, while qualified Dublin Core could provide better semantic complexity than unqualified Dublin Core, MODS was still a better fit for the Aquifer project. In addition, the Metadata Working Group recognized practical challenges to implementing qualified Dublin Core. First, local practice frequently uses Dublin Core in combination with other metadata standards, and "nonstandard use" is more common for Dublin Core than for other metadata formats (Palmer and Knutson, 2004, p. 5). Second, the nature of qualified Dublin Core as defined in multiple XML Schemas that must be combined has resulted in inconsistent implementation over OAI PMH, with fifteen different namespaces for qualified Dublin Core in use by OAI PMH data providers at this writing.[8]

It was clear that simply requiring metadata providers to expose MODS records would not be sufficient for truly interoperable metadata. Even metadata that works well within its local context can have interoperability issues when shared outside of that environment. Work on the DLF/NSDL *Best Practices* and related research had highlighted challenges to the shareability of metadata such as inclusion of information inappropriate outside of the local environment (such as technical or preservation information) and lack of contextual information that allows users to make sense of a record outside of its local environment (Shreeves, Riley, and Milewicz, 2006). The Metadata Working Group after some discussion determined that its first task was to develop a set of guidelines that would aid data providers in implementing more shareable MODS records. The guidelines were imagined as a specific application of the principles outlined in the DLF/NSDL *Best Practices* to the MODS metadata format.

Development of the *Guidelines*

Despite the benefits of using an existing framework to assist with the development of guidelines for MODS designed to be contributed to the Aquifer initiative, the development of these guidelines took a significant amount of effort. A first draft was created between June and December 2005, with the release of a draft for public review at the end of this period. Feedback on the draft resulted in both major and minor changes, with the stable 1.0 version entitled *Digital Library Federation / Aquifer Implementation Guidelines for Shareable MODS Records* released in November 2006 (DLF Aquifer Metadata Working Group, 2006).

The structure of the *Guidelines* mirrors that of the MODS Schema and User Guidelines, with one section for each MODS top-level element. Each MODS element and attribute is listed, together with a textual description of usage for Aquifer purposes, example XML syntax, a description of how an aggregator would likely use the data, a mapping to simple Dublin Core, and a reference to the relevant section(s) of the *DLF/NSDL Best Practices*. A template for each top-level element entry can be seen in Figure 1. The usage guidelines for each element and attribute in the Aquifer MODS *Guidelines* are specified using terminology inspired by RFC2119 (Bradner, 1997) to express obligation, as follows:

> "REQUIRED" designates an item that is an absolute requirement of the guidelines.
>
> "REQUIRED IF APPLICABLE" designates an item that is an absolute requirement of the guidelines if it is applicable to the resource being described.
>
> "RECOMMENDED" designates an item that an implementer may ignore, but only if she has fully weighed the implications of doing so.

> "RECOMMENDED IF APPLICABLE" designates an item that is applicable to the resource being described and an implementer may ignore, but only if he has fully weighed the implications of doing so.
>
> "OPTIONAL" designates an item that an implementer may use at his own discretion.
>
> "NOT RECOMMENDED" designates an item that an implementer may use, but only after she has fully weighed the implications of doing so. This item is discouraged. (DLF Aquifer Metadata Working Group, 2006, p. 6).

Figure 1. Template for the top-level element entries in the Aquifer MODS *Guidelines*

<element name>

MODS Element	Attributes	Sub-elements
<element name>	attributes	<sub-element>

SUMMARY OF REQUIREMENTS

[Brief summary of requirements]

DEFINITION FROM *MODS USER GUIDELINES*

[Definition from http://www.loc.gov/standards/mods/v3/mods-userguide.html]

DISCUSSION OF USE

[Discussion of typical use of MODS element and content guidelines.]

Attributes:

Attribute name [RFC2119 terminology to express obligation]

[Discussion of use of specific attribute.]

Sub-elements:

<sub-element> [RFC2119 terminology to express obligation]

[Discussion of use of specific element. If there are attributes to specific sub-elements, these are found under the relevant sub-element.]

EXAMPLES OF <element> USE

<examples>

USE BY AGGREGATORS

[Discussion of use by aggregators.]

MAPPING TO DUBLIN CORE

[Guidance on mapping to Dublin Core. MODS examples above expressed in Dublin Core:]

<dc:examples>

RELATIONSHIP TO DLF/NSDL BEST PRACTICES FOR SHAREABLE METADATA

[Pointer to relevant section of the DLF/NSDL *Best Practices for Shareable Metadata*.]

The Metadata Working Group's activities developing the *Guidelines* followed a process of collaboration common to many group projects, incorporating a mix of formal documents, iterative document development on a collective wiki, somewhat less formal phone calls guided by an agenda, and informal email conversations. In nearly all conversations, regardless of the method, there was a tension between complexity and simplicity that drove to the heart of the effort. On the one hand, the goal was to make available a method by which to provide richer metadata via OAI PMH. On the other hand, our recommendations needed to be simple enough to share, document, and implement so as not to pose an undue burden on the data provider.

Members of the Working Group brought to the project a wide variety of opinions regarding how ambitious the Aquifer metadata guidelines should be, showing many different positions between the extremes of prescriptiveness and leniency. In order to achieve consensus decisions, group members encouraged one another to argue from evidence, whether from experience with their own institution's collections or from discussions with those hoping to implement the *Guidelines* upon completion. This

methodology resulted in avoiding the "holy war" arguments common to many metadata issues by focusing on rationale rather than only on an end product. The vast majority of the Working Group's decisions were easy ones, allowing us to spend significant time carefully considering the issues in cases where decisions were not as easy. The difficult cases illustrated thorny issues in the broader metadata landscape, representing areas in which there is no community consensus. Three of these cases will be discussed here: subject vs. genre, content vs. carrier, and dates of resources.

Subject vs. Genre

Previous work in metadata aggregations has suggested that the genre of a resource can be a useful access point. Encoding genre in MODS is simple, as <genre> is a top-level element. Despite the ease of encoding this data, the Working Group struggled to determine the appropriate role for genre information in the *Guidelines*.

The dividing line between the concepts of subject and genre is an unclear one. Like the interplay in design between form and function, the topical nature of a work often informs and is informed by its genre. To further complicate the issue, LCSH practice frequently includes a form/genre subdivision as part of a precoordinated topical heading. Similarly, the boundary between genre, type, form, and style is murky and perhaps drawn differently in different disciplines. The *Thesaurus for Graphical Materials II* from the Library of Congress Prints & Photographs Division, for example, is a vocabulary for genre and physical characteristics of graphical materials, not distinguishing between the two.[9] In some controlled vocabularies, genre is almost indistinguishable from the physical form of a resource, such as in the Library of Congress' *Basic Genre Terms for*

Cultural Heritage Materials, which includes such terms as Albums, Drawings, and Pamphlets.[10] In others, genre clearly refers to the intellectual content of a resource, such as in the Library of Congress Motion Pictures, Broadcasting & Recorded Sound Division's *Moving Image Genre List*, which includes such terms as Adventure, Comedy, and Mystery.[11] In still other disciplines, such as music, genre vocabularies (as distinct from subject vocabularies) are not well-developed or are not widely adopted. Genre terms therefore are often not trivial to provide in metadata records, as the very notion of genre is marred by the lack of a clearly understood definition in the cataloging community. For many institutions, the primary area in which genre is reflected is in subdivisions of topical headings.

Given the potential utility of genre headings, it would have been simple for the Working Group to say that <genre> would be a required element in Aquifer MODS records. Such an approach, however, could have any of several negative consequences:

1. Prospective Aquifer contributors with inconsistent or absent genre information in existing records would balk at this requirement, either limiting participation or requiring contributors to invest a great deal of effort in retooling metadata for contribution.
2. Data of questionable quality would be supplied by Aquifer contributors solely for the purpose of meeting the requirement, diluting the value of the element across aggregated collections.
3. The requirement would be ignored, leading to an overall devaluing of the *Guidelines* in the overall Aquifer metadata workflow.

Despite the anticipated difficulty in requiring the <genre> element, the Working Group was concerned that specifying a lesser obligation would lead to few contributors making use of the element, and therefore the inability of the Aquifer portal to effectively use this information for discovery across collections. Further complicating the issue was a change in how MODS handled genre information. In MODS 3.1, <genre> was a top-level element only. In MODS 3.2 and 3.3, <genre> is also allowed as a subelement of <subject>.

The ultimate solution by the Working Group took multiple forms:

1. An obligation of "Recommended" for the top-level <genre> element
2. A requirement that if a genre is supplied, it is from a controlled vocabulary, and the vocabulary is named in the MODS record.
3. A strong recommendation in the discussion of the <subject> element to prefer the top-level <genre> element over the <subject>/<genre> sublement approach.
4. Language in both the <genre> and <subject> element sections to indicate the desirability of providing genre information, with an awareness that researchers in different areas would expect and benefit from it differently.

Content vs. carrier

The Aquifer initiative was designed to provide access for scholars to materials available on the Web. The Metadata Working Group crafted the *Guidelines* with this in mind, stating in the introduction that "resources to be described are digital (either born digital or digitized from analog originals) cultural heritage and humanities-based

materials in keeping with the Aquifer collection focus on American life and culture" (DLF Aquifer Metadata Working Group, 2006, p. 5). Issues of how to describe various versions of a resource (for example, a digitized item and its analog original) were therefore the subject of much discussion by the Working Group.

One of the signature subtleties and difficulties in metadata work is keeping a descriptive record focused on a single object. When dealing with digital surrogates, the challenge of doing so consistently is formidable. Some metadata applies specifically to a physical item, some only to a digital surrogate, and some to the intellectual content of a resource spanning all versions. For example, metadata on physical location would naturally apply only to an analog original, technical metadata regarding the digitization process only to a digital surrogate; while the place or person depicted in an image would apply to both versions.

The Dublin Core Metadata Initiative has long maintained a dictum of a one-to-one relationship between record and object (Hillmann, 2005, section 1.2). This approach has recently been made more formal with the maturation of the DCMI Abstract Model, in which:

- A *description set* is a set of one or more *descriptions*, each of which describes a single *resource*.
- A *description* is made up of one or more *statements* (about one, and only one, *resource*) and zero or one *described resource URI* (a *URI* that identifies the *described resource*). (Powell et al., 2007)

Some standards such as MARC and MODS have provided for a looser approach, allowing information on both an original and a digital copy within a single record at the

discretion of the metadata creator. One MODS element in particular, <relateditem>, is designed with this flexibility in mind.

The initial draft release of the *Guidelines* took a strict approach to distinguishing the description of a digitized item from its analog original, using the top-level MODS elements for the description of a digital object and the intellectual content it contains, and putting information about an analog original in the <relatedItem> element. This approach was the feature for which the Working Group received the most feedback in the comment period on the draft *Guidelines*. Commenters highlighted three general problems with the draft *Guidelines*: first, that the document itself was internally inconsistent in following the stated approach; second, that it is overly difficult to separate out data related to the content of a resource rather than its carrier, that is, MODS is not explicitly designed to support fully FRBRized approaches (IFLA Study Group, 2008); and third, that this level of intellectual rigor was both unnecessary and unsustainable in the current metadata aggregation environment. There was disagreement among commenters on this issue, however; while most believed the original approach taken was problematic, a few strongly supported the draft approach strictly separating description of multiple versions.

Having heard these comments, the Metadata Working Group created a document describing four possible approaches to the multiple versions problem (seen in Table 1), and presented these at a Birds of a Feather session at the Spring 2006 DLF Forum for further feedback. In each of these approaches, the elements outside of <relateditem> were considered the "main record." Descriptive information independent of format was considered "content" while technical or administrative information pertaining to either analog original or digital surrogate was considered the "carrier." Born-digital materials

were here treated as an edge-case, to be dealt with in the context of whatever approach worked best for the knotty mixed analog/digital paradigm.

Table 1. Content vs. Carrier Approaches Presented to Digital Library Federation Community

Approach 1. Ensure all recommendations conform to original approach	Main record contains data about content and digital carrier <relatedItem> contains data about analog carrier
Approach 2. Strict(ish) interpretation of one-to-one - Digital Primary	Main record contains data about digital carrier <relatedItem> contains data about content and analog carrier
Approach 3. Strict(ish) interpretation of one-to-one - Analog Primary	Main record contains data about content and analog carrier <relatedItem> contains data about digital carrier
Approach 4. Content in main record, each carrier (analog, digital, etc.) in a separate <relatedItem>	Main record contains data about content One <relatedItem> contains data about analog carrier Other <relatedItem> contains data about digital carrier

In addition to these four options presented to the DLF community, the Metadata Working Group considered two additional choices. The first was the simplest of the options to implement: simply put all data in the main record and not use <relatedItem> at all to describe multiple versions. We chose not to present this option to the DLF community because at the time we believed the benefits of a more structured approach to be worth the extra effort to implement it. The second was more complex than any of the other options, employing a fully FRBR-ized approach where FRBR Work attributes are described in the main record, and Expression, Manifestation, and Item attributes each in their own <relatedItem>. This second approach would require the use of displayLabel attributes on <relatedItem> to distinguish between the FRBR entities, an approach the

Working Group believed was far from ideal, and therefore was also not presented to the DLF community.

The discussion of the four options at the Spring 2006 DLF Forum was lively. The lack of parallelism between approaches 2 and 3 emerged as a topic of discussion, raising questions about the rationale behind these choices. Approaches 1 and 2 were sharply criticized, as they include information that would be exposed by aggregators within the <relateditem> element rather than the main record, making them less likely to be understood by aggregators other than those that were specifically designed around the *MODS Guidelines*.

In considering the various possible approaches, the Metadata Working Group compared each according to the following criteria:

1. Be appropriate for the target content – cultural heritage materials
2. Include recommendations for all data elements likely to be useful in an aggregated environment
3. Be conceptually clear
4. Provide predictable places for data elements within records for both digitized and born-digital materials
5. Provide for easy conformance with legacy metadata

In the end, the Working Group, with strong support from several commenters on the draft *Guidelines*, decided that the simplest approach was best. This simplest approach, where all relevant data was put in the main record regardless of which version it applied to, although not presented formally as an option to the DLF community, was selected for

the 1.0 release of the *Guidelines* in November 2006. Following the discussion at the Spring 2006 DLF Forum, the Working Group realized that while there are great benefits to a more structured approach, any of the more complex options would be prohibitively difficult to implement. When describing the chosen approach, the Working Group promoted selectivity in the amount of data to include, encouraging implementers to leave out information that would be unlikely to be of benefit in present-day aggregations—for example, leaving out the date an object was digitized in favor of including only the date the original analog object was created. The Working Group believed this hybrid approach was the best fit with the current state of metadata aggregations (few if any can provide advanced services on multiple versions of a resource) and was the most accessible for the majority of potential Aquifer contributors. It was not without reservations that we took this approach, however, as the possibilities for more advanced functionality by aggregators based on more rigorous data loomed large in our minds.

Dates

Dates have long been a complex issue in cultural heritage metadata. One reason for this complexity stems from the uncertainty surrounding the materials typically described by these types of institutions. With published works, both a printed copyright date and a production date may exist, or either of these may not be printed on the item but might be known from another authoritative source. Yet increasingly cultural heritage institutions are creating and sharing metadata for materials for which the date of creation is difficult to know or happened over a relatively long time span. Historic letters may be undated, a photograph collection may have been taken over a defined period of time with

little indication of when individual photographs fall in the span, or a work of art may have taken many years to complete.

The Metadata Working Group, with very little debate, agreed that the *Guidelines* should recommend supplying a date for a resource whenever possible, taking its cue from the archives and museum communities, as the presence of a date for searching and sorting was believed to be of significant importance to end-users. With this approach in mind, we also determined that the *Guidelines* should suggest that at least one date should be supplied in a machine-readable form, although we recognized that this might represent a divergence from local practice in many cases.

The goals of supplying a machine-readable date and of distinguishing between multiple dates were supported by a number of features in MODS, and the use of these features was outlined in detail in the *Guidelines*. First was the instruction that one and only one date in the record should be marked with the keyDate="yes" attribute, as a signal to aggregators that this date is the one that should be used for processing. It is this date that must be in machine-readable form, and it should represent the date mostly likely to be of interest to an end-user. In keeping with the approach taken for the content vs. carrier issue, the *Guidelines* recommend only including dates that are likely to be of use for discovery of a resource. The MODS <dateIssued>, <dateCreated>, <copyrightDate>, and <dateOther> elements were recommended for use, while the <dateCaptured>, <dateValid>, and <dateModified> elements were not recommended. (DLF Aquifer Metadata Working Group, 2006, p. 31).

The *Guidelines* recommend the use of other features of MODS in order to ensure the provision of a machine-readable date. The qualifier attribute on any of the date

elements can have a value of approximate, inferred, or questionable, removing the need for brackets, question marks, "ca." or other notations within the date value to indicate these cases, as seen in Figure 1. Date ranges in MODS do not necessarily have to be entered as a single value; rather, the *Guidelines* recommend making use of the point attribute to indicate the start and end of a date range across a pair of date elements. When these features are used, dates can then be marked as using the "w3cdtf" encoding (YYYY-MM-DD format), further indicating to the aggregator how the date value should be processed, also as seen in Figure 2. These recommendations are among those that potentially represent the greatest challenges to metadata providers, but the Working Group agreed that dates were one of the areas most deserving of improvements in metadata practice.

Figure 2. Date encoding in MODS for machine-readability

Approximate date:

```
<dateCreated encoding="w3cdtf" qualifier="approximate" keyDate="yes">1912</dateCreated>
```

instead of

```
<dateCreated>[ca. 1912]</dateCreated>
```

Date range:

```
<dateCreated encoding="w3cdtf" point="start">1912</dateCreated>
<dateCreated encoding="w3cdtf" point="end">1919</dateCreated>
```

instead of

```
<dateCreated>1912-1919</dateCreated>
```

Supplements to the *Guidelines*

While the Working Group was able to come to some degree of consensus on each of the recommendations in the *Guidelines*, our decisions to some degree represent our best educated guesses as to what Aquifer contributors can easily provide and what the Aquifer initiative would need in order to provide high-quality services to scholars. The Mellon-funded ASHO project provided for a period of formal assessment of Aquifer services following development of the project portal and local implementation scenarios. The Metadata Working Group therefore decided not to pursue ongoing revision of the *Guidelines* throughout the ASHO project period; rather we determined the best course of action was to keep the *Guidelines* stable during this time and embark on a substantive revision following the formal assessment, taking into account the results of this assessment and the experiences of Aquifer contributors using the *Guidelines*.

At the same time, we realized the Guidelines, at over 100 pages in length, were too large to serve as an efficient introduction to the issues for potential new Aquifer contributors. We therefore created two short documents to summarize and supplement the guidelines: a Levels of Adoption statement and an FAQ. These supplements are short and concise, allowing potential Aquifer contributors to get a quick view of the guidelines' requirements and recommendations, and providing the Working Group with a means of reacting to adopter comments and changes to MODS in a timely manner.

The Levels of Adoption document defines five categories of conformance to the *Guidelines*, enumerating the metadata elements that comprise each level and the end-user functionality each level supports. They therefore provide a different view of the full *Guidelines*, grouping recommendations together by what basic functions they support. The five Levels are:

1. Minimum for participation: Allows users to cite the resource
2. Minimum for doing anything useful: Allows users to perform basic searches and filtering
3. Allows more advanced functionality: Allows users to browse and group search results
4. Adopt all required guidelines (and some recommended): Allows users to perform more precise searches
5. Completely adopt all recommendations: Allows users to effectively evaluate resources

Like the full *Guidelines*, the Levels of Adoption represent the Metadata Working Group members' best guesses regarding the difficulty of adhering to guidelines and the utility of various metadata elements for specific end-user functions, informed by the members' individual experiences as metadata providers and aggregators. We expect that the ASHO project assessment activities, discussed later, will provide us with specific feedback that we can use to conduct a revision of the Levels of Adoption for use in the future.

The FAQ, as the second supplemental document to the *Guidelines*, is the most dynamic of the Metadata Working Group's documents. It was initially populated with entries covering the rationale behind the recommendations in the *Guidelines* and to link to tools created to assist potential Aquifer contributors in preparing their metadata for sharing. As we received feedback on the *Guidelines*, the Metadata Working Group added

entries for notable and recurring questions raised, generally regarding the usage of specific MODS elements.

Tools to Assist with Implementing the *Guidelines*

Given the relatively ambitious requirements set forth in the *Guidelines*, the Metadata Working Group realized that it would not be enough to provide descriptions of what metadata for Aquifer should look like, but that we also needed to provide tools to assist potential contributors in creating that metadata. These tools fall into three categories: Aquifer-specific mappings, technical tools, and software-specific procedural documentation.

Mapping: MARCXML to Aquifer MODS

As the Working Group expected a significant amount of metadata in MARC to be automatically converted to MODS for sharing with Aquifer, we took as a high priority the development of a MARC-to-MODS stylesheet customized to meet the *Guidelines* as fully as possible. We took as the starting point for our work the pre-existing MARCXML-to-MODS XSLT stylesheet from the Library of Congress.[12]

The large number of fields and subfields in MARC, the complexity of the transformations, and the need to get a useable stylesheet out quickly forced us to focus our efforts on mappings that affected instructions in the *Guidelines* marked as Required. Many mappings already met instructions in the *Guidelines*, but some required elements and attributes needed to be added or changed in the stylesheet, and in a few cases the stylesheet would benefit from the addition of a new MARC tag. As part of our work, we

also deleted some mappings to information that was not recommended for use by the *Guidelines* or not useful in the aggregated environment, such as tag 583 for processing and preservation action notes.

Due to some fundamental differences between MARC and MODS, it is sometimes possible to provide effective mappings to Aquifer-style MODS only when MARC records follow certain practices. In some cases, information needed in the MODS record does not have an exact equivalent in MARC. There is no way of specifying, for example, which URL in a MARC record should get the primary display="yes" attribute in the MODS <location>/<url> element. In these cases, our revised stylesheet makes a best guess, using data such as indicators and order of tags as clues. In other cases, our revised stylesheet does not make an attempt to provide a MODS element, especially when that data already exists in the record in another form. As MARC cataloging practices vary widely, we realized in developing the Aquifer-specific MARC-to-MODS stylesheet that it could not work perfectly for all MARC records. The mappings in our stylesheet follow whenever possible the most common MARC practice when multiple approaches exist.

Altogether, 34 changes to the Library of Congress stylesheet were made; 9 deletions, 13 added mappings (adding MARC subfield mappings or adding mappings to MODS attributes), and 23 changes.[13] The changes are documented by comments at the beginning of the stylesheet and by an online Introduction.[14] Even at this early stage, we have seen evidence that our revised MARC to MODS stylesheet is useful to Aquifer contributors. The Library of Congress, which makes MODS records for many of its online collections available for harvesting via OAI PMH, has switched from using its

original MARC-to-MODS stylesheet to using the version customized for Aquifer MODS, as the Aquifer initiative is the primary external user of the Library of Congress MODS records.

Mapping: EAD to Aquifer MODS

Archival finding aids presented a special challenge to the Metadata Working Group. Although metadata for non-digitized archival collections was considered out of scope for the ASHO phase of the Aquifer initiative, the Working Group realized that many primary source materials in DLF member institutions are described with archival finding aids. Encoded Archival Description (EAD) is the schema used widely by archivists to encode and deliver online finding aids for collections, and although EAD is an XML language, EAD finding aids are not commonly shared with OAI PMH-based metadata aggregations.

Despite the lack of short-term plans for inclusion of non-digitized archival collections in Aquifer, the Metadata Working Group believed it was important to build on the significant progress in promoting shareable metadata and to extend our scope to archival collections as well. For aggregations aimed primarily at a humanities-based research community, inclusion of archival finding aids, which often serve as pointers to research resources and frequently contain contextual information about collection materials that may help scholars and others to find those resources, is a reasonable longer-term goal. The Metadata Working Group therefore proceeded to develop a preliminary EAD-to-Aquifer-style-MODS mapping, with a two-fold rationale. First was that the Working Group members exhibited considerable expertise in this area that could

be leveraged during its time of targeted work, even if it would not be immediately useful to the ASHO project. Second was that no widely distributed EAD-to-MODS stylesheet exists. The creation of such a stylesheet could potentially provide a resource for archivists in linking up their metadata with the broader best practices of the digital library community, and encourage the inclusion of facsimiles in digital format and born-digital originals from archival holdings in shared metadata aggregations.

While inclusion of archival finding aids in resources targeted for use by humanities-based scholars seems desirable over the long term, it is unclear how transparent these finding aids, which are themselves metadata objects about collections, are to the general user of aggregated metadata services. The primary challenge to the understanding of archival collections by novice users, and to the mixing of archival and bibliographic metadata in a single system, is the nature of multi-level nature of description that is core to archival practice. Regardless of how standards-conformant a given archive is for describing digital facsimiles and born-digital resources, access and use of these materials in metadata aggregations is desirable over the long term. Much work remains to be done in this area, however. Integration of archival materials into this shared environment can be facilitated by an understanding of how the data content standards archivists use (for U.S. archivists this is typically *Describing Archives: A Content Standard*, or DACS) maps through EAD to well-articulated item-level shareable data structures like MODS or Dublin Core. The Working Group's crosswalk from EAD to the Aquifer flavor of MODS is an important intellectual tool for helping archivists to understand how their descriptive data content standards map in ways that promote shareability into standards that are promulgated within the digital library community.

This crosswalk is currently expressed as a mapping for collection-level metadata in EAD to MODS, but can serve as a model and an inspiration to archivists in mapping collection-, series-, file-, and item-level metadata in the future, as EAD makes the same set of elements available at any level of a multi-level description. The EAD-to-MODS crosswalk is currently in a draft stage, but after some review by others working in this area, the Metadata Working Group hopes to make it publicly available some time in 2008.

Technical Tools

The requirement to expose MODS records via OAI PMH proved to be a bigger barrier to participation than the Aquifer planning group originally anticipated. Few stand-alone OAI PMH data provider software packages and virtually none included as a module of a larger digital library content management system support exposure of MODS via OAI PMH out of the box. Moreover, few potential Aquifer contributors were able to devote development resources to making this improvement, even when it was technically possible within their local system. With the urging of the Metadata Working Group, the ASHO development team was able to set up an OAI PMH Static Repository Gateway[15], with which metadata contributors could register XML files containing MODS records. In this scenario, contributors would export metadata from their local systems in its native format, convert this metadata to MODS following the *Guidelines* to the greatest degree feasible, ensure the output conforms with the OAI PMH Static Repository specification, and place the resulting file into a stable, web-based location for harvesting. With this additional method of creating MODS records for contribution to Aquifer, by August

2008, four additional repositories were able to contribute metadata during the ASHO project phase.

A second major tool to assist contributors in creating MODS records conforming to the *Guidelines* is still in the development phase: a web-based record analysis tool. This tool will provide contributors with a report outlining which requirements and recommendations from the *Guidelines* a MODS record meets, and which Level of Adoption the record as a whole falls into. While baseline conformance to the MODS Schema is testable with an XML validator, simply creating a valid MODS record is not enough to provide the advanced services envisioned for Aquifer-developed metadata aggregations. The Metadata Working Group is in the final stages of developing checks for each of the MODS Guidelines in the Schematron XML assertion language[16]. The Schematron language is designed to provide for validation of an XML document to a specified set of requirements, above and beyond what a DTD or XML Schema defines. We are creating a Schematron rule for each MODS guideline, and these rules are then grouped by Level of Adoption. Most rules simply check for the presence or absence of an element, but a few use more advanced tests such as restriction of a value to a defined list beyond those enforced by the MODS Schema or only allowing one instance of a repeatable element. The Schematron rules will then serve as the basis of a web-based validation service that shows, for each Level of Adoption, which guidelines have not been met by a tested MODS record, and which level of obligation (Required, Recommended, etc.) the guidelines that have not been met represent. This service should provide a useful tool for current and potential contributors to quickly analyze how closely their current records match the *Guidelines* and to assess the effort involved should they

choose to devote resources to adapting that metadata. It could also be useful in the planning stages for new projects, as metadata specifications are developed.

Software-specific Procedural Documentation: CONTENTdm

In addition to tools to assist with the analysis and sharing of metadata records as separate entities, the Metadata Working Group realized it would be beneficial to also provide assistance in the form of procedural documentation for specific, commonly-used software packages used for digital library objects. The obvious first choice for this approach was CONTENTdm[17], as it is widely used in this sector, and although it can share simple Dublin Core and qualified Dublin Core records via OAI PMH, it does not have the capability to expose MODS records for harvesting.

Expertise in the Metadata Working Group is particularly strong in the areas of working with legacy data and the transformation from one XML format to another. The Working Group is leveraging this expertise in creating documentation and tools to assist CONTENTdm users in producing Aquifer-style MODS from native and even customized CONTENTdm metadata, which is generally a form of qualified Dublin Core. This documentation, currently under development, consists of instructions for selecting an appropriate export format for the metadata, a basic stylesheet for transforming the standard CONTENTdm elements into MODS, and instructions for adding features to the stylesheet to capture local customizations to the core CONTENTdm metadata. We expect to make this procedural documentation publicly available by the end of summer 2008.

Future for Aquifer metadata activities

While the ongoing development of tools to assist institutions in becoming Aquifer participants promises to take a great deal of the Metadata Working Group's time in the near future, the Working Group will have other responsibilities as well. Now that the ASHO portal is available to scholars and each of the local implementation scenarios are in place, the Aquifer team has begun a series of formal assessment activities, led by the Services Working Group. Each of the methods provided by Aquifer for interacting with the ASHO materials will be evaluated as to its effectiveness in meeting scholars' needs. A variety of methodologies will be employed, ranging from surveys and focus groups to task-based testing and user observation.

The Metadata Working Group expects to learn a great deal from the Aquifer ASHO assessment activities, as the quality, structure, and functions implemented on top of metadata in online collections can have a significant effect on their usefulness. The Working Group hopes to use results from assessment activities to inform decisions throughout the metadata lifecycle, from initial creation to sharing with Aquifer and indexing and display decisions for each of the end-user access mechanisms.

Assessment activities are expected to inform a revision of the MODS *Guidelines*. The writing of the *Guidelines* involved a great many decisions about how metadata should be structured in order to best meet scholars' needs for use of resources. The Working Group made these decisions based on individual members' experience with metadata in a shared environment, both as metadata providers and as metadata aggregators, and by applying general principles of shareable metadata specifically to MODS. We believe that the *Guidelines* that emerged from this environment largely meet their goals or supporting higher-level aggregation services. However, the *Guidelines* are

at present lacking external validation as to their effectiveness, having been by necessity written before functional requirements for the metadata portal or any of the local implementation scenarios were designed, and before any user requirements gathering had been performed. ASHO assessment activities could provide both this validation and information on areas that deserve revisiting, by demonstrating the functionality on which scholars most depend and the degree to which the metadata in the system supports this functionality. It will also be important to assess the degree to which relevant resources were unavailable to scholars because inadequate metadata made them poorly discoverable. The Metadata Working Group plans to analyze the assessment findings from the portal and from each of the local implementation scenarios to learn more about the effectiveness of various recommendations made by the MODS Guidelines and subsequent Levels of Adoption framework. We hope to adjust the Guidelines near end of the ASHO project in March 2009 to take what we have learned from the assessment activities into account.

A more recent focus of the Metadata Working Group has been the design of "data processing" rules for turning harvested metadata into effective formats for search and browse. These rules were developed based on expectations for harvested MODS records implementing recommendations in the *Guidelines*, and through an examination of actual harvested data using XQuery technology. Developing the data processing rules required making difficult decisions regarding converting pre-coordinated subject headings into segments suitable for faceted browsing, selecting major access points for browse and search indexes, and normalizing uncertain dates and date ranges for indexing and sorting. As in the creation of the MODS *Guidelines*, each of these decisions was made with the

benefit of the hands-on system-building experience of Working Group members and would benefit from validation in a production environment. Assessment of the ASHO portal and each of the local implementation scenarios will indicate which of the data processing rules are effective and which are in need of revision.

Aquifer, as a Digital Library Federation initiative, is scheduled to end in March 2009, concurrent with the end of the Mellon-funded ASHO project phase. The Metadata Working Group expects to complete all planned tool development and a revision of the *Guidelines* by that time, as well as completing the oversight of a shorter-term project funded by the Gladys Krieble Delmas Foundation to evaluate various metadata remediation and enhancement techniques. Each of these activities should result in openly-available documentation that can be used and built upon by future initiatives.

Metadata harvesting initiatives are entering a new era. Simple indexing and normalization of simple Dublin Core records is not likely to meet the ever-increasing demands of our users. The Aquifer ASHO project, by leveraging the advanced capabilities of Digital Library Federation members to harvest MODS records prepared with sharing in mind, has made progress in building new frameworks and expectations for aggregation services. The iterative nature of the project, incorporating metadata planning services together with assessment of end-user services, is designed to contribute in a significant way to the state of the art in digital libraries. The lessons learned as part of the Aquifer initiative should inform future services from the Digital Library Federation and elsewhere.

References

Arms, William Y., Dushay, Naomi, Fulker, Dave, & Lagoze, Carl. (2003). A case study in metadata harvesting: the NSDL. *Library Hi Tech*, 21(1), 228-237.

Bradner, Scott. (March 1997). *Key words for use in RFCs to indicate requirement levels.* BCP 14, RFC 2119. Retrieved June 7, 2008, from http://www.ietf.org/rfc/rfc2119.txt

Dempsey, Lorcan. (June 24, 2005). *In the Flow*. Blog post at Lorcan Dempsey's weblog. Retrieved June 7, 2008, from http://orweblog.oclc.org/archives/000688.html

DLF Aquifer development for interoperability across scholarly repositories: American Social History Online. Executive summary. (n.d.). Retrieved June 7, 2008, from http://www.diglib.org/aquifer/ASHOvisionExSum.pdf

DLF Aquifer Metadata Working Group. (November 2006). *Digital Library Federation / Aquifer implementation guidelines for shareable MODS records.* Version 1.0. Retrieved June 7, 2008, from http://wiki.dlib.indiana.edu/confluence/download/attachments/24288/DLFMODS_ImplementationGuidelines_Version1-2.pdf?version=1

DLF/NSDL Working Group on OAI PMH Best Practices. (October 2007). *Best practices for OAI PMH data provider implementations and shareable metadata*. (Sarah L. Shreeves, Jenn Riley, & Kat Hagedorn, Eds.) Washington, D.C.: Digital Library Federation.

Hagedorn, Kat. (2003). OAIster: a 'no dead ends' OAI service provider. *Library Hi Tech*, 21(2), 170-181.

Halbert, Martin. (2003). The Metascholar Initiative: AmericanSouth.Org and MetaArchive.Org. *Library Hi Tech* 21(2), 182-198.

Hillmann, Diane. (2005). *Using Dublin Core*. Retrieved June 7, 2008, from http://www.dublincore.org/documents/usageguide/

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1997, with amendments in 2008). *Functional Requirements for Bibliographic Records*. Munich: K.G. Saur. Retrieved July 2, 2008, from http://www.ifla.org/VII/s13/frbr/frbr_2008.pdf

Kott, Katherine. (May/June 2005). DLF Aquifer: tapping new sources for scholarship. *CLIR Issues*, 45. Retrieved June 7, 2008, from http://www.clir.org/pubs/issues/issues45.html#dlf

Lagoze, Carl & Van de Sompel, Herbert. (2003). The making of the Open Archives Initiative Protocol for Metadata Harvesting. *Library Hi Tech*, 21(2), 118-128.

Lagoze, Carl, Krafft, Dean, Cornwell, Tim, Dushay, Naomi, Eckstrom, Dean, & Saylor, John. (2006). Metadata aggregation and "automated digital libraries": a retrospective on the NSDL experience." In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2006), June 11-15, 2006* (pp. 230-239). Chapel Hill, NC, New York, NY: ACM.

Palmer, Carole L. & Knutson, Ellen M. (2004). Metadata practices and implications for federated collections. *Proceedings of the American Society for Information Science and Technology*, 41(1), 456-462.

Powell, Andy, Nilsson, Mikael, Naeve, Ambjörn, Johnston, Pete, & Baker, Thomas. (2007). *DCMI abstract model*. Retrieved June 7, 2008, from http://www.dublincore.org/documents/abstract-model/

Shreeves, Sarah L., Knutson, Ellen M., Stvilia, Besiki , Palmer, Carole L., Twidale, Michael B., & Cole, Timothy W. (2005). Is "quality" metadata "shareable" metadata? The implications of local metadata practice on federated collections. In Hugh A. Thompson (Ed.), *Proceedings of the twelfth national conference of the Association of College and Research Libraries, Minneapolis, MN, April 7-10 2005* (pp. 223-237). Chicago: Association of College and Research Libraries.

Sarah L. Shreeves, Riley, Jenn, & Milewicz, Liz. (August 2006). Moving towards shareable metadata. *First Monday,* 11(8). Retrieved June 7, 2008, from http://firstmonday.org/issues/issue11_8/shreeves/index.html

Shreeves, Sarah L., Kaczmarek, Joanne, & Cole, Timothy W. (2003). Harvesting cultural heritage metadata using the OAI protocol. *Library Hi Tech*, 21(2), 159 - 169.

Wendler, Robin. (2004). The eye of the beholder: challenges of image description and access at Harvard. In Diane Hillmann & Elaine Westbrooks (Eds.), *Metadata in practice* (pp. 51-69). Chicago: ALA Editions.

Appendix A. Aquifer Metadata Working Group Membership

Jenn Riley (Indiana University): 2005-present; Chair, 2007-present

Sarah L. Shreeves (University of Illinois at Urbana-Champaign): 2005-present; Chair, 2005-2007

Laura Akerman (Emory University): 2006-present

John Chapman (University of Minnesota): 2005-present

Melanie Feltner-Reichert (University of Tennessee): 2006-present

Kat Hagedorn (University of Michigan): 2007-present; ASHO Core Team Liaison, 2006

Bill Landis (California Digital Library): 2005-2006, 2007-present

Tracy Meehleib (Library of Congress): 2006-present

Elizabeth Milewicz (Emory University): 2005-2006

David Reynolds (Johns Hopkins University): 2005-present

Gary Shawver (New York University): 2005-present

---

[1] http://www.diglib.org
[2] http://wiki.dlib.indiana.edu/confluence/x/4F4
[3] http://www.dlfaquifer.org
[4] http://www.openarchives.org/pmh/
[5] These OAI PMH aggregations or project reports can be found at: Emory - http://www.metascholar.org/index.html; UIUC - http://oai.grainger.uiuc.edu/ and http://imlsdcc.grainger.uiuc.edu/; University of Michigan - http://www.oaister.org/; California Digital Library - http://www.cdlib.org/inside/projects/amwest/.
[6] The authors would like to recognize that with the development of the Dublin Core Abstract Model there is a greater potential for expressing semantically rich records using the Dublin Core element set.
[7] A list of Aquifer Metadata Working Group members may be found in Appendix A.
[8] A summary of metadata formats in use by OAI PMH data providers can be found at the University of Illinois OAI-PMH Data Provider Registry, Distinct Metadata Schemas report, http://gita.grainger.uiuc.edu/registry/ListSchemas.asp.
[9] http://www.loc.gov/rr/print/tgm2/. It should also be noted that in October 2007 the TGM I (subject terms) and TGM II were merged into a single vocabulary, although separate vocabulary files are still available.
[10] http://memory.loc.gov/ammem/techdocs/genre.html
[11] http://www.loc.gov/rr/mopic/miggen.html
[12] http://www.loc.gov/standards/mods/v3/MARC21slim2MODS.xsl
[13] The current version of the MARCXML to Aquifer MODS stylesheet can be found at http://wiki.dlib.indiana.edu/confluence/x/K4AQ
[14] http://wiki.dlib.indiana.edu/confluence/x/MYAQ
[15] http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm
[16] http://www.schematron.com/
[17] http://www.contentdm.com/