

Metadata Requirements for Acme University's Proposed Institutional Repository

April Younglove
ayounglo@emporia.edu
School of Library and Information Management
Emporia University
May 2007

Objective:

To create a recommendation for a single metadata standard for Acme University's proposed Institutional Repository (IR). This recommendation will be used to aid in the creation of a Request For Proposal (RFP) for the IR project.

Background:

Acme University's Institutional Repository will require mandatory self-archiving by faculty. Thus, it will include work from many disciplines, possibly in many formats. Faculty will be the primary creators of metadata in the repository, and will need training in how to create this metadata. The repository will include pre-prints and post-prints, and the IR project must work with the objectives of the University's existing [digital preservation policy](#).

Digital preservation policy objective:

"Acme's preservation role is guided by its key objective to preserve and maintain all Acme and significant non-Acme intellectual property to ensure they are available for current and future educational use. This objective applies to both digital and non-digital information resources, although the Library recognizes that it will use different methods and draw on different skills, procedures and partnerships, for managing digital and non-digital collections."¹

Initially, the repository plans to include text and images, and in the long term it intends to include audio and video formats as well. In the future, the University may also develop a comprehensive Digital Access Management (DAM) system that could include student exams, school literature, financial records etc. Acme's IR should be open to working with a future DAM system and to using its repository for materials that are not strictly e-prints or faculty research (e.g. it could include student theses, classroom projects, or other creative and scholarly work generated by the school).

Technical Service goals for Acme's IR:

¹ A full copy of the policy is available at: <http://www.acme.edu/policy/digpres.html>

- planning for long term preservation of digital objects
- providing for perpetual access and support (sustainability)
- creating highest quality metadata (enables indexing and enhances search results)
- promoting access and collaboration through interoperability
- enabling staff to easily contribute digital objects and metadata
- promoting access to staff work with maximum “Google-ability” and straightforward linking between staff pages and IR deposits
- maintaining security of content (only authorized users may add or edit content)
- allowing for future change and extensibility

These goals are intended to reflect the National Information Standards Organization (NISO) Framework for metadata²:

1. Appropriateness to described resource. Good metadata should be appropriate to the materials in the collection, users of the collection, and intended, current and likely use of the digital object.
2. Interoperability. Good metadata supports interoperability.
3. Vocabularies. Good metadata uses standard controlled vocabularies to reflect the what, where, when and who of the content.
4. Use Terms. Good metadata includes a clear statement on the conditions and terms of use for the digital object.
5. Authenticity/Persistence. Good metadata records are objects themselves and therefore should have the qualities of good objects, including "archivability," persistence, unique identification, etc. Good metadata should be authoritative and verifiable.
6. Object management. Good metadata supports the long-term management of objects in collections.³

Recommendation:

Based on the specifications given to me by the University and on our technical service goals for Acme’s IR, I recommend the use of Metadata Object Description Schema (MODS).

Considerations:

The following metadata formats were suggested for by different departments within Acme University for my consideration: qualified and unqualified Dublin Core (DC), Text Encoding Initiative (TEI), Global Information Locator Service (GILS), ONline Information eXchange (ONIX), Metadata Object Description Schema (MODS) and (MARCXML). We decided to review only standards that

² NISO. A Framework of Guidance for Building Good Digital Collections.
<http://www.niso.org/framework/Framework2.html#metadata>

³ University of Illinois at Urbana-Champaign. Digital Content Creation Team Metadata Guide.
<http://www.library.uiuc.edu/digproj/dcct/meta.php>

are XML based, because, “XML has become the de-facto standard for representing metadata descriptions of resources on the Internet.”⁴

Briefly,

- **DC:** (<http://dublincore.org/>) the simplest and one of the oldest metadata formats, also called the “lingua franca for metadata” by the Open Archive Initiative’s (OAI) website.⁵ Only unqualified DC (that is, DC without subfields) is OAI compliant.
- **TEI:** (<http://www.tei-c.org/>) intended for the markup of texts. Heavy on the structural metadata, light on the descriptive metadata.
- **GILS:** (<http://www.gils.net/about.html>) largely used by land surveyors, satellite companies, environmental libraries etc. to create locator records⁶
- **ONIX:** (<http://www.editeur.org/>) standard metadata format for book sellers. Designed to represent physical objects such as books, DVDs etc., but not born-digital information (ie. a .wav file).
- **MODS:** (<http://www.loc.gov/mods>) works with MARC (Machine Readable Catalog – the standard library records format), developed by the library of Congress and MARC -- MODS is a bibliographic element set heavily based on MARC, and may be regarded as a somewhat simplified and rationalized version of it. MODS is more easily mapped to other standards than MARCXML.
- **MARCXML:** (<http://www.loc.gov/standards/marcxml/>) basically, whole MARC records in XML format. Preferred by highly trained librarians, very complex for beginners and non-catalogers. Limited in many of the same ways that MARC itself is limited.

Some questions that I asked myself when looking at each format included: How would we use this format in our IR? Will metadata in this format be easily harvestable by major search engines and how much information will be lost if crosswalking is required? How difficult would it be to train faculty to use this format? What other institutions use this format for IR? Is this format archival?

Interoperability

In some cases, it became immediately clear to me that certain metadata formats would not be useful for our repository. They lack the NISO requirement of being the most appropriate format for the described resource. While some of these metadata formats might work well with one discipline’s content, those same formats might work poorly with the needs of other departments. For instance, TEI might be the ideal structure for the literature department, but might not be very

⁴ Hunter, J. (2003). Working towards MetaUtopia: A Survey of Current Metadata Research. *Library Trends*, 52 (2), 318-344. <http://espace.library.uq.edu.au/view.php?pid=UQ:7871>

⁵ OAI-PMH Implementation Guidelines - Guidelines for Repository Implementers -- Dublin Core and Other Metadata Formats. <http://www.openarchives.org/OAI/2.0/guidelines-repository.htm#MinimalImplementation-DC>

⁶ Caplan, P. (2003). *Metadata Fundamentals for All Librarians*. Chicago: ALA. pp.110-115.

useful for the music or art departments since it is not designed to describe objects outside of text. Because our repository will include both textual and visual elements and intends to include auditory and other materials in the future, any of our chosen standards must easily be able to accommodate all of these media types. With TEI, for instance, “The set of metadata elements it includes to describe the digital manifestation of the text is deemphasized over the structural and seamantic [sic] markup of the resource content. As a result, in most cases it will be necessary to prepare descriptive records in other formats (such as MARC) to allow for resource discovery and aggregation.”⁷ It is a disadvantage to have a format that does not easily translate to become web readable.

The following rubric shows how that each of the above standards would work with external aggregation and packaging technologies that make IR data available to search engines:

	SRU⁸	OAI-PMH⁹	METS¹⁰
Unqualified Dublin Core	Yes	Yes	Yes
TEI	No – must be crosswalked to MARCXML	With crosswalking to unqualified DC	Yes
GILS	No – must be crosswalked to MARCXML	With crosswalking to unqualified DC	Yes
MODS	Yes	With crosswalking to unqualified DC	Yes
ONIX	No – must be crosswalked to MARCXML	With crosswalking to unqualified DC	Yes
MARC21XML	Default – if no standard specified, records will be crosswalked into MARCXML	With crosswalking to unqualified DC	Yes
Qualified Dublin Core	Yes	With crosswalking to unqualified DC	Yes

⁷ University of Illinois at Urbana-Champaign. Digital Content Creation Team Metadata Guide: TEI. <http://www.library.uiuc.edu/digproj/dcct/meta.php>

⁸ A list of SRU supported metadata types is here: <http://z3950.loc.gov:7090/voyager> Explanation of default types and supported types here: <http://www.loc.gov/standards/sru/simple.html>

⁹ OAI-PMH Implementation Guidelines - Guidelines for Repository Implementers -- Dublin Core and Other Metadata Formats. <http://www.openarchives.org/OAI/2.0/guidelines-repository.htm#MinimalImplementation-DC>

¹⁰ Under MD type required, METS currently lists “other,” indicating that they can accept any standard: http://ark-dev.cdlib.org:8086/mets-samp/schema_test/version16/mets.v1-6#mdWrap

In addition to accommodating different internal needs, our chosen metadata format must also interoperate well with external data harvesters. Search/Retrieve via URL (SRU) and the Open Archives Initiatives (OAI) are open access aggregation resources for library metadata. They help take library metadata out of catalogs where it is invisible to outside search engines, and either summarize or provide direct access to the data. By building on the search and retrieve protocol of the Z39.50 standard, SRU points web browsers to URLs where a resource or information about a resource can be gathered and made searchable.¹¹ SRU is still a fairly new aggregator, but is rapidly growing in popularity and support. Another data harvesting tool, the most widely used and referenced, OAI, “helps provide access to multiple digital collections by providing a framework that allows libraries and other service providers to harvest metadata about collections according to rules specified in the Protocol for Metadata Harvesting (PMH).”¹² The difference between SRU and OAI is that OAI “is not designed to allow users to interactively search for items that meet a specific information need. Rather, it only provides a simple framework that allows organizations to download batches of records from an archive.”¹³ It is important that data in our repository be available to these aggregation resources if we want our institution and faculty work to be searchable by engines like Google.

While METS is not truly an aggregation technology, it is a highly extensible open standard that enables search engines to access otherwise inaccessible data. It can either hold its own internal descriptive metadata, or point within its records to external data created in other standards like DC, MODS, and MARC.¹⁴ Librarian Karen Coyle describes METS as being like a digital binding and cover for the computerized files that make up a virtual book. Within this binding, METS has a sort of copyright page that provides the technical data about “file formats, the technology used in scanning if the item began its life on paper, and the digital transformations and compression that have been used on the files.”¹⁵ METS consists of six elements that contain or point to the digital object’s metadata: a header, descriptive metadata, administrative metadata, file section, structural map and behavior section. With the exception of the header and structural map, all elements are optional and can be stored externally in any format.

TEI, GILS, and ONIX are too specialized and not harvestable enough for our needs. At this point, only MARCXML, MODS and qualified or unqualified Dublin Core will be considered.

Crosswalking

¹¹ Dahl, M., Banerjee, K. & Spalti, M. (2006). *Digital Libraries: Integrating Content and Systems*. pp. 41-44. Oxford: Chandos Publishing.

¹² Ibid. p.44

¹³ Ibid. p.45

¹⁴ Tennant, R. (2002, April 15). Digital libraries – Metadata as if libraries depended on it. *Library Journal*, np. <http://www.libraryjournal.com/article/CA206408.html>

¹⁵ Coyle, Karen. (2005). Understanding Metadata and its purpose. *Journal of Academic Librarianship*, 31(2), 160-163. Preprint. http://www.kcoyle.net/jal2_Metadata.html

Clearly, formats that might interoperate well with some outside parties but not with others are not optimal. ONIX, for instance, can map to MARCXML and Dublin Core, and would clearly interoperate well with online book traders.¹⁶ However, its need to crosswalk in order to be used by either SRU or OAI is a drawback. Every time a metadata record is crosswalked, part of the record's granularity is lost -- the more different the initial standard from the standard it is being crosswalked too, the more the data loss.¹⁷

Ideally, one would chose the most lossless format – that is, the format that is supported by the most amount of content harvesters and that needs the least amount of damaging crosswalking. In this case, it would appear from the above rubric that unqualified Dublin Core would be the obvious choice. However, in today's complex web environment, one can expect that one's metadata will at some point need to be crosswalked. If one's standard is not data rich to begin with, than it is far worse to make that standard even less rich than it is to begin with a rich standard and then translate that into a simpler format.¹⁸ Unqualified DC may be the “lingua franca of the metadata world” and may not lose much content in the process of harvesting, but DC records tend not have that much substance to begin with. DC is the simplest DTD with only 15 fixed fields and almost no defined rules for content. This is good because it means that even non-catalogers can quickly organize metadata by using it. However, this same simplicity can be its biggest liability. Even OAI suggests adopting an internal standard other than unqualified DC.¹⁹ Therefore, while realizing that at some point our metadata must be translated into unqualified DC, as an institution we should instead look at implementing MODS, MARCXML or Qualified Dublin Core as our metadata standard.

Staff and faculty buy-in

Because our standard will most likely be used by both librarians and faculty, it is important to choose a metadata standard that is robust enough to accommodate both populations. Our library already uses MARC records internally (as do almost all libraries). Consequently, our catalogers would most likely favor a format that maintains as much of their MARC data as possible. Translating records straight from MARC into DC for sharing over the web would automatically require the library to sacrifice most of the granularity that it has already created. For instance, Branschofsky et al (2003) describes how that staff at MIT realized that

¹⁶ According to MIT's Metadata Reference Guide, all of the major adapters of ONIX so far have been from the retail book industry:

<http://libraries.mit.edu/guides/subjects/metadata/standards/onix.html>

¹⁷ Tennant, R. (2002). The importance of being granular. *Library Journal*, 127(9), 32-4.

¹⁸ Dahl, M., Banerjee, K. & Spalti, M. (2006). *Digital Libraries: Integrating Content and Systems*. pp. 128. Oxford: Chandos Publishing.

¹⁹ According to OAI's website (<http://www.openarchives.org/OAI/2.0/guidelines-eprints.htm>): “Many repositories store their metadata in some other format, and dynamically convert to DC in response to harvester requests . . . Although repositories are strongly encouraged to expose richer, possibly community-specific metadata formats, there is no requirement to do so.”

they would need to go back and incorporate MARC data into their Institutional Repository, “In the initial stages of the project a qualified form of the Dublin Core metadata set was developed for use within the system. Subsequently it became evident that metadata would also have to be imported from existing MARC sources for batch loads of scanned digital items.”²⁰ Other formats, like MARCXML and MODS, can more accurately translate MARC data into XML than DC.²¹ Another reason to consider what librarians would most prefer is that even though they will not be generating the initial data going into our IR (thesis and faculty writings), they may be called upon to instruct and support the faculty in using the IR. In addition, in the future librarians may either take over some of the duties of the DAM system or wish to expand the capabilities of their catalog to include information objects, that is, both the metadata about a described resource as well as the resource itself.²²

Having an extensible standard means choosing a standard that is user centered, not collection centered.²³ In this case, we need to recall that when faculty deposit documents into the IR, they will be self-archiving, and this means that we do not want to implement a standard that is as arcane and as non-intuitive as MARCXML would be. As Goldsmith and Knudson (2006) point out in their article “Repository librarian and the next crusade,” “Despite the limited naming conventions, the sheer number of tag/indicator/subfield combinations further suggested that the complexity of this standard [MARCXML] could be problematic.”²⁴

Our goal is to make the process of adding metadata as quick and painless for faculty members as possible. As cataloger Karen Coyle says, “it is not the object itself that determines the metadata but the needs and purposes of the people who create it and those who it will serve.”²⁵ Ideally, faculty would not have to directly interact with XML tags, but would be able to fill out a form with simple prompts for title and subject information and drop down options to control vocabulary (in MODS, for instance, one could pre-determine item types to choose from: Cartographic, Still Image etc.). However, a certain amount of simplicity is required for even this strategy to work. Even though using the IR will

²⁰ Branchofsky, M. et al. (2003). Evolving Metadata Needs for an Institutional Repository: MIT’s DSpace. 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice – Metadata Research & Applications. <http://dc2003.ischool.washington.edu/Archive-03/03branchofsky.pdf>

²¹ Coyle, Karen. (2005). Understanding Metadata and its purpose. *Journal of Academic Librarianship*, 31(2), 160-163. Preprint. http://www.kcoyle.net/jal2_Metadata.html

²² CCSDS Recommendation for an OASIS Reference Model. Blue Book, Issue 1. January 2002. <http://public.ccsds.org/publications/archive/650x0b1.pdf>

²³ Dahl, M., Banerjee, K. & Spalti, M. (2006). *Digital Libraries: Integrating Content and Systems*. pp. 95-133. Oxford: Chandos Publishing.

²⁴ Goldsmith, B. and Knudson, F. (September 2006). Repository Librarian and the Next Crusade: The Search for a Common Standard for Digital Repository Metadata. *D-Lib Magazine*. 12 (9) np. <http://www.dlib.org/dlib/september06/goldsmith/09goldsmith.html>

²⁵ Coyle, Karen. (2005). Understanding Metadata and its purpose [preprint]. *Journal of Academic Librarianship*, 31(2), 160-163. http://www.kcoyle.net/jal2_Metadata.html

be mandatory, it is important to generate support from faculty for its use. An IR that is easy to use and well supported saves time and money down the road because it will require less training, will eliminate the need to constantly prompt faculty to participate, and will take up less time for them to use.²⁶ As Amanda Wilson (2007) writes in her article “Towards Releasing the Metadata Bottleneck: a Baseline Evaluation of Contributor-Supplied Metadata,” “Metadata creation is one of the most expensive components of digital projects. Organizational expertise (the correct structure, syntax, and use of metadata elements) and the subject expertise (the appropriate semantic description of a resource’s content for users) are both needed to create a high-quality metadata record.”²⁷ Wilson concludes that faculty can, in fact, create excellent quality metadata, but suggests that this is most possible when the faculty is provided with easy to use electronic forms.

Industry standards

Because we need a standard that fits the needs of both our faculty and our library staff, MARCXML will be too complex and not friendly enough to faculty. Therefore, it will no longer be considered. In examining the remaining standards, Dublin Core Qualified and MODS, I decided that one way to evaluate them is based on their prevalence within the field of IR metadata. Determining what the industry standard is for academic institutional repositories will help us choose a format that will not become obsolete in a few years. It will also help us collaborate with other institutions if we encounter problems or if we choose to embark on joint projects with others.

In order to evaluate the prevalence of a particular metadata format, I first located several comprehensive lists of IRs.²⁸ Then I eliminated IRs that were not similar to ours (ie. they were not-academic) and then did a random sample of every 10th repository to see what kind of metadata it used. Unfortunately, not all institutions choose to use the same acronyms for the same standards and many do not appear to distinguish between unqualified Dublin Core and qualified Dublin Core (qualified DC enables users to break data down into smaller chunks and to be more specific than unqualified DC does). This was problematic in that unqualified DC appears to be universal for academic institutions that wish to have data harvested by OAI.

²⁶ Waters, M. R. (2004). OAI Standards, LEarning About Digital Institutional Repositories -- Creating an Institutional Repository: LEADIRS Workbook. *MIT Libraries*.
<http://www.openarchives.org/OAI/2.0/guidelines-repository.htm>

²⁷ Wilson, A. (2007). Towards Releasing the Metadata Bottleneck: a Baseline Evaluation of Contributor-Supplied Metadata. *Libr Resour Tech Serv* 51(1): 16-28.

²⁸Registry of Open Access Repositories (ROAR) <http://archives.eprints.org/> and The University of Illinois OAI-PMH Data Provider Registry at:
<http://gita.grainger.uiuc.edu/registry/ListRepolds.asp?self=1>

Many institutions used multiple standards. Among the most common (after DC) are: variants of DC (such as qualified DC, OAI DC and NSDL DC), MODS, MARCXML, rfc1807, and etdms (a metadata standard for theses). Some libraries that have adopted MODS for at least a major chunk of sets in their IR included the Indiana University Digital Library Program <http://oai.dlib.indiana.edu/phpoai/oai2.php> and the University of Chicago Library Metadata Repository <http://oai.lib.uchicago.edu/>. Although use of MODS certainly was not as widespread as DC, its use increased in IRs that were established more recently. In addition, it was typically noted that use of MODS was paired with METS.

Conclusion:

In the end, I decided upon MODS because it strikes a good balance between a complex MARC standard like MARXML and a standard that is too general, like DC. MODS, “is supported by the Library of Congress and is expressed as an XML schema, using language-based tags instead of MARC's numerical ones . . . MODS retains MARC's support for a wide range of controlled vocabularies.” IT also “retains MARC's provisions for recording unique identifiers, datestamps, level of verification, and institutional origin. MODS was created partly in response to a need for a version of MARC that was better suited for use in digital library environments. It has potential for use as a Z39.50 Next Generation format and as a METS extension schema.”²⁹ Although both Qualified Dublin Core and MODS are human readable, which is important for archival reasons (so that data can be repackaged in the future should the need arise),³⁰ Dublin Core lacks the robustness of MODS. MODS has better structural as well as descriptive capabilities.³¹

As Andy Powell (2005) of Bath University writes, the most common uses of DC in IRs “leads to problems for the consumers of metadata from those systems . . . it is difficult or impossible in many cases to reliably tie identifiers and metadata records to individual ‘manifestations’ of eprints (i.e. different formats), largely because of the widely varying practices across institutions. In practice this means that it is often difficult for software robots to move reliably from the harvested metadata record to the full-text of an eprint.”³² Although qualified DC includes elements that help refine its data, these elements do not have any specific rules to control the varying practices Powell describes. A content creator may list an author’s first name first or last name first, or any other way he or she sees fit. Even the use of Qualified Dublin Core does not solve this problem. Unless a

²⁹ University of Illinois at Urbana-Champaign. Digital Content Creation Team Metadata Guide. <http://www.library.uiuc.edu/digproj/dcct/meta.php>

³⁰ CCSDS Recommendation for an OASIS Reference Model. Blue Book, Issue 1. January 2002. <http://public.ccsds.org/publications/archive/650x0b1.pdf>

³¹ Ibid.

³² Powell, A. (2005). Notes about possible technical criteria for evaluating institutional repository (IR) software. *University of Bath*. <http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/ir-software.pdf>

content creator is using the Dublin Core Best Practices guidelines,³³ the content will be difficult to crosswalk.

Even in the best case scenario, DC lacks flexibility when it comes to newer digital standards. Eprints, for instance, are so new that they do not have a place in DC's Best Practices guidelines yet. Unlike DC, MODS allows for rich description of complex digital objects and its tags, or elements, are particularly applicable to digital resources. In addition, it also works well with hierarchical METS objects and is compatible with open source softwares like Fedora, Dspace, and GNUprints.³⁴ Before closing, I would like to reiterate the technical services of Acme's proposed IR to ensure that MODS meets the criteria for a metadata standard in our repository. These goals are:

- planning for long term preservation of digital objects
- providing for perpetual access and support (sustainability)
- creating highest quality metadata (enables indexing and enhances search results)
- promoting access and collaboration through interoperability
- enabling staff to easily contribute digital objects and metadata
- promoting access to staff work with maximum "Google-ability" and straightforward linking between staff pages and IR deposits
- maintaining security of content (only authorized users may add or edit content)
- allowing for future change and extensibility

Because MODS is granular, complex, sustainable and extensible, I believe it will best allow us to meet those goals.

³³ CDP Metadata Working Group. (2006). Dublin Core Metadata Best Practices Version 2.1.1 <http://www.cdpheritage.org/cdp/documents/CDPDCMBP.pdf>

³⁴ <http://www.openarchives.org/tools/tools.html>