Dublin Core, DSpace, and a Brief Analysis of Three University Repositories

Mary Kurtz

This paper provides an overview of Dublin Core (DC) and DSpace together with an examination of the institutional repositories of three public research universities. The universities all use DC and DSpace to create and manage their repositories. I drew a sampling of records from each repository and examined them for metadata quality using the criteria of completeness, accuracy, and consistency. I also examined the quality of records with reference to the methods of educating repository users. One repository used librarians to oversee the archiving process, while the other two employed two different strategies as part of the self-archiving process. The librarian-overseen archive had the most complete and accurate records for DSpace entries.

The last quarter of the twentieth century has seen the birth, evolution, and explosive proliferation of a bewildering variety of new data types and formats. Digital text and images, audio and video files, spreadsheets, websites, interactive databases, RSS feeds, streaming live video, computer programs, and macros are merely a few examples of the kinds of data that can be now found on the Web and elsewhere. These new dataforms do not always conform to conventional cataloging formats. In an attempt to bring some sort of order from chaos, the concept of metadata (literally "data about data") arose. Metadata is, according to ALA, "structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities."[1]

Metadata is an attempt to capture the contextual information surrounding a datum. The enriching contextual information assists the data user to understand how to use the original datum. Metadata also attempts to bridge the semantic gap between machine users of data and human users of the same data.

Dublin Core

Dublin Core (DC) is a metadata schema that arose from an invitational workshop sponsored by the Online Computer Library Center (OCLC) in 1995. "Dublin" refers to the location of this original meeting in Dublin, Ohio, and "Core" refers to that fact DC is set of metadata elements that are basic, but expandable.

DC draws upon concepts from many disciplines, including librarianship, computer science, and archival preservation.

The standards and definitions of the DC element sets have been developed and refined by the Dublin Core Metadata Initiative (DCMI) with an eye to interoperability. DCMI maintains a website (http://dublincore.org/documents/dces/) that hosts the current definitions of all the DC elements and their properties.

DC is a set of fifteen basic elements plus three additional elements. All elements are both optional and repeatable. The basic DC elements are:

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

The additional DC Elements are:

16. Audience
17. Provenance
18. Rights Holder

DC allows for element refinements (or subfields) that narrow the meaning of an element, making it more specific. The use of these refinements is not required. DC also allows for the addition of nonstandard elements for local use.

DSpace

DSpace is an open-source software package that provides management tools for digital assets. It is frequently used to create and manage institutional repositories.

First released in 2002, DSpace is a joint development effort of Hewlett Packard (HP) Labs and the Massachusetts Institute of Technology (MIT). Today, DSpace's future

Mary Kurtz (mhkurtz@gmail.com) is a June 2009 graduate of Drexel University's School of Information Technology. She also holds a BS in Secondary Education from the University of Scranton and an MA in English from the University of Illinois at Urbana–Champaign. Currently, Kurtz volunteers her time in technical services/cataloging at Simms Library at Albuquerque Academy and in corporate archives at Lovelace Respiratory Research Institute (www.lrri.org), where she is using DSpace to manage a diverse collection of historical photographs and scientific publications.

is guided by a loose grouping of interested developers called the DSpace Committers Group, whose members currently include HP Labs, MIT, OCLC, the University of Cambridge, the University of Edinburgh, the Australian National University, and Texas A&M University.

DSpace version 1.3 was released in 2005 and the newest version, DSpace 1.5, was released in March 2008. More than one thousand institutions around the world use DSpace, including public and private colleges and universities and a variety not-for-profit corporations.

DC is at the heart of DSpace. Although DSpace can be customized to a limited extent, the basic and qualified elements of DC and their refinements form DSpace's backbone.[2]

How DSpace works: a contributor's perspective

DSpace is designed for use by "metadata naive" contributors. This is a conscious design choice made by its developers and in keeping with the philosophy of inclusion for institutional repositories. DSpace was developed for use by a wide variety of contributors with a wide range of metadata and bibliographic skills. DSpace simplifies the metadata markup process by using terminology that is different from DC standards and by automating the production of element fields and XML/HTML code.

DSpace has four hierarchical levels of users: users, contributors, community administrators, and network/ systems administrators.

The user is a member of the general public who will retrieve information from the repository via browsing the database or conducting structured searches for specific information.

The contributor is an individual who wishes to add their own work to the database. To become a contributor, one must be approved by a DSpace community administrator and receive a password. A contributor may create, upload, and (depending upon the privileges bestowed upon him by his community administrator), edit or remove informational records. Their editing and removal privileges are restricted to their own records.

A community administrator has oversight within their specialized area of DSpace and accordingly has more privileges within the system than a contributor. A community administrator may create, upload, edit, and remove records, but also can edit and remove all records available within the community's area of the database. Additionally, the community administrator has access to some metadata about the repository's records that is not available to users and contributors and has the power to approve requests to become contributors and grant upload access to the database. Lastly, the community administrator sets the rights policy for all materials included in the database and writes the statement of rights that every contributor must agree to with every record upload.

The network/systems administrator is not involved with database content, focusing rather on software maintenance and code customization.

When a DSpace contributor wishes to create a new record, the software walks them through the process. DSpace presents seven screens in sequence that ask for specific information to be entered via check buttons, fill-in textboxes, and sliders. At the end of this process, the contributor must electronically sign an acceptance of the statement of rights.

Because DSpace's software attempts to simplify the metadata-creation process for contributors, its terminology is different from DC's. DSpace uses more common terms that are familiar to a wider variety of individuals. For example, DSpace asks the contributor to list an "author" for the work, not a "creator" or a "contributor." In fact, those terms appear nowhere in any DSpace. Instead, DSpace takes the text entered in the author textbox and maps it to a DC element—something that has profound implications if the mapping does not follow expected DC definitions.

Likewise, DSpace does not use "subject" when asking the contributor to describe their material. Instead, DSpace asks the contributor to list keywords. Text entered into the keyword field is then mapped into the subject element. While this seems like a reasonable path, it does have some interesting implications for how the subject element is interpreted and used by contributors.

DC's metadata elements are all optional. This is not true in DSpace. DSpace has both mandatory and automatic elements in its records. Because of this, data records created in DSpace look different than data records created in DC. These mandatory, automatic, and default fields affect the fill frequency of certain DC elements—with all of these elements having 100 percent participation.

In DSpace, the title element is mandatory; that is, it is a required element. The software will not allow the contributor to proceed if the title text box is left empty. As a consequence, all DSpace records will have 100 percent participation in the title element.

DSpace has seven automatic elements, that is, element fields that are created by the software without any need for contributor input. Three are date elements, two are format elements, one is an identifier, and one is provenance. DSpace automatically records the time of the each record's creation in machine-readable form. When the record is uploaded into the database, this timestamp is entered into three element fields: dc.date.available, dc.date.accessioned, and dc.date.issued. Therefore DSpace records have 100 percent participation in the date element. For previously published materials, a separate screen asks for the original publication date, which is then placed in the dc.date.issued element. Like title, the original date of publication is a mandatory field, and failure to enter a meaningful numerical date into the textbox will halt the creation of a record.

In a similar manner, DSpace "reads" the kind of file the contributor is uploading to the database. DSpace automatically records the size and type (.doc, .jpg, .pdf, etc.) of the file or files. This data is automatically entered into dc.format.mimetype and dc.format.extent. Like date, all DSpace records will have 100 percent participation in the format element. Likewise, DSpace automatically assigns a location identifier when a record is uploaded to the database. This information is recorded as an URI and placed in the identifier element. All DSpace records have a dc.identifier.uri field.

The final automatic element is provenance. At the time of record creation, DSpace records the identity of the contributor (derived from the sign-in identity and password) and places this information into a dc.provenance element field. This information becomes a permanent part of the DSpace record; however, this field is a hidden to users. Typically only community and network/systems administrators may view provenance information. Still, like date, format, and identifier elements, DSpace records have automatic 100 percent participation in provenance.

Because of the design of DSpace's software, all DSpace-created records will have a combination of both contributor-created and DSpace-created metadata.

All DSpace records can be edited. During record creation, the contributor may at any time move backward through his record to alter information. Once the record has been finished and the statement of rights signed, the completed record moves into the community administrator's workflow. Once the record has entered the workflow, the community administrator is able to view the record with all the metadata tags attached and make changes using DSpace's editing tools. However, depending on the local practices and the volume of records passing through the administrator's workflow, the administrator may simply upload records without first reviewing them.

A record may also be edited after it has been uploaded, with any changes being uploaded into the database at the end of editing process. In editing a record after it has been uploaded, the contributor, providing he has been granted the appropriate privileges, is able to see all the metadata elements that have attached to the record. Calling up the editing tools at this point allows the contributor or administrator to make significant changes to the elements and their qualifiers, something that is not possible during the record's creation. When using the editing tools, the simplified contributor interface disappears, and the metadata elements fields are labeled with their DC names. The contributor or administrator may remove metadata tags and the information they contain and add new ones selecting the appropriate metadata element and qualifier from a slider. For example, during the editing process, the contributor or administrator may choose to create dc.contributor.editor or dc.subject.lcsh options—something not possible during the record-creation process.

In the examination of the DSpace records from our three repositories, DSpace's shaping influence on element participation and metadata quality will be clearly seen.

The repositories

DSpace is principally used by academic and corporate nonprofit agencies to create and manage their institutional repositories. For this study, I selected three academic institutions that shared similar characteristics (large, public, research-based universities) but which had differing approaches to how they managed their metadata-quality issues.

The University of New Mexico (UNM) DSpace repository (DSpaceUNM) holds a wide-ranging set of records, including materials from the university's faculty and administration, the Law School, the Anderson School of Business Administration, and the Medical School, as well as materials from a number of tangentially related university entities like the Western Water Policy Review Advisory Commission, New Mexico Water Trust Board, and Governor Richardson's Task Force on Ethic Reform.

At the time of the initial research for this paper (spring 2008), DSpaceUNM provided little easily accessible on-site education for contributors about the DSpace record-creation process. What was offered—a set of eight general information files—was buried deep inside the library community. A contributor would have to know the files existed to find them.

By summer 2009, this had changed. DSpaceUNM had a new homepage layout. There is now a link to "help sheets and promotional materials" at the top center of the homepage. This link leads to the previously difficult-to-find help files.

The content of the help files, however, remains largely unchanged. They discuss community creation, copyrights, administrative workflow for community creation, a list of supported formats, a statement of DSpaceUNM's privacy policy, and a list of required, encouraged, and not required elements for each new record created. For the most part, DSpaceUNM help sheets do not attempt to educate the contributor in issues of metadata quality. There is no discussion of DC terminology, no attempts to refer the contributor to a thesaurus or controlled vocabulary list, nor any explanation of the record-creation or editing process.

This lack of contributor education may be explained in part because DSpaceUNM requires all new records to be reviewed by a subject area librarian as part of the DSpace community workflow. Thus any contributor errors, in theory, ought to be caught and corrected before being uploaded to the database.

The University of Washington (UW) DSpace repository (ResearchWorks at The University of Washington) hosts a narrower set of records than DSpaceUNM, with the materials limited to the those contributed by the university's faculty, students, and staff, plus materials from the UW's archives and UW's School of Public and Community Health.

In 2008, ResearchWorks was self-archiving. Most contributors were expected to use DSpace to create and upload their record. There is no indication in the publicly available information about the record creation workflow if record reviews were conducted before record upload. The help link on the ResearchWorks homepage brought contributors to a set of screen-by-screen instructions on how to use DSpace's software to create and upload a record. The step-through did not include instructions on how to edit a record once it had been created. No explanation of the meanings or definitions of the various DC elements was included in the help files. There also were no suggestions about the use of a controlled vocabulary or a thesaurus for subject headings. By 2009, this link had disappeared and the associated contributor education materials with it.

The Knowledge Bank at Ohio State University(OSU) is the third repository examined for this paper. OSU's repository hosts more than thirty communities, all of which are associated with various academic departments or special university programs.

Like ResearchWorks at UW, OSU's repository appears to be self-archiving with no clear policy statement as to whether a record is reviewed before it is uploaded to the repository's database.

OSU makes a strong effort to educate its contributors. On the upper-left of the Knowledge Bank homepage is a slider link that brings the contributor (or any user) to several important and useful sources of repository information: About Knowledge Bank, FAQs, Policies, Video Upload Procedures, Community Set-Up Form, Describing Your Resources, and Knowledge Bank Licensing Agreement.

The existence and use of metadata in Knowledge Bank are explicitly mentioned in the FAQ and Policies areas, together with an explanation of what metadata is and how metadata is used (FAQ), and a list of supported metadata elements (Policies). The Describe Your Resources section gives extended definitions of each DSpace-available DC metadata element and provides examples of appropriate metadata-element use.

Knowledge Bank provides the most comprehensive contributor education information of any of the three repositories examined. It does not use a controlled vocabulary list for subject headings, and it does not offer a thesaurus.

Data and analysis

I chose twenty randomly selected full records from each repository. No more than one record was taken from any one collection to gather a broad sampling from each repository. I examined each record for the quality of its metadata.

Metadata quality is a semantically slippery term. Park, in the spring 2009 special metadata issue of Cataloging and Classification Quarterly, suggested that most commonly accepted criteria for metadata quality are completeness, accuracy, and consistence.[3] Those criteria will be applied in this analysis.

For the purpose of this paper, I define completeness as the fill rate for key metadata elements. Because the purpose of metadata is to identify the record and to assist in the user's search process, the key elements are title, contributor/creator, subject, and description.abstract—all contributor-generated fields. I chose these elements because these are the fields that the DSpace software uses when someone conducts an unrestricted search.

Table 1 shows the fill rate for the title element is 100 percent for all three repositories. This is to be expected because, as noted above, title is mandatory field.

The fill rate for contributor/creator is likewise high: 16 of 20 (80 percent) for UNM, 19 of 20 (95 percent) for UW, and 19 of 20 (95 percent) for OSU. (OSU's fill rate for creator and contributor were summed because OSU uses different definitions for creator and contributor element fields than do UNM or UW. This discrepancy will be discussed in greater depth in the consistency of metadata terminology below.)

The fill rate for subject was more variable. UNM's subject fill rate was 100 percent, while UW's was 55 percent, and OSU's was 40 percent.

The fill rate for the description.abstract subfield was 12 of 80 (60 percent) at UNM, 15 of 20 (75 percent) at UW, and 8 of 20 (40 percent) at OSU. (See appendix A for a complete list of metadata elements and subfields used by each of the three repositories.)

The relatively low fill rate (below 50 percent) at the OSU KnowledgeBank in both subject and description .abstract suggests a lack of completeness in that repository's records.

Accuracy in metadata quality is the essential "correctness" of a record. Correctness issues in a record range from data-entry issues (typos, misspellings, and inconsistent date formats) to the correct application of metadata definitions and data overlaps.[4]

Accuracy is perhaps the most difficult of the metadata

Table 1. Metadata Fields and their Frequencies

Element	Univ. of N.M.	Univ. of Wash.	Ohio State Univ.
Title	20	20	20
Creator	0	0	16
Subject	20	11	8
Description	12	16	17
Publisher	4	4	8
Contributor	16	19	3
Date	20	20	20
Type	20	20	20
Identifier	20	20	20
Source	0	0	0
Language	20	20	20
Relation	3	1	6
Coverage	2	0	0
Rights	2	0	0
Provenance	**	**	**

**provenance tags are not visible to public users

quality criteria to judge. Local practices vary widely, and DC allows for the creation of custom metadata tags for local use. Additionally, there is long-standing debate and confusion about the definitions of metadata elements even among librarians and information professionals.[5] Because of this, only the most egregious of accuracy errors were considered for this paper.

All three repositories had at least one record that contained one or more inaccurate metadata fields; two of them had four or more inaccurate records.

Inaccurate records included a wide variety of accuracy errors, including poor subject information (no matter how loosely one defines a subject heading, "the" is not an accurate descriptor); mutually contradictory metadata (record contained two different language tags, although only one applied to the content); and one in which the abstract was significantly longer and only tangentially related than the file it described. Additionally, records showed confusion over contributor versus creator elements. In a few records, contributors entered duplicate information into both element fields. This observation supports Park and Childress's findings that there is widespread confusion over these elements.[6]

Among the most problematic records in terms of accuracy were those contained in UW's Early Buddhist Manuscripts Project. This collection, which has been removed from public access since the original data was drawn for this paper, contained numerous ambiguous, contradictory, and inaccurate metadata elements.[7]

While contributor-generated subject headings were specifically not examined for this paper, it must be noted that was a wide variation in the level of detail and vocabulary used to describe records. No community within any of the repositories had specific rules for the generation of keyword descriptors for records, and the lack of guidance shows.

Consistency can be defined as the homogeneity of formats, definitions, and use of DC elements within the records. This consistency, or uniformity, of data is important because it promotes basic semantic interoperability. Consistency both inside the repository itself and with other repositories makes the repository easier to use and provides the user with higher quality information.

All three repositories showed 100 percent consistency in DSpace-generated elements. DSpace's automated creation of date and format fields provided reliably consistent records in those element fields. DSpace's automatic formatting of personal names in the dc.contributor.author and dc.creator fields also provided excellent internal consistency. However, the metadata elements were much less consistent for contributor-generated information.

Inconsistency within the subject element is where most problems occurred. Personal names used as subject heading and capitalization within subject headings both proved to be particular issues. DSpace alphabetizes subject headings according to the first letter of the free text entered in the keyword box. Thus the same name entered in different formats (first name first or last name first) generates different subject-heading listings. The same is true for capitalization. Any difference in capitalization of any word within the free-text entry generates a separate subject heading.

Another field where consistency was an issue was dc.description.sponsorship. Sponsorship is problem because different communities, even different collections within the same community, use the field to hold different information. Some collections used the sponsorship field to hold the name of a thesis or dissertation advisor. Some collections used sponsorship to list the funding agency or underwriter for a project being documented inside the record. Some collections used sponsorship to acknowledge the donation of the physical materials documented by the record. While all of these are valid uses of the field, they are not the same thing and do not hold the same meaning for the user.

The largest consistency issue, however, came from a comparison of repository policies regarding element use and definition. Unaltered DSpace software maps contributor-generated information entered into the author textbox during the record-creation process into the dc.contributor.author field. However, OSU's DSpace software has been altered so that the dc.contributor.author field does not exist. Instead, text entered into the author textbox during the record-creation process maps to dc.creator. Although both uses are correct, this choice does create a significant difference in element definitions. OSU's DSpace author fields are no longer congruent with other DSpace author fields.

Conclusions

DSpace was created as repository management tool. By streamlining the record creation workflow and partially automating the creation of metadata, DSpace's developers hoped to make institutional repositories more useful and functional while time providing an improved experience for both users and contributors. In this, DSpace has been partially successful.

DSpace has made it easier for the "metadata naive" contributor to create records. And, in some ways, DSpace has improved the quality of repository metadata. Its automatically generated fields ensure better consistency in those elements and subfields. Its mandatory fields guarantee 100 percent fill rates in some elements, and this contributes to an increase in metadata completeness. However, DSpace still relies heavily on contributor-generated data to fill most of the DC elements, and it is in these contributor-generated fields that most of the metadata quality issues arise. Nonmandatory fields are skipped, leading to incomplete records. Data entry errors, a lack of authority control over subject headings, and confusion over element definitions can lead to poor metadata accuracy. A lack of enforced, uniform naming and capitalization conventions leads to metadata inconsistency, as does the localized and individual differences in the application of metadata element definitions.

While most of the records examined in this small survey could be characterized as "acceptable" to "good," some are abysmal. To improve the inconsistency of the DSpace records, the three universities have tried differing approaches. Only UNM's required record review by a subject area librarian before upload seems to have made any significant impact on metadata quality. UNM has a 100 percent fill rate for subject elements in its records, while UW and OSU do not. This is not to say that UNM's process is perfect and that poor records do not get into the system—they do (see appendix B for an example). But it appears that for now, the intermediary intervention of a librarian during the record-creation process is an improvement over self-archiving—even with education—by contributors.

References and notes

1. Association of Library Collections & Technical Services, Committee on Cataloging: Description & Access, Task Force on Metadata, "Final Report," June 16, 2000, http://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html (accessed Mar. 10, 2007).

2. A voluntary (and therefore less-than-complete) list of current DSpace users can be found at http://www.dspace.org/index.php?option=com_content&task=view&id=596&Itemid=180. Further specific information about DSpace, including technical specifications, training materials, licensing, and a user wiki, can be found at http://www.dspace.org/index.php?option=com_content&task=blogcategory&id=44&Itemid=125.

3. Jung-Ran Park "Metadata Quality in Digital Repositories: A Survey of the Current State of the Art," *Cataloging & Classification Quarterly* 47, no. 3 (2009): 213–28.

4. Sarah Currier et al., "Quality Assurance for Digital Learning Object Repositories: Issues for the Metadata Creation Process," *ALT-J: Research in Learning Technology* 12, no. 1 (2004): 5–20.

5. Jung-Ran Park and Eric Childress, "DC Metadata Semantics: An Analysis of the Perspectives of Informational Professionals," *Journal of Information Science* 20, no. 10 (2009): 1–13.

6. Ibid.

7. For a fuller discussion of the collection's problems and challenges in using both DSpace and DC, see Kathleen Forsythe et al., *University of Washington Ealy Buddhist Manuscripts Project in DSpace* (paper presented at DC-2003, Seattle, Wash., Sept. 28–Oct. 2, 2003), http://dc2003.ischool.washington.edu/Archive-03/03forsythe.pdf (accessed Mar. 10, 2007).

Index to Advertisers

Appendix A. A list of the most commonly used qualifiers in each repository

University of New Mexico

dc.date.issued (20)
dc.date.accessioned (20)
dc.date.available (20)
dc.format.mimetype (20)
dc.format.extent (20)
dc.identifier.uri (20)
dc.contributor.author (15))
dc.description.abstract (12)
dc.identifier.citation (6)
dc.description.sponsorship (4)
dc.subject.mesh (2)
dc.contributor.other (2)
dc.description.sponsor (1)
dc.date.created (1)
dc.relation.isbasedon (1)
dc.relation.ispartof (1)
dc.coverage.temporal (1)
dc.coverage.spatial (1)
dc.contributor.other (1)

University of Washington

dc.date.accessioned (20)
dc.date.available (20)
dc.date.issued (20)
dc.format.mimetype (20)
dc.format.extent (20)
dc. identifier.uri (20)
dc.contributor.author (18)
dc.description.abstract (15)
dc.identifier.citation (4)
dc.identifier.issn (4)
dc.description.sponsorship (1)
dc.contributor.corporateauthor (1)
dc.contributor.illustrator (1)
dc.relation.ispartof (1)

Ohio State University

dc.date.issued (20)
dc.date.available (20)
dc.date.accessioned (20)
dc.format.mimetype (20)
dc.format.extent (20)
dc.identifier.uri (20)
dc.description.abstract (8)
dc.identifier.citation (4)
dc.subject.lcsh (4)
dc.relation.ispartof (4)
dc.description.sponsorship (3)
dc.identifier.other (2)
dc.contributor.editor (2)
dc.contribtor.advisor (1)
dc.identifier.issn (1)
dc.description.duration (1)
dc.relation.isformatof (1)
dc.description.statementofresponsibility (1)
dc.description.tableofcontents (1)

Appendix B. Sample Record

dc.identifier.uri
http://hdl.handle.net/1928/3571

dc.description.abstract
President Schmidly's charge for the creation of a North Golf Course Community Advisory Board.

dc.format.extent
17301 bytes

dc.format.mimetype
application/pdf

dc.language.iso
en_US

dc.subject
President

dc.subject
Schmidly

dc.subject
North

dc.subject
Golf

dc.subject
Course

dc.subject
Community

dc.subject
Advisory

dc.subject
Board

dc.subject
Charge

dc.title
Community_Advisory_Board_Charge

dc.type
Other