*Cataloging & Classification Quarterly*, 47:309–325, 2009

ISSN: 0163-9374 print / 1544-4554 online
DOI: 10.1080/01639370902735020

Routledge
Taylor & Francis Group

# Repository Metadata: Approaches and Challenges

JOHN W. CHAPMAN
*University of Minnesota Libraries, Minneapolis, Minnesota, USA*

DAVID REYNOLDS
*Johns Hopkins University, Baltimore, Maryland, USA*

SARAH A. SHREEVES
*University of Illinois Urbana-Champaign, Urbana, Illinois, USA*

*Many institutional repositories have pursued a mixed metadata environment, relying on description by multiple workflows. Strategies may include metadata converted from other systems, metadata elicited from the document creator or manager, and metadata created by library or repository staff. Additional editing or proofing may or may not occur. The mixed environment brings challenges of creation, management, and access. In this article, repository efforts at three major universities are discussed. All three repositories run on the DSpace software package, and the opportunities and limitations of that system will be examined. The authors discuss local strategies in light of current thinking on metadata creation, user behavior, and the aggregation of heterogeneous metadata. The contrasts between the mission of each repository effort will show the importance of local customization, while the experience of all three institutions forms the basis for recommendations on strategies of benefit to a wide range of librarians and repository planners.*

*KEYWORDS metadata, DSpace, institutional repository, heterogeneity*

Received August 2008; revised September 2008; accepted October 2008.
Address correspondence to John W. Chapman, Product Manager, Cataloging and Metadata Services, OCLC, 6565 Kilgour Place, Dublin, OH 43017. E-mail: chapmanj@oclc.org

## INTRODUCTION

An institutional repository (IR) collects, manages, and disseminates materials produced at an institution. Lynch[1] describes institutional repositories as "essentially an organizational commitment to the stewardship of . . . digital materials, including long-term preservation where appropriate, as well as organization and access or distribution." An institutional repository can contain a range of materials including administrative records, but tends to focus on the research and scholarship of faculty, students, and staff, as well as other materials that reflect the intellectual environment of a campus. As of October 2008, OpenDOAR[2] lists over 1,000 institutional repositories from around the world.

The hallmark of the IR environment, particularly as conceptualized early in its development, is that the IR would be populated by the faculty and researchers of the campus; that is, researchers or their proxies would deposit and describe or self-archive their own papers and data.[3] For librarians and repository managers, this opportunity to distribute the descriptive workload is attractive. The creator or manager of a resource, rather than a cataloger, often is the most knowledgeable about a resource's content and can identify appropriate (although often uncontrolled) subject terms to describe their work. In practice, of course, self-archiving has not proven to be a successful way to fill IRs. McDowell in her study to evaluate the deployment of institutional repositories found that from 2005 to 2006 the median annual increase was 366 items, or essentially one new deposit per day.[4] Only a percentage of these are actual self-archived deposits. Many institutional repository managers and librarians deposit published research on behalf of faculty and bring into the repository other appropriate material produced on their campuses. Grey literature such as technical reports, working papers, and occasional papers, as well as digitized historical research, images, data sets, and audio and visual material all have a place in the repository. The backfiles of this type of material are often ingested in bulk using metadata scraped from Web sites or mapped from spreadsheets and databases. Thus, many IRs ingest metadata through multiple workflows and contain a mixed metadata environment. Metadata is mapped and converted from existing systems, elicited from the document creator or manager, or created by library or repository staff. Few institutional repositories in our experience, however, have put extra effort into augmenting, correcting, or editing the metadata coming into the repository. This may be because of staffing shortages at an institution, a sense that full text indexing by both the software in use and by search engines mitigates the need to augment the metadata, or because material in the IR is felt to be of lesser priority than other material.

This mixed metadata environment means that institutional repositories face a number of challenges that more controlled environments do not (or face to a much lesser degree) including inconsistency of the metadata across

the entire repository and the lack of authority control and complex controlled vocabularies. Repositories often include metadata coming from a range of disciplines, each of which have different citation traditions and different emphases on the type of information they share. Because metadata is coming into the repository from many different streams, including directly from researchers themselves, it can be difficult to enforce consistent use of metadata and entry of metadata values except at rather rudimentary levels (i.e., required fields and use of lists of terms). Metadata can be sparse or lack important contextual information particularly when that context is held at a collection level. The breadth and depth of disciplines across an academic institution means that use of controlled subject terms is possible at only the highest levels. Authority control for author names is also difficult. Use of the Library of Congress Name Authority File is problematic because many authors in institutional repositories have no entry, as they tend to be authors of journal articles and conference papers, not books or monographs. Use of the campus-level directory can aid in some cases, but often faculty leave or publish under a name different from their directory name leaving gaps in its usefulness for authority control. There exists no standard to uniquely identify authors; this gap is a growing problem for institutional and disciplinary repositories as well as for journal publishers and aggregators.

There are several software platforms available for IRs including Eprints, DSpace, and Digital Commons. As the IRs discussed in this article all use DSpace, the following is a brief description of the software. DSpace is an open-source repository software platform, developed jointly by the Massachusetts Institute of Technology (MIT) Libraries and Hewlett-Packard.[5] Currently, the software has oversight from a board of "maintainers" that respond to requests and reports from the user community to guide its development. The software is Java-based, and runs most easily on a PostgreSQL database. Server, storage, and memory requirements are medium-level, meaning that it may be a challenge for the smallest libraries to run; staff with UNIX skills and database administration experience will be required.

Although DSpace can handle multiple item formats, including audio, image, and video, it is most commonly used as a repository for text documents. One of its distinguishing features is its ingest process, which allows delegated users to submit their own materials to the repository. This process can be customized to include workflows for editing or checking metadata, approving submission before public availability, and adding default metadata values. The default metadata schema in use is a qualified Dublin Core based loosely on the Dublin Core Library Application Profile.[6]

High-level communities in DSpace contain collections or subcommunities that themselves hold collections. Collections consist of records describing items, which can consist of one or more digital files, or bitstreams. Metadata, rights and permissions, and database structure are heavily dependent on this hierarchical model.

In this article, repository efforts at three major universities are described. The authors discuss local strategies in light of current thinking on metadata creation, user behavior, and the aggregation of heterogeneous metadata. The contrasts among the missions of each repository program illustrate the importance of local customization and practice, whereas the similarities among the experience of all three institutions form the basis for recommendations on strategies of benefit to a wide range of librarians and repository planners. As noted earlier, all three repositories use the DSpace software.

## CASE STUDIES

### Metadata and the University of Minnesota Digital Conservancy

The University of Minnesota Digital Conservancy is responsible for two major instances of DSpace. The first is AgEcon Search,[7] a repository for prepublication papers and published journal articles in agricultural economics and applied economics. The second DSpace instance supports the main institutional repository for the University of Minnesota.[8] For the rest of this discussion, the terms University Digital Conservancy or UDC will refer to this second instance. The instances are functionally distinct, but the staff administering the software is the same, so expertise is shared between both programs.

Working within the DSpace architecture has posed challenges for each program. In the case of AgEcon Search, these challenges are largely related to the mix of entire runs of formally published journal articles with pre-published manuscripts. In the case of the University Digital Conservancy, a different mix—of university administrative materials and scholarly works—has proved less problematic, while the distributed workflow and variety of existing metadata have been the main concerns.

The community/sub-community/collection hierarchy on which DSpace is designed has proven troublesome on multiple levels—that of metaphor, that of practice, and that of technological architecture. In metaphor, the difficulty is that the terms community and sub-community are an ill fit for many materials and their parent collections. In practice, the rigidity of collection and community membership means that associating an item with multiple collections is not as transparent as it should be. Moreover, on the level of technological architecture, the way that certain administrative, rights, and relational metadata is stored—preferring implicit inheritance through the hierarchy over explicit information in item records—has been a source of frustration.

AgEcon Search is administered by two librarians, one within the University Libraries organization and the other employed by the Department of Economics. The resource is maintained by those two organizations and receives sponsorship from the American Agricultural Economics Association.

As AgEcon Search is largely run outside of the direction of the University Digital Conservancy's Working Group, it will not be discussed at great length here. One particular metadata peculiarity is worth mentioning, however. The repository includes previously published journal articles, and the desire to support flexible re-use of citation information for those articles led to the extension of the metadata framework. In the out-of-the-box metadata schema, DSpace supports a bibliographic citation in a single field. The UDC staff extended the schema with additional pagination and publication information, including starting and ending pages, page count, and issue information.

## The University Digital Conservancy: Collections

The University Digital Conservancy differs from many institutional repository efforts due to its strong focus on materials collected by the University Archives. These materials include administrative records, publications, and other materials from all departments of the institution. This collecting responsibility, already a difficult task due to the size and organizational complexity of the University, has only been made more complex as born-digital materials proliferate. The UDC was seen as a strategic opportunity to make the University Archives a more visible partner on campus.

This approach is coupled with a more common scholarly communications initiative[9] that seeks to promote awareness of open access and copyright among the faculty and staff of the University of Minnesota. While individual submission of articles has not been common, a number of academic papers created through centers or collaboratives have been submitted to the Conservancy.

The University Digital Conservancy has sought to collect materials in several ways:

- seeking out existing task forces, interdisciplinary efforts, and other groups with a strong tradition of publishing and engaging them on the merits of preservation and access;
- building relationships with faculty—a scholarly communications effort that promoted open access and increased visibility;
- promoting the UDC as the "digital arm" of the University Archives, to build ongoing deposit agreements with central University departments; and
- ad hoc identification (by liaison librarians 'in the field') of collections of departmental publications or resources for which the UDC could provide organization, access, and preservation.

These strategies, each of which has been successful to some degree, have resulted in nearly 7,000 items being added to the Digital Conservancy. What

follows are mini case studies that discuss the descriptive work done for differing types of collections.

*Senate Committee minutes.* The University of Minnesota faculty governance has several Senate Committees that discuss and determine policy in specific areas: faculty affairs, academic freedom, and educational policy among them. The University administrative offices provide these files to the University archives in electronic form, as word processing documents. Currently, these files come in a slow trickle as they are created, but in the past, these were entered in batches by copy-and-pasting from a spreadsheet.

Basic metadata such as author and publisher are straightforward, as these are University creations. As these are consistent across entire collections, they can be entered as default values through the DSpace administrative interface. The filename contains date information, so some text manipulation is performed to put this data into a text string, which is appended to a boilerplate title, that is, "scfa-12–02-89.pdf" becomes "Minutes: Senate Committee on Faculty Affairs: December 2, 1989." In addition, some minutes contain a brief statement on the first page regarding major points of discussion. Although it was tempting to enter these as subjects, there is no control or consistency as to how they are recorded, so these were included in a description element.

The Senate Committee minutes were ingested into DSpace using a combination of simple copy-and-pasting and basic data entry, as the date fields in the default ingest form are separated into month, day, and year values. Here the common Web browser feature of field autocomplete is quite welcome, as the values that have been entered already are easily selected for subsequent entries.

*Institute for Math and its Applications.* The Institute for Math and its Applications (IMA)[10] has made available a preprint series that extends back to 1982. In this case, the Institute for Math and its Applications had existing metadata for its collections, including title, author, date, and identification number. The full-text indexing capabilities of DSpace were the selling point for the Conservancy; otherwise, they had a workable HTML-based index. However, early PDFs in this collection were simple image files rather than converted word processing files, making text indexing impossible. A retrospective OCR project was launched to help solve this.

For ingest, text processing was used to convert information from existing Web sites into a spreadsheet. With this collection, however, manual-cut-and-pasting was seen as too laborious—there were over 2,000 papers, many different authors, and the existing identification numbers. For this collection, the spreadsheet was used as the basis for batch import process. This required resolving filenames to full directory locations and making sure that the files were live and available for upload.

*Board of Regents minutes.* The Board of Regents minutes have a similar format to the Senate Committee minutes, However, this was a purely

retrospective effort, seeking to put online the bound volumes of minutes comprising their paper-only publication run, which ended in the early 1990s. The print minutes also were accompanied by an extensive card index, maintained by the University Archives. The index tracks several dozen themes, each one represented by series of cards listing discussion topics in that area, referencing the meeting by volume and page number, and arranged in chronological order. For the purposes of this project, a smaller group of popular themes were selected, and the cards representing those themes were transcribed. Once the document references were processed and sorted, it was trivial to concatenate all the references into a brief description of major points of discussion.

Taking advantage of this preprocessing of the descriptive information, and with the benefits of spreadsheet consistency outlined in the earlier discussion of the Senate Committee minutes, it was possible to sustain a rate of ingest of 1 document per 90 seconds over several employee-hours.

*Strategic positioning files.* These documents were created as part of a University-wide effort to examine and overhaul several areas of its operations and policies. The collection comprises a number of different types of documents, including presentation slides, executive summaries, and informational brochures. In most cases, little to no metadata was supplied with the documents; and some did not have explicit authorship or date information.

Metadata was inferred from several sources, including metadata saved as part of the file properties, and documents representing other versions of the document in question. In some cases, the DSpace facility of assigning multiple bitstreams to one object was used to collect a full report and its accompanying executive summary, when it was extracted and published as a separate document. Ingest was time-consuming, between 5 and 10 minutes per document, due to the examination that each document required.

The experience of metadata creation at the University of Minnesota has encouraged staff to pursue automated or streamlined ways to create metadata. The lack of effective metadata management or editing tools within DSpace means that mistakes are especially expensive to correct. Forethought and rigor in designing collection-level templates and attention to detail when filling out spreadsheets is essential.

## JScholarship at The Johns Hopkins University

JScholarship (http://jscholarship.library.jhu.edu), the Johns Hopkins institutional repository, is the home for research materials created by faculty and staff from the university, the medical institutions, and other affiliates such as the Applied Physics Lab. After a lengthy pilot phase, JScholarship officially launched in February 2008. This DSpace-based repository is a service developed and operated jointly by the Sheridan Libraries (arts and sciences and

engineering) and the Welch Medical Library. The directors of both libraries and several key staff members serve as the Oversight Group for JScholarship. This group establishes high-level policies for the repository and provides guidance to the IR manager in areas such as content recruitment and assessment.

During the planning phase of the project, the Oversight Group discussed the merits of having library staff perform most submission activities, including metadata creation. Because cataloging department staff already create metadata for both analog and digital library resources, they could be relied on to create consistent descriptions of repository objects. On the other hand, members of the research community who are creating the materials that will populate JScholarship have the domain knowledge necessary to create the most accurate subject descriptors. Also, involving the research communities in the submission process may increase their sense of investment in JScholarship. Ultimately, the Oversight Group decided to leave the submission process and metadata creation to the various research communities, with library staff acting only in a training and advisory role. So far, each community has created its metadata at the time of submission, but the library is experimenting with harvesting existing metadata to use for batch ingestion of digitized library collections.

In addition to serving as the metadata authority, each research community establishes many of the policies for its collections. Although there are a few policies that apply throughout JScholarship, the individual communities have several options for metadata and quality control. The Oversight Group believes that a community-driven set of policies is best because each community has its own needs. Some may consider it necessary to establish strict policies about appropriate content and enforce them through an approval system. Other communities prefer to leave content questions up to the individual researchers. Metadata requirements follow a similar pattern. One community may use a controlled vocabulary for subject terms whereas another will simply use uncontrolled keywords. The Oversight Group has placed a very low barrier to participation in JScholarship. The only metadata required for all collections are title and date, and no controlled vocabulary is specified for subject terms or author names.

## JScholarship Metadata Approaches

The lack of rigid requirements for metadata and quality control has led to a number of different scenarios in the short time that the service has been available. Following are three approaches to metadata policies that have been established by different JScholarship communities.

*The Center for Africana Studies* takes an interdisciplinary approach to the study of African peoples throughout the world. In addition to offering

coursework for both an undergraduate major and minor, the center offers research opportunities for graduate students, and sponsors conferences, workshops, and symposia. Within JScholarship, the Center elected to create collections for center research, faculty articles, and working papers. Due to its interdisciplinary nature, faculty who teach and carry out research in the Center have their primary appointments in other departments. This decentralization of researchers motivated the Center to centralize its JScholarship submission activities in the director's office. An administrative assistant in the director's office gathers research, uploads files, and creates the metadata for each of the Center's collections. Because all of these activities are performed by a single person, there is no need to add the additional approval layer favored by communities with multiple submitters. This keeps the process simple and brings a certain amount of consistency to the metadata. The interdisciplinary nature of the collections does not lend itself to using a specialized controlled vocabulary for subject terms. Although a wide-ranging thesaurus such as Library of Congress Subject Headings would work with these materials, the Center has opted to use keywords from the articles themselves.

*The Hopkins Population Center* facilitates interdisciplinary population research throughout the Johns Hopkins University. Faculty associates in the Bloomberg School of Public Health, the School of Medicine, and the School of Nursing produce most of the research. The Center's contributions to JScholarship include working papers, conference proceedings, and journal articles. The Population Center was one of the first pilot collections in JScholarship and the first group outside the library to test the submission interface. Unlike the Center for Africana Studies, the Population Center used a distributed submission model for part of the pilot. Instead of having a single person perform the submission, metadata creation, and approval, they had students perform some of the submission and basic metadata tasks. The submissions were then checked and enhanced by a liaison librarian from the Welch Medical Library. Overall, they found the interface to be satisfactory, but they also offered some suggestions that have been incorporated into the current version. In addition, the Population Center is the only community so far to use a controlled vocabulary for subject terms. Because they already have their own thesaurus for their POPLINE database, they decided to use those terms in the JScholarship metadata. Submitters suggested that it would be a significant improvement to be able to select POPLINE descriptors directly from the metadata submission page. This feature was not incorporated into the pilot, but will be considered for a future enhancement.

While library personnel submitted *electronic theses and dissertations* (ETDs) during the JScholarship pilot phase, repository staff are now in the process of setting up a new workflow for these materials. The Johns Hopkins Graduate Board mandated that students submit an electronic copy of their thesis or dissertation to JScholarship beginning in the spring 2009 semester.

JScholarship personnel have worked closely with the Graduate Board to determine both metadata standards and workflow issues. Students will be submitting their own ETDs to JScholarship after their department certifies them as eligible. No controlled vocabulary will be required, but students are encouraged to use subject terms appropriate to their field of study. The Graduate Board also suggested additional metadata elements such as the name of the student's advisor, degree granted, and granting department. After the student has submitted his or her thesis, the library will check both the ETD and the accompanying metadata before making the materials available. An embargo function that will enable the student to restrict access to the ETD for a set period is under development. This will be a new function for DSpace, and will be shared with the rest of the DSpace community.

## Illinois Digital Environment for Access to Learning and Scholarship (IDEALS) at the University of Illinois at Urbana-Champaign

This section briefly describes the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS), a set of collections and services that serves as the institutional repository for the University of Illinois at Urbana-Champaign (UIUC) and discusses some of the decisions and strategies for attempting to ensure a minimum of metadata quality and consistency. Also identified are areas that need further work.

### Profile of IDEALS

IDEALS collects, manages, and preserves the research and scholarship produced at UIUC as well as material that reflects the intellectual environment of the university; it does not collect, however, administrative records, standard curricular material, or digitized special collections except where these fall under the general collection policy mentioned earlier. Faculty, staff, and graduate students may deposit directly into IDEALS; a faculty member must sponsor undergraduate work. The material in IDEALS ranges from faculty pre- and post-prints, entire journal runs, data sets, technical reports, working papers, video, audio, and a selection of student work. IDEALS has been in production since the summer of 2006. It currently contains over 7,500 items and has seen over 240,000 downloads between December 2006 and July 2007. Downloads have been filtered as much as possible for spiders and mass downloaders.

IDEALS is a joint initiative of the University Library and CITES, the academic computing organization on the UIUC campus. It is staffed by two full-time employees, a program manager and a technical lead, both of whom are based in the Library. In addition, a third of an additional programmer

is attached to the program to provide back-up support for the technical infrastructure and an additional quarter of a librarian assists in faculty liaison activities. Other staff in the Library and in CITES are involved as needed. Organizationally IDEALS sits under the Office of Information Technology Planning and Policy in the Library.

IDEALS, like most DSpace installations, is organized into communities and collections that tend to correspond to research and academic units on campus. This is not a strict guideline; for example, IDEALS contains a top-level community for the blog and related works of a scholar in residence. Communities are established in consultation with the units and individuals involved and may be managed according to the needs of that unit or individual within the bounds of IDEALS' policies and procedures.[11] In addition, IDEALS allows faculty, graduate students, and staff who do not belong to an already established community to deposit their work into an open community called UIUC Research and Scholarship [Uncategorized]. This allows depositors to bypass the step of setting up a community specifically for their unit.

## Metadata in IDEALS

IDEALS has done little to modify the metadata native to DSpace beyond attempting to align it better with Dublin Core terms. Although the metadata in DSpace was meant to conform to the Dublin Core Library Application Profile and to use the Dublin Core terms namespace, it does not actually do either and does not allow, for example, a means to pull in changes to Dublin Core terms. For example, DSpace by default qualifies the element identifier with the term citation. In order to conform to Dublin Core terms, IDEALS uses the term bibliographicCitation.[12] Management of the metadata terms is difficult as well. DSpace includes a metadata registry that has limited utility; it consists of the name of an element, the name of a qualifier (if any), and a general scope note that, in IDEALS, includes the definition, whether or not the term is required, where it is repeatable, and whether it is system supplied.

Metadata is entered into IDEALS through a variety of means, but the two primary routes are via direct entry by a depositor or via a batch upload. Batch uploads generally occur when a unit has a medium to large number of items (such as a working paper series from a department) to deposit. In most cases, the unit also has metadata describing these items stored in a database or spreadsheet. IDEALS staff develop an initial crosswalk between the native metadata and the schema internal to IDEALS. This is shared with the unit and modified as needed before uploaded into IDEALS. Often metadata is constructed by drawing from multiple places. For example, in order to create article level metadata for the backfiles of *Library Trends*, metadata was pulled from:

- MARC records describing the series;
- Vendor created metadata notating author(s), article title, and page ranges; and
- TEI header information containing the volume, issue, and date of the article.

The quality and consistency of the metadata in IDEALS has long been a concern, but has been balanced by the need to provide the quickest and easiest deposit process possible. IDEALS requires only the title, the type of resource (using a broad type vocabulary), the date issued, and one uncontrolled subject term. There are various other fields available for use (author, genre, language, description, citation information, copyright information, series name, peer-review status, publication status, publisher, and sponsor), but none of these are required. As noted earlier, IDEALS allows units and individuals to define policies and guidelines for their own communities and collections, so units can make decisions to enforce use of certain fields as well as to enforce use of controlled vocabularies, but at this point in IDEALS, neither of those constraints can be system enforced.

As a result the metadata in IDEALS can vary wildly (certainly there is sometimes a vast difference between user deposited metadata and the metadata created through the mapping work described earlier), but there can also be subtle differences that affect services that IDEALS might want to offer. As an illustrative example, Table 1 shows two user deposited records that each describe a different paper from a conference proceedings. Although there is much similarity between these records, note particularly the differences in the type.genre field and the identifier.bibliographicCitation field. If IDEALS wanted to provide a service where appropriate citations were created automatically for each of these papers, Record One would not be treated as a conference paper but as an article and Record Two would have incomplete information because of the lack of page numbers, place names in the citation. Of course, neither of these records has the information parsed out in such a way to easily and automatically create a citation, but the inconsistency of the data is problematic from the start.

There was some discussion early in the formation of IDEALS whether catalogers should provide remediation services for metadata in IDEALS. Although the metadata librarian at UIUC does assist with some of the metadata mappings for bulk uploads, an early decision was made to not do remediation on metadata on an item-by-item basis. Two factors were key in making this decision. First, it was not clear what the staffing implications were likely to be for the cataloging unit and due to chronic staffing shortages, the original IDEALS Metadata Working Group felt that it would not be a good use of resources. Second, the interface for editing metadata on an item-by-item basis is difficult to use due to a poor user interface. In addition, there was a general feeling that because of the nature of the institutional repository,

**TABLE 1** Comparison of Two User Deposited Metadata Records in IDEALS

| | Record One | Record Two |
|---|---|---|
| **creator** | Kendall, Lori | Maitre, Matthieu<br>Guillemot, Christine<br>Morin, Luce |
| **date.issued** | 2007-01-07 | 2006–10 |
| **identifier.url** | http://hdl.handle.net/2142/705 | http://hdl.handle.net/2142/159 |
| **description.abstract** | With the availability of relatively easy-to-use tools for online video creation and distribution, people are increasingly producing videos not just for artistic expression, but also as a form of communication... | The compression efficiency of Distributed Video-Coding (DVC) suffers from the necessity of transmitting a large number of key-frames, which are intra-coded... |
| **language** | en | en |
| **publisher** | IEEE | IEEE |
| **rights** | Copyright owned by IEEE | Copyright owned by IEEE |
| **subject** | • Video<br>• Masculinities<br>• Persistent conversation<br>• Animation | • Video coding<br>• Stereo vision |
| **title** | Colin Mochrie vs. Jesus H. Christ: Messages About Masculinities and Fame in Online Video Conversations | 3D scene modeling for Distributed Video Coding |
| **type** | text | text |
| **type.genre** | article | conference paper |
| **description.status** | published or submitted for publication | published or submitted for publication |
| **description.peerReview** | is peer reviewed | is peer reviewed |
| **identifier.bibliographicCitation** | hicss, p. 76b, 40th Annual Hawaii International Conference on System Sciences (HICSS'07), 2007 | International Conference on Image Processing (ICIP) |

access to resources would principally occur through search engines and full text indexing.

Given the constraints IDEALS was working under—lack of resources to do extensive metadata remediation and the need to make deposit as easy as possible—IDEALS staff looked for other ways to increase quality and consistency of the metadata. Most of these have involved making changes to DSpace itself under the premise that the system should make it easier to enter consistent metadata. IDEALS has made two major changes that help to increase metadata quality. The first was to allow IDEALS staff to rearrange the steps for the deposit process to make the metadata slightly easier to enter.[13] This means, for example, that users can upload their documents before entering metadata. In this way, they can open the document in a separate window as they fill in pieces of the metadata. IDEALS can also more easily rearrange and collapse metadata entry pages. This code has been incorporated into the 1.5 version of DSpace.

The second customization was to provide a means for communities to develop their own controlled vocabularies for subjects and authority files for author and publisher names as they deposit items. As a depositor enters an author, for example, a drop-down menu appears with names containing the entered string that already exist in the IDEALS database; as the string lengthens the number of names drops. If the name has been entered before, the depositor can select that version. The same process occurs for subject terms, although here the menu is divided between terms that are used within the community in which the deposit is occurring and terms used generally throughout IDEALS. This utility has been in IDEALS since fall 2007, and IDEALS staff has yet to do an analysis of the effectiveness of this tool. However, there has been informal feedback that depositors, particularly those who are attempting to use a set of controlled vocabulary, have found this quite useful.

## FUTURE DEVELOPMENT NEEDS

By examining the common experiences of Minnesota, Illinois, and Johns Hopkins, the authors have observed some gaps in metadata infrastructure that need significant research and development work either within DSpace or within the institutional repository community in general. Although the experiences described here have been with DSpace, many of these same issues would apply to other software platforms as well.

### Expansion of Metadata Standard Support

The diversity of disciplines and formats of content in institutional repositories makes it difficult to use a blanket metadata format and allow only high-level

controlled vocabularies to be used across the institutional repository. Ideally, IRs should allow for the selection of an appropriate metadata schema and appropriate controlled vocabularies at the community or collection level, rather than having to shoehorn existing metadata and descriptive practices into a single Dublin Core (DC) schema. Standard schemas such as Metadata Object Description Schema (MODS) and the Visual Resources Association's VRA Core are widely supported and provide more granular description than simple DC. By integrating these metadata standards into the IR environment, it would enable batch ingestion of a wider variety of materials without having to write additional transformational programs. It would also enable each IR community to choose a metadata standard that better fits its needs and content.

In addition to providing more options for descriptive metadata, IRs should support preservation, structural, and rights metadata. DSpace currently provides a limited amount of preservation metadata in the form of filetype identification and a checksum for each bitstream submitted. There is also provenance data captured for each event such as submission, approval, and edited metadata. All of this data is created and captured by DSpace itself, but there is a strong need to make use of additional user-supplied metadata. Preservation metadata enhances the library's ability to manage activities related to a digital item's format, authenticity, and stability over time. Repository managers need to know such things as how an item was created, what has been done to it and by whom, and whether or not a digital object can still be appropriately rendered by current technology. These metadata can best be used when integrated into the same repository that stores the digital objects themselves.

## Support for Expression of Relationships between Bitstreams and Items

Out of the box, DSpace does not allow explicit relationships between items. This becomes particularly problematic when needing to identify versions of the same article, or to link between datasets and articles that refer to them. There is hope that implementation of the Open Archives Initiative Object Reuse and Exchange (OAI ORE) standard will help mediate this deficiency, but even with the implementation of ORE resource maps, this information is still lacking in the basic descriptive metadata. Richer structural metadata would enable the IR to present a clear picture of the relationship between the various components of a complex digital object. DSpace allows an item to comprise more than one bitstream, but it provides only a brief description label for each one. The user is left with only a list of discrete files without much notion of how they relate to each other. Good structural metadata would help a user navigate through a logical sequence of files that might

be unique to a particular resource. It would also establish clear relationships between supporting files such as research data and the published papers based on them.

## Author Identifier Standard

Authority control and management of authors in institutional repositories is notoriously difficult. Yet it is essential that an IR be able to accurately identify the researchers who deposit their materials for preservation, access, and rights reasons. Traditional library catalogs have struggled with author identity through the establishment of name authority records, but this approach may not extend well to the IR environment. DSpace has no such author identity component, so IR managers or communities are forced to devise their own methods to address this. The institutional repository community should continue to work with the disciplinary repositories and journal publishing community to move toward open author identifiers.

As discussed earlier, the University of Minnesota, UIUC, and Johns Hopkins University face common issues for metadata within their institutional repositories. Each institution has developed internal strategies to cope with some of the shortcomings of DSpace and the model of institutional repositories in general. Each institution has done some work to change the internal metadata structure of DSpace: Minnesota to allow more granular description of journal issues; Johns Hopkins to include required information for ETDs; and Illinois to align DSpace more closely with the Dublin Core standard. Each institution has had to compromise on enforcing consistency of metadata and on the use of controlled vocabulary. Investment by the DSpace community and by the repository community in areas like persistent author identification and increased support of a range of metadata formats could raise the quality of metadata in repositories. Even though there has not yet been a study of quality metrics for metadata in an institutional repository, metadata tool investment would help to minimize the amount of customization each institution has to do in order to produce metadata that meets their requirements.

## NOTES

1. Clifford A. Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," *ARL: A Bimonthly Report* 226 (2003), http://www.arl.org/resources/pubs/br/br226/br226ir.shtml.

2. OpenDOAR, "Directory of Open Access Repositories (OpenDOAR)," OpenDOAR, http://www.opendoar.org/.

3. Raym Crow, *The Case for Institutional Repositories: A SPARC Position Paper* (Washington, D.C.: SPARC, 2002), http://www.arl.org/sparc/bm%7Edoc/ir_final_release_102.pdf.

4. Cat S. McDowell, "Evaluating Institutional Repository Deployment in American Academe Since Early 2005: Repositories by the Numbers, Part 2," *D-Lib Magazine* 13, no. 9/10 (2007), http://www.dlib.org/dlib/september07/mcdowell/09mcdowell.html.

5. DSpace Foundation. "DSpace FAQ," DSpace Foundation, http://www.dspace.org/index.php/FAQs/.

6. Dublin Core Metadata Initiative Libraries Working Group, "DC-Library Application Profile," Dublin Core Metadata Initiative, http://dublincore.org/documents/library-application-profile/.

7. University of Minnesota Department of Applied Economics and Libraries, "AgEcon Search: Research in Agricultural and Applied Economics," University of Minnesota, http://ageconsearch.umn.edu.

8. University of Minnesota Libraries and Office of Information Technology, "University of Minnesota Digital Conservancy," University of Minnesota, http://conservancy.umn.edu.

9. University of Minnesota Libraries, "Transforming Scholarly Communication," University of Minnesota, http://www.lib.umn.edu/scholcom/.

10. Institute for Mathematics and Its Applications, "Institute for Mathematics and Its Applications," University of Minnesota, http://ima.umn.edu.

11. Illinois Digital Environment for Access to Learning and Scholarship, "IDEALS Resources and Information," University of Illinois at Urbana-Champaign, http://services.ideals.uiuc.edu/wiki/.

12. Dublin Core Metadata Initiative Usage Board, "DCMI Metadata Terms," Dublin Core Metadata Initiative, http://dublincore.org/documents/dcmi-terms/.

13. Timothy G. Donohue, "Configurable Submission System for DSpace" (paper presented at the annual Open Repositories conference, San Antonio, TX, January 23, 2007).