2009

Metadata Implementation for Building Cross-Institutional Repositories: Lessons Learned from the Liberal Arts Scholarly Repository (LASR)

Jane Costanza
Trinity University, jcostanz@trinity.edu

R. Cecilia Knight

Hsianghui Liu-Spencer

Repository Citation
Costanza, J., Knight, R.C., & Liu-Spencer, H. (2009). Metadata implementation for building cross-institutional repositories: Lessons learned fromthe liberal arts scholarly repository (LASR). *Journal of Library Metadata,* 9(1-2), 153-166. doi: 10.1080/19386380903176063

Metadata Implementation for Building Cross-Institutional Repositories: Lessons Learned from the Liberal Arts Scholarly Repository (LASR)

Metadata Implementation for Building Cross-Institutional Repositories: Lessons Learned from the Liberal Arts Scholarly Repository (LASR)

Jane Costanza, R. Cecilia Knight, Hsianghui Liu-Spencer

ABSTRACT: Institutional repositories are an exciting innovation in scholarly communication and liberal arts institutions have a unique opportunity to create repository collections that reflect their tradition. However, the challenges of cost, staffing, infrastructure, standardized metadata, and content recruitment that are part and parcel of developing institutional repositories may be daunting to individual liberal arts institutions. The idea that multiple, like-minded institutions could join forces to share their efforts, unique challenges, and maximize their efficiencies grew into the Liberal Arts Scholarly Repository (LASR). Initial steps in this collaboration included the development of a group mission and a statement of collection policies. Technical specifications and metadata best practices were developed to facilitate searching and to ensure the interoperability of the repository. This article outlines the history of the project and the process of collaborating on metadata standards.

KEYWORDS. Metadata, institutional repositories, collaboration, DSpace, OAI/PMH

Jane Costanza (jcostanz@trinity.edu) is Head of Cataloging and Associate Professor at Trinity University in San Antonio, Texas.

R. Cecilia Knight (knight@grinnell.edu) is Catalog Librarian and Associate Professor at Grinnell College in Grinnell, Iowa.

Hsianghui Liu-Spencer (hliuspen@carleton.edu) is Cataloging/Metadata Librarian at Carleton College in Northfield, Minnesota

Introduction

The Liberal Arts Scholarly Repository (LASR) is an openly accessible repository that documents the scholarship and creative activity of students and faculty of small, selective liberal arts colleges. LASR provides individual scholars at these institutions the opportunity to discover, explore, and share the ideas and experiences that are fundamental to liberal education. As participants in this project, Bucknell University, Carleton College, Grinnell College, St. Lawrence University, Trinity University, the University of Richmond, and Whitman College have joined together to address universal issues we would otherwise face individually. The unified effort of participants from multiple campuses also engenders support from our broader campus constituencies that may see the value of a cross-institutional collection concept. LASR intends to showcase significant student work – such as senior theses and collaborative projects – that demonstrates the accomplishments of students for a worldwide audience. Additionally, the institutions participating in LASR function as a community of practice exploring the benefits of cross-institutional collaboration in capturing, preserving and providing perpetual access to liberal arts scholarship through a shared repository.

In June of 2008, working with the Longsight Group for technical support, LASR initiated a DSpace[1] shared repository, which included the development of a Drupal[2] portal for access to LASR content. Initially, LASR began to develop a set of policies and practices that would support a viable shared repository, including the development of metadata standards and practices to ensure effective searching across the repository. In order to improve discovery of resources and to ensure interoperability within a cross-institutional context, it is important that

[1] http://www.dspace.org/

[2] http://drupal.org/

metadata be consistent, sufficient, and compatible. To investigate further, a metadata working group, consisting of cataloging/ metadata librarians at three of the participating institutions, was charged to explore the following topics:

- What are the current LASR participant metadata practices within institutional repositories?
- Who is vetting the metadata?
- Do they have resources to do/continue that?
- At what step is metadata evaluated?
- Do institutions allow self-submission or restrict to mediated submission?
- Is all metadata OAI compliant?
- Will metadata migrate to other platforms?
- Can metadata be harvested from other repositories?

Background of LASR

The institutions that formed the initial LASR group had participated in the National Institute for Technology and Liberal Education (NITLE) DSpace3 pilot service implemented in Spring of 2007. NITLE[4] is a membership organization that helps undergraduate centered academic institutions develop technological expertise, keep current with new technology in education, and provides a variety of types of support for these institutions to explore

[3] http://dspace.nitle.org/
[4] http://www.nitle.org/

technological applications. Each institution enrolled in the NITLE DSpace had its own section of the repository to experiment with in terms of metadata applications, workflow practices, file types, and collection structure. However, the NITLE DSpace repositories included much larger universities as well as smaller schools and the objectives across these institutions were not consistent. Therefore, in 2007, LASR made a proposal to the NITLE Southern Region Instructional Innovation Fund for partial support of a series of two programs focusing on working policies and practices essential to implementing a successful shared repository specifically for liberal arts colleges.

The first program, funded by a NITLE grant, was held at Trinity University in San Antonio, Texas in January, 2008. The members worked toward creating a shared mission, collection statement, and a general vision that would allow the LASR group to move forward. The authors agreed to review 8 to 12 other NITLE DSpace institutions' collections and submissions to gain a sense of how participants were applying metadata, what types of items were being considered for inclusion in the repository, and what the special needs might be at each campus. In addition, members agreed to survey potential LASR participants on submission and metadata practices. This allowed us to consider possible collection models, and to see what we might learn from the NITLE DSpace "community of practice".

For the second program, we developed a metadata best practices draft document in preparation for a meeting in April, 2008 in Richmond, Virginia. At this meeting, we revisited the collections document and the mission, and continued to build consensus on the creation of a portal. The meeting was facilitated by repository experts from the University of Toronto

Libraries. The Longsight Group was available to answer questions about what was possible now, or with a little effort, in the near future.

Research Methodology

The initial objective of the LASR Metadata Working Group (Working Group) was to create a metadata standard that would support successful retrieval of information and offer consistent description of materials across different collections. As a first step in this process, Working Group members reviewed the use of Dublin Core (DC) metadata in several of the schools participating in the NITLE DSpace pilot in order to evaluate the consistency in use of metadata across institutions. After analyzing these metadata samples, the Working Group developed a list of questions for potential LASR institutions. These questions focused on specific practices or issues noted in the analysis of NITLE DSpace institutions. The results of the survey were used to draft a "Metadata Best Practices" document to be used by all LASR participants.

Analysis of Metadata Practices at LASR Institutions

Working Group members analyzed a random sampling of 5 individual repository records from 11 repositories in the NITLE DSpace pilot. As a result of this review, the Working Group noticed inconsistencies in the use of certain fields as well as in the content of the fields. These included (1) inconsistency in capitalization in the title field; 2) the lack of subject heading or mixed use of Library of Congress Subject Headings (LCSH) ; 3) the inconsistency in the use of institutional affiliation; 4) the lack of rights management statements; and, 5) inconsistency in the use of physical description.

Capitalization, especially in the title field, occurred inconsistently within and across collections. Some schools followed the AACR2 standard, capitalizing the first word of the title only and any other proper nouns, while others used standard citation practice and capitalized all important words. In exploring further, we found that the schools using standard citation practice purposefully utilized it to enhance interoperability with citation management programs.

The lack of subject headings and/or keywords in records was a surprise as well. The schools that included subject headings either used LCSH or a localized schema. We found a mixture of LCSH, keywords, and localized schema within the unqualified dc.subject field. One of the possible reasons for this is that DSpace utilizes the unqualified dc.subject in the default submission template. To add a qualified subject, one would have to re-edit the record to add a dc.subject.lcsh field. This example is one record with both keywords and LCSH subject headings:

- dc.subject fish
- dc. subject dams
- dc.subject Dams – Environmental aspects – New York (State) {this is LCSH}

There was also a great deal of inconsistency in describing the institutional affiliation in records. A few schools input the institution’s name in the dc.publisher field, while others left it blank, or input a particular department within the school, for example, “Chemistry department.” Other schools used description.sponsorship or dc.contributor.other field to input either the school’s name or a departmental affiliation. In a shared repository, as in the wider world, context is essential. A department within a larger institution needs to be identified clearly as part of the larger unit.

Examples:

- dc.publisher Grinnell College
- dc.description.sponsorship Chemistry Department
- dc.description.sponsorship Trinity University
- dc.description.sponsorship Submitted in Partial Fulfillment of the Requirement for the Degree of Bachelor of Arts in Latin American Studies, History, Middlebury College
  - dc.contributor.other Trinity University (San Antonio, Tex.). Department of Biology.

Some schools input a physical description of the original form into the dc.description field, while most did not. For example,

- dc.description 120 p. Includes illustrations
- dc.description 71 leaves. Includes bibliographic references.

In addition, very few schools included a general rights statement or a statement/note of restricted viewing on those items that had restrictions on viewing the resource.

Some of the reasons for this apparent inconsistency surely lie in local decisions or constraints. Who is submitting? Who is creating metadata? What data is available to the submitter? What happens when data is migrated from other platforms? How is it mapped? Does this inconsistency affect the retrievability of materials through the portal? How does OAI/PMH compliancy fit? In our metadata best practices, what will we want to require? And what will we recommend? To answer questions developed during an analysis of metadata sample, a survey of potential LASR participants was conducted.

Survey of LASR Institutions

In April 2008, the Working Group distributed a survey on metadata and workflow to the seven LASR participating institutions. Nine individuals responded to the survey. Of the 22 survey questions, the Working Group selected 12 metadata-focused questions for consideration. Figure I. shows 10 questions with Yes/No replies; and Figure II shows two with multiple choice answers.

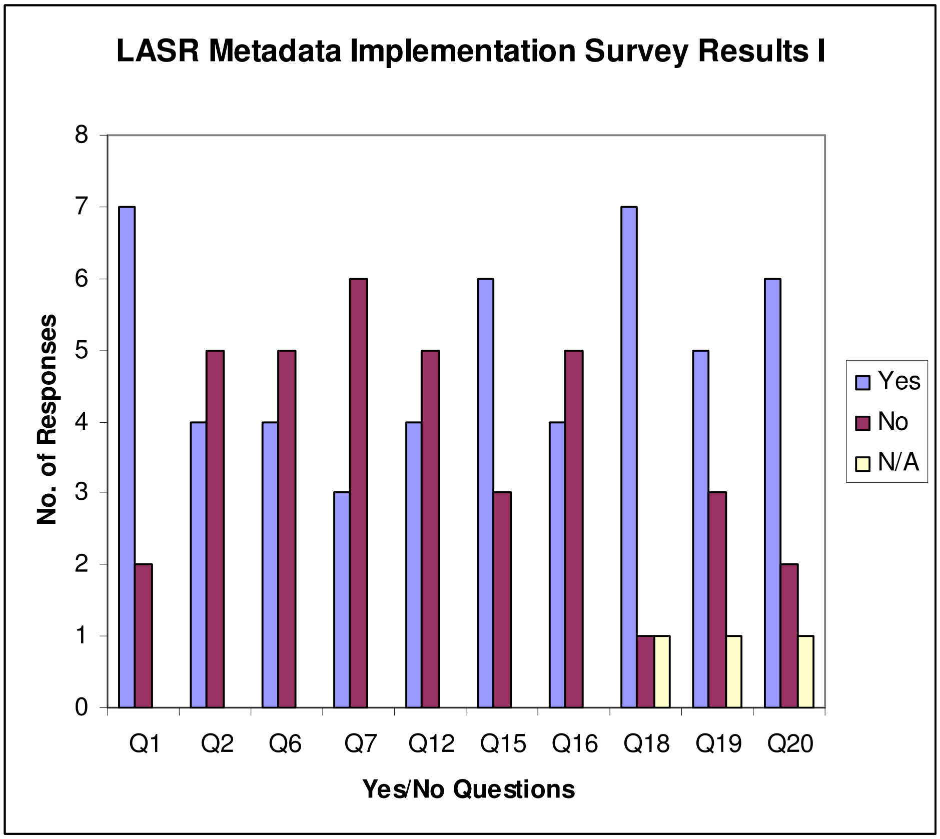

Figure I.

Q1. Is there an approval process for content being submitted to the scholarly repository?

Q2. Are there any issues with your current practice (approval process)?

Q6. Do you specify file types you will accept for repository usage?

Q7. What types of editing, proofreading, or enhancing are done to documents placed in the scholarly repository? (No-Editing, Yes-Some)

Q12. Do you offer options for materials to be restricted to a particular audience? (e.g. campus network only)

Q15. Is there an approval process for metadata?

Q16. Have you developed any specific metadata to use with the scholarly repository?

Q18. Does your metadata change by type of document? (e.g. do you use the same fields for images as for theses, etc.)

Q19. Do you use or plan to use controlled vocabularies for subjects?

Q20. Do you use or plan to use a controlled vocabulary for names, both personal and corporate?

More than half of our respondents have a submission approval process in place at their institution, but they noted some concerns with current practices. One serious issue is that the submission process is too cumbersome for faculty or students to use. Just over half of the respondents specified that the type of materials submitted were of concern. In addition, 67% (6 out of 9) of the respondents indicated that materials are placed in the repository without much editing.

An approval process for metadata is in place in 67% (6 out of 9) of the surveys. Further, 44% (4 out of 9) of the surveys indicated that specific metadata had been developed for use in the repository. The replies to question No. 18 indicates that there is a variety of types of documents and a need to alter metadata in order to truly describe the object. More than half of the respondents recognize and have a preference to utilize controlled vocabularies for subjects, personal names, and corporate names.

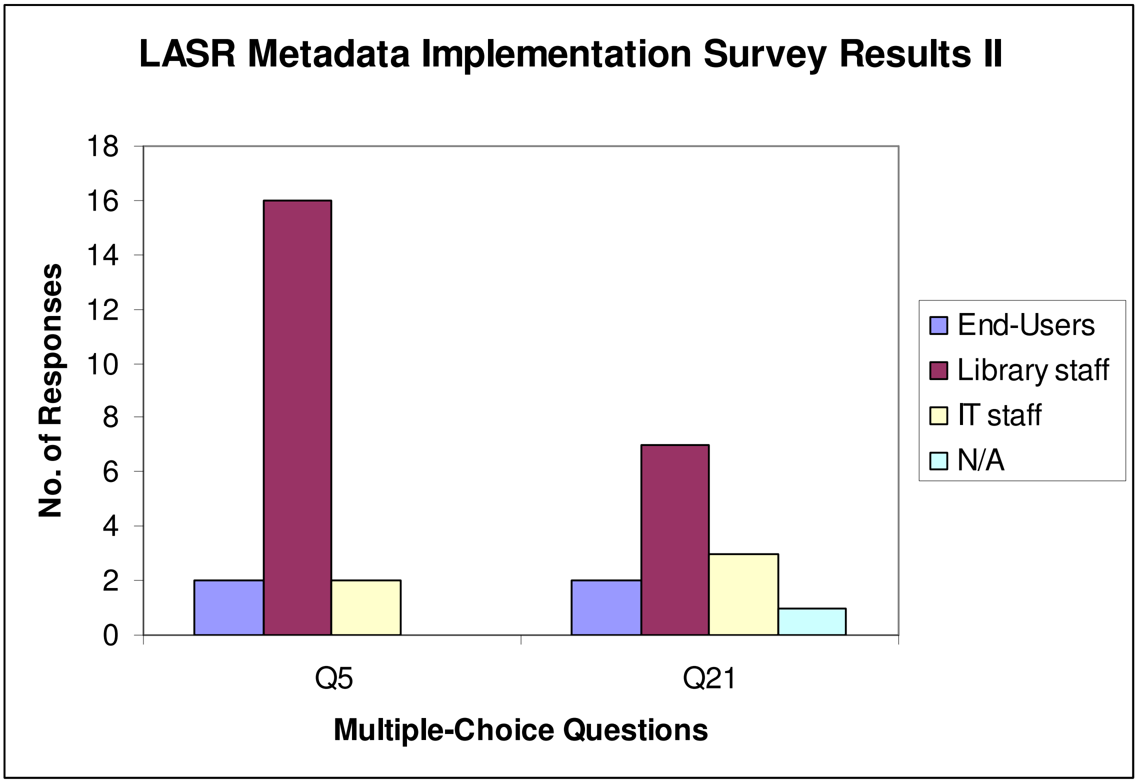

Figure II.

Q5. Who adds items to the scholarly repository?

Choose all that are authorized. (Choices include end-users, library staff, IT staff, or N/A)

Q21. Who is authorized to enter the above-mentioned metadata?

Choose all that are authorized. (Choices include end-users, library staff, IT staff, or N/A)

Figure II shows those responsible for adding items and entering metadata for new submissions. It is not surprising to learn that library staff (either professional or paraprofessional) have dominated this process, 80% and 54% respectively.

Metadata Best Practices

The initial metadata draft included some very specific fields to facilitate description and access for discrete types of data to be included in the LASR collection(s). Every Dublin Core qualifier was put to work along with some new and different qualifiers and even a few additional metadata tags utilized by other institutions. Our special metadata fields included:

- audience
- audience: educationallevel (for curricular materials)
- instructionalmethod (for curricular materials, learning objects, etc.)

However, the consultants from the University of Toronto Libraries and the Longsight Group advised the Working Group that this approach to a metadata standard, while very elegant in concept, would not succeed in a shared repository environment

The shared repository, in the virtual world, would not be limited to one repository. If an institution had materials deposited in a system on their campus, or on a remote server, that metadata would be harvested for the LASR repository and results returned along with the content stored in the LASR DSpace repository. Content could also be deposited in the LASR DSpace instance and not be part of LASR if either permissions, or, the collections policy did not permit it to be included. Clearly, a complex metadata standard would not help to facilitate collaboration and participation between LASR participants.

After a great deal of conversation within the Working Group, it was decided that a limit should be placed on the number of required elements in the metadata. Currently, the following fields are required in the metadata:

- dc:title
- dc:date.accessioned
- dc:date.issued
- dc:rights
- dc:type
- dc:relation.ispartof Liberal Arts Scholarly Repository

While the members of the Working Group were fully cognizant of the importance of metadata in searching across the collections, it is also clear that difficulties in the submission process could hamper the growth and development of the repository. The Working Group did not want standards for completeness and enhanced discoverability to prevent the submission of materials due to a difficult or frustrating process. Still, the Working Groups contends that the development of standard guidelines will ensure some reduction in variation in metadata content/structure within our community of institutions. In addition, future LASR collaborators may be more willing to participate with a simplified metadata scheme.

Metadata Vetting Practices

The LASR metadata will be vetted using workflows developed to meet the needs of each participating institution. A variety of different practices could therefore be followed for metadata creation and vetting. These could range from self-submission by contributors to the equivalent of best practices for original cataloging in OCLC (with one expert creating the metadata and a second expert checking it over before the submission becomes public).

As workflows are developed, guidelines will be made available to facilitate the process for institutions joining the LASR group, to share local practices and to encourage other institutions to share as well.

Staffing and Support

Most of the LASR institutions have small overburdened staff. Institutional repository work is often considered an “add-on” task. Technological support is also in tight supply on many campuses, with local server space and support of a local institutional repository / digital assets management system not at the top of the campus list. This collaborative effort is an attempt to extend what any of us might manage to do alone into something of real use and interest to the wider community, especially other liberal arts institutions.

The submission of materials items by individual campus community members is not expected to be overwhelming in volume. Most repositories have faced larger challenges in recruiting content than in keeping up with submissions.

Metadata Evaluation

Metadata evaluation practices will be decided upon by individual LASR institutions. Still, LASR institutions will be encouraged to maintain metadata best practices and to supply the best possible metadata for their content in order to make the repository as usable and as useful as possible.

Policies may need to be developed to allow the enhancement of metadata by experts at other LASR institutions to facilitate specific types of searching and data comparison.

Mediated Versus Self-Submission

Each institution will determine whether all submissions will be mediated or whether selected items can be self-submissions. The DSpace interface will be a factor in making this decision, since the typical interface for DSpace is not easy to navigate. There is a great deal of information that the novice individual would need to know, including formats for fields, file types that are supported by the platform, best practices for preservation, etc. With this in mind, some institutions may wish to mediate all repository submissions.

However, an effort is being made to facilitate the submission process for the novice user. This effort should reduce user frustration, reduce the time required to complete the submission process, and also gather as much information from the contributor as possible. We expect that most institutions will set up workflows that require an expert to "approve" the submission either before it becomes public, or as a later step. This would allow controlled vocabularies for subjects to be implemented, standard forms of entries for departments to be provided, and other enhancements to be made.

In order to facilitate the self-submission process, LASR member institutions are working together to revise the submission screens for use by novice submitters. A self-submission form might ask for title, author, and description or keywords. Pull-down menus could be provided that would offer standard data to be included in each field. Prompts for data would be as clear as possible to maximize success. Then, the user could deposit their documents in a “drop” or “mailbox.” After approval of the submission, a technical expert would reformat the documents if necessary before loading them into the repository.

OAI/PMH Compliancy

As mentioned earlier, after we created our first draft of the Metadata best practices document and began working with the Longsight Group to develop a list of functions for the Drupal portal, the Working Group learned that our special LASR fields were not OAI/PMH compliant and that only the Dublin Core Metadata Schema (DCMS) would be fully harvestable and harvested for this project. If we had known that all LASR content was going to be included in the LASR DSpace instance, and also believed that we would stay with this DSpace instance into the foreseeable future, we might have considered using our specific metadata fields for the powerful searching capabilities that they could have provided.

Knowing that we would be harvesting metadata at the Dublin Core level and searching the repository based on this metadata caused us to reconsider our approach to the metadata for the project. We did keep our expanded “type” list since we were assured that since the values were content that they would be harvested. The LASR type list is expanded from the 22 in NITLE DSpace (Animation, Article, Book, Book chapter, , Dataset, Learning object, Image,

Image 3-D, Map, Musical score, Plan or blueprint, Preprint, Recording acoustical, Recording musical, Recording oral, Software, Technical report, Thesis, Video, Working paper, Other) to give greater granularity to the materials that that we expect to have deposited based on our mission and collections statements and on what we observed in our examination of the NITLE DSpace collections. The LASR type list currently adds: Assessment material,Collection, Curricular materials, Event, Instructional materials, Mixed Media, Moving image, Painting, Photograph, Physical object, Presentation, Print, Sculpture, Service, Student paper, Sound, Syllabus, and Text. This expanded list of types will require a corresponding data dictionary to be included with the metadata best practices document in the LASR portal.

The Dublin Core Metadata Schema (DCMS) has 15 repeatable fields. The current OAI/PMH protocol harvests based on these 15 fields. DCMS qualified expands the metadata to provide for greater granularity in labeling and in searching. The recognized Dublin Core qualifiers are ignored, or passed over by the OAI/PMH and "dumbed down" to the 15 DC elements, but they do not cause a conflict and the content of the fields is harvested. There is reason to expect that the OAI/PMH protocol will be further developed to utilize DCMS Qualified in the future. In turn, we decided to make use of DC qualified within the LASR DSpace with the belief that the metadata could be re-harvested to make use of the searching capabilities provided for by DCMS Qualified metadata.

Since we are working within a shared, hosted, DSpace instance, it is absolutely essential that our metadata and content be available for extraction and migration to other platforms if needed. This is another reason that special fields had been set aside. Programming costs for extracting non-standard metadata would have been added to the expense for any of the participating members who later withdrew. It would also potentially create issues for moving to a new platform at a future date.

In designing a system where the collaboration revolves around metadata harvesting, rather than a system strictly based on depositing materials to one system, LASR will be able to be able to expand and grow and will welcome future members with a low threshold of preparation. This will allow our initial contributions to become the building blocks of a significant collection of liberal arts scholarship.

Conclusions

Clearly, the LASR project is both product and process oriented. The creation of a repository requires not only the creation of an open access archive, but also facilitates in the development of inter-institutional collaborative relationships with other professionals. Developing metadata best practices for the LASR shared repository provided the member institutions with an opportunity to collaboratively explore the issues of metadata best practices and the meaning of interoperability. In fact, the collaboration between the institutions to develop the infrastructure may be as important as the repository, since these relationships can be carried forward into future projects and initiatives.

As the Working Group conducted our survey of metadata usage in NITLE DSpace, we noted that there are several differences in needs for the materials submitted. Most of the materials are not published and lack the type of standardization that publishers traditionally provide. Titles are sometimes difficult to determine and the dates associated with the items are unclear, except for the date that the item is added to the repository. Because some of the materials are locally created, there may be information about the items that are not explicitly on or part of the items in any physical way; for example, the item was part of an exhibition, or won a departmental prize, or is a senior thesis. As catalogers we look to the source, while repositories cross over into the world of archives and consider context. The known, but not specified by the item, information is frequently the impetus for the inclusion of the material in the repository and must be addressed.

The scope of the materials also makes a difference in subject access. In dealing with a book we can find authorized subject headings that generally cover the topical content of the item. In dealing with a student paper about a local initiative we are challenged to deal with both the general and the very specific nature of the material and provide access related to the potential interest in the material.

Names are a particular challenge for student materials. Each campus has its own approach to dealing with student names and what the official form should be. We find that as students become published authors that we rarely have used the form of name that they settle on later in their careers.

As we worked with the metadata and struggled with finding a balance of the local and wider world we were forced to pull back. DCMS Qualified offered us more granularity for issues such as the name of a student's adviser for their project or paper. We are using dc.contributor.advisor which harvests at the contributor aspect. Faculty will not want authorship of student papers to be inaccurately attributed to them and may well resist allowing student papers to be made widely available.

As catalogers, the Metadata Working Group found any comfort with ambiguity was stretched by the LASR project. Experimenting in an area that lacks a clear standards organization has been a learning experience for all of the participants. Exploring the web sites of the players (participants, creators, developers) and monitoring blogs and various online publications is different from checking the index in AACR2 and then consulting the MARC format. There is both the possibility and likelihood that things will change and at times, change rapidly. We believe that we have chosen our direction wisely and that we have hung our hopes on developments that will come to fruition and thrive.

References

Baca, Murtha. (1998). *Introduction to Metadata: Pathways to Digital Information*. Los Angeles, CA: Getty Information Institute.

Caplan, Priscilla. (2003). *Metadata Fundamentals for All Librarians*. Chicago: American Library Association.

Collaborative Digitization Program. (2005). *Dublin Core Metadata Best Practices*. http://www.cdpheritage.org/cdp/documents/CDPDCMBP.pdf

Dublin Core Metadata Initiative. (1995). *Dublin Core Metadata Initiative*. http://www.dublincore.org/

Dunsire, Gordon. (2008). Collecting Metadata from Institutional Repositories. *OCLC Systems & Services*, 24(1), 51-58.

Hulse, Bruce, Joan F. Cheverie, & Claire T. Dygert. (2007). ALADIN Research Commons, a Consortial Institutional Repository. *OCLC Systems & Services*, 23(2), 158-169.

LASR Dspace Site. (2007). https://dspace.lasrworks.org/

LASR Portal Site. (2007). http://lasr.longsight.com/

Shreeves, Sarah L ., Jenn Riley, & Kat Hagedorn. (2007). Best Practices for OAI PMH Data Provider Implementations and Shareable Metadata. *DLF/NSDL Working Group on OAI PMH Best Practices.* Washington, D.C.: Digital Library Federation.