

# Metadata Use in OAI-Compliant Institutional Repositories

Miles Efron  
School of Information  
University of Texas, Austin  
1 University Station D7000  
Austin, TX 78712  
miles@ischool.utexas.edu

## ABSTRACT

This preliminary study examines the use of Dublin Core (DC) metadata in Institutional Repositories (IRs) that participate in the Open Archives Initiative (OAI). The study concerns evaluating IRs with respect to the quality of their indexing. In particular, I argue that in order to support successful retrieval of information, IR developers should endeavor to populate their indexes with well-structured metadata. The current study provides an initial motivation for this argument by analyzing the degree to which presently available IR metadata evidences such structure.

## Categories and Subject Descriptors

H.3.7 [Digital Libraries]: Information Systems: Information Storage and Retrieval: Digital Libraries.

## General Terms

Measurement, Reliability, Standardization.

## Keywords

Metadata; Institutional Repositories; Digital Repositories; Evaluation; Open Archives Initiative; DSPACE.

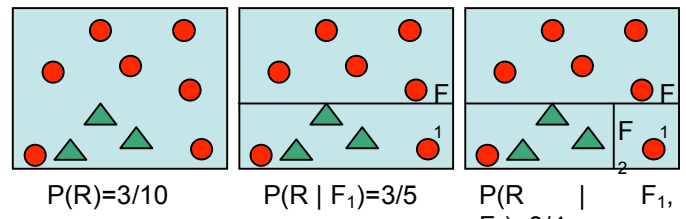
## 1. INTRODUCTION

To facilitate information retrieval, institutional repositories (IRs) capture metadata—structured information—about their collections. This study examines empirically how IRs use metadata in their collection indexing. More specifically, I analyze the collections of 23 IRs that have exposed their data through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [1]. OAI compliance requires repositories to expose their data using unqualified Dublin Core (DC) metadata [2]. I examine the degree to which each repository in my sample utilizes the 15 DC elements. I argue that better information retrieval is likely to result if a system's catalogers impose meaningful structure on their indexes by making use of multiple metadata elements. With respect to IR evaluation, my aim is to describe how a sample of established collections has deployed the Dublin Core, with an eye towards giving system developers a baseline for interpreting metadata usage in their own repositories.

## 2. MOTIVATION

It is intuitively plausible that capturing information about multiple fields of metadata would aid retrieval. But a brief example will put us on firmer ground as to why it is desirable for IR developers to encourage catalogers (whoever they may be—authors, professional librarians, etc.) to structure metadata records by using a broad palette of elements.

Figure 1. A Hypothetical Classification Problem



Imagine that a searcher wishes to retrieve relevant documents from a small repository. Figure 1 represents this scenario schematically. Further, imagine that the repository contains 10 documents: three relevant (triangles) and seven non-relevant (circles). The left panel of Figure 1 shows that without any knowledge of the user's information need, the system has a 3/10 chance of returning a relevant document from a single draw.

The middle panel shows that if the user is allowed to specify an aspect of his information need (in this case that relevant documents lie below the line  $F_1$ ), the probability of finding a relevant document increases to 3/5. By specifying two aspects of his need (below  $F_1$  and to the left of  $F_2$ ), the user can increase his chances of success again.

The lines in Figure 1 act analogously to metadata elements in search and retrieval. For instance,  $F_1$  might correspond to whether or not a document is in a particular format, while perhaps documents to the left of  $F_2$  are on a particular subject. The point is that a well organized metadata space (in the sense defined by Wasson and Wiley [4]) adds information to the documents it houses. It is desirable to use a variety of elements, then, because these elements allow searchers to articulate which portions of the space are of interest to them.

It is worth noting that the type of IR evaluation that I explore in this paper is concerned with much lower-level analysis than many established guidelines for IR metadata quality. This paper treats the deployment of particular metadata elements, a much more narrow focus than is offered by, for instance, the RLG/NARA Audit Checklist for Certifying Digital Repositories (cf. Sections C.2-C.3) [3].

### 3. DATA COLLECTION

To analyze the use of metadata in institutional repositories, I harvested a sample of repository records. My sample consisted of the holdings of the 23 IRs listed on the DSPACE website's roster of OAI-compliant collections (<http://wiki.dspace.org/OaiInstallations>) as of April 7, 2006.

It must be admitted that this sample was non-optimal on several accounts. Most obviously, many IRs do not use DSPACE, do not participate in OAI, or for other reasons, do not appear on this list. In other words, I only treat a sample of repositories that identify themselves as OAI-compliant, a sample that may put greater effort into DC metadata creation than the general population of IRs. However, due to the preliminary nature of this study, I felt that the technical ease that OAI compliance brings to the harvesting process outweighed an imperfect sampling strategy.

A more crucial criticism, however, lies in the OAI's reliance on unqualified Dublin Core as the metadata language of choice. The majority of DSPACE-based repositories use *qualified* DC. Thus this study will necessarily elide much of the structure that those collections have imposed on their indexes.

With these limitations in mind, I collected data by writing software using OCLC's OAIHarvester2, a freely available Java API [5]. The harvesting took place on April 7, 2006.

### 4. DATA ANALYSIS

Of the 23 sampled repositories, 4 returned XML that was ill-formed. Ill-formedness errors included improperly nested or unclosed elements as well as illegal character inclusion. Excluding these from my sample left 19 repositories, yielding a total of 86,522 records. The average number of records per IR was 4807. The average number of DC elements per record was 18.28, with variance=9.249.

Table 1 shows summary statistics for the fifteen Dublin Core elements. The table does not count the frequency of each element in each record, but rather it records whether or not a given element occurs in a given record. The elements are ordered by their mean average occurrence in the data. Thus the *date* element appeared at least once in 99.7% of the records I analyzed. Surprisingly, none of the repositories I queried supplied DC *source* markup. That *date* and *identifier* should appear so frequently is to be expected, as these elements can be generated automatically.

**Table 1. DC Element usage (Proportion of records with at least one entry)**

Element	Mean	Median	Var
date	0.997	0.998	0.000
ID	0.996	0.998	0.000
title	0.990	0.998	0.001
lang.	0.941	0.998	0.052
format	0.937	0.997	0.041
relation	0.865	0.987	0.081
type	0.865	0.987	0.081
creator	0.793	0.955	0.094
descr.	0.695	0.886	0.108
subject	0.647	0.660	0.105
publisher	0.532	0.643	0.128
contrib.	0.352	0.111	0.163
rights	0.187	0.001	0.090
coverage	0.058	0.000	0.045
source	0.000	0.000	0.000

**Table 2. DC Element Usage (Average number of elements per record)**

Element	Mean	Median	Var
# Tags	18.281	18.2172	9.249
# Non-0	9.4139	9.33826	0.927
date	3.1846	3	0.103
subject	2.8813	2.56169	4.712
format	2.7497	2.73333	1.249
ID	1.6942	1.69416	0.247
creator	1.4463	1.4191	0.667
descr.	1.1003	1.00956	0.659
title	1.0572	1.00358	0.017
lang.	0.9622	0.99846	0.057
type	0.9278	0.99691	0.133
relation	0.7096	0.45918	0.477
contrib.	0.6899	0.2075	0.792
publisher	0.5366	0.64286	0.125
rights	0.188	0.00119	0.088
coverage	0.1532	0	0.322
source	0	0	0

Table 1 also shows the median proportion of documents across repositories that show each element. The divergent means and medians for the elements near the middle of the distribution

suggest that the usage across repositories is not normally distributed. This would be important for parametric inferential statistical analysis, but is less crucial here, as both measures of central tendency give roughly the same ordering of DC elements.

Table 2 lists the average number of times that each element was used per record, per repository. For example, the mean average number of *date* elements per item across the 19 repositories was 3.1846. On average, records had 18.281 tags, with 9.4139 unique elements (non-zero elements) defined.

Table 2 suggests that most non-zero elements occur about once per record. Several elements, however, tended to occur more frequently. Of these the *subject* element is perhaps most interesting with respect to this study. While the average frequency of the *subject* element was 2.88, according to Table 1 nearly 35% of the polled records had zero *subjects*. Thus many records must have had significantly more than 2 or even 3 *subjects* defined. The high variance of the *subject* element in both tables supports this idea. System evaluators interested in making cross-collection comparisons, then, might pay special attention to each repository's subject information.

## 5. IMPLICATIONS

Tables 1 and 2 reflect well on the searchability of the IRs analyzed in this study. In an earlier study [6], Jewel Ward found that the general population of OAI-compliant collections had only eight DC elements defined, on average. On the other hand, the institutional repositories in this sample showed an average of over 18 elements. Whether this difference is due to greater diligence by IR metadata contributors or some other factor will be addressed in future work.

Additionally, institutional repository records were, for the most part, quite detailed. Ten out of the DC's fifteen elements were more likely than not to appear in a record in this sample. The average number of DC elements per record was over 18, with almost 10 unique tags per record on average.

But these data also leave room for improvement in IR indexing. The metadata I analyzed could be improved in two obvious ways. First, of the 23 sampled IRs, four responded to OAI *list item* requests with malformed XML or other error-causing data. Inspection of these data showed the in many cases, the offending XML was a poorly formatted character entity. Others simply returned XML with un-closed tags or asymmetrically nested elements. Assuring that an IR delivers on the open standards that it purports to embrace is an obvious point of attention during repository evaluation.

A second opportunity for improvement (or at least further analysis) lies in interrogating IR deployment of the less frequently used DC elements such as *rights*, *coverage* and *source*. The infrequency of these elements in my sample raises the question of

why they were so often omitted. Do contributors understand these elements? Does the DSPACE software support their utilization (e.g. rights are often covered by mechanisms other than DC metadata)? Do *coverage* and *source* fail to apply to IR metadata needs? These elements were clear outliers in my sample and so they invite the attention of IR evaluators.

Of course, this work provides only the coarsest analysis of metadata usage in institutional repositories. In upcoming work I plan to analyze not only the distributions of metadata elements, but also the information they convey, in an information-theoretic sense. My goal in this work will be to build models of metadata use in IRs that will help the task of repository evaluation.

Even this preliminary research, however, suggests that the matter of evaluating IRs with respect to information retrieval is important and non-trivial. In large part, I found encouraging use of Dublin Core metadata in the IRs sampled here. But the Dublin Core is a rudimentary, weakly expressive standard in comparison to other archival metadata standards such as METS. Would IR contributors and developers be able to sustain metadata quality in a more rigorous format? Also, the repositories surveyed here were relatively small. An important question in coming years will be how will users find materials in larger, more complex repositories? What type and quality of metadata will be required to support such collections? Answering these questions will be essential for the IR research community, and will require us to undertake the difficult (but productive) work of information retrieval evaluation.

## 6. REFERENCES

- [1] Lagoze, C., et al. *The Open Archives Initiative Protocol for Metadata Harvesting*. 2002 [accessed 2006 April 7]; Available from: <http://www.openarchives.org/OAI/openarchivesprotocol.html>.
- [2] *Dublin Core Metadata Element Set, Version 1.1: Reference Description*. 2004 [accessed 2006 April 7]; Available from: <http://dublincore.org/documents/dces/>.
- [3] RLG/NARA Certification Task Force. *Audit Checklist for Certifying Digital Repositories*, 2005 [accessed 2006 April 7]; Available from: <http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf>.
- [4] Wason, T.D. and D. Wiley, *Structured metadata spaces*. *Journal of Internet Cataloging*, 2000. 3(2/3): p. 263-277.
- [5] Young, J. *OAIHarvester2*. 2006 [accessed 2006 April 7]; Available from: <http://www.oclc.org/research/software/oai/harvester2.htm>.
- [6] Ward, J. A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative, in *2003 Joint Conference on Digital Libraries (JCDL'03)* 2003. Houston: ACM/IEEE.