

A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative

Jewel Ward

University of North Carolina at Chapel Hill

School of Information and Library Science

wardj@ils.unc.edu

Abstract

This research describes an empirical study of how the unqualified Dublin Core Metadata Element Set (DC or DCMES) is used by 100 Data Providers (DPs) registered with the Open Archives Initiative (OAI). The research was conducted to determine whether or not the DCMES is used to its full capabilities. Eighty-two of 100 DPs have metadata records available for analysis. DCMES usage varies by type of DP. The average number of Dublin Core elements per record is eight, with an average of 91,785 Dublin Core elements in each DP. Five of the 15 elements of the DCMES are used 71% of the time. The results show the unqualified DCMES is not used to its fullest extent within DPs registered with the OAI.

1. Introduction

The authors of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [5] require administrators of digital libraries acting as DPs to expose a set of the repository's metadata using the unqualified DCMES [3], either in addition to or in lieu of a native metadata format. This requirement is intended to facilitate cross-domain resource discovery and digital library interoperability.

Although the DCMES is an accepted standard that provides for interoperability between disparate information communities, it is not without its critics, who would like to see the OAI technical committee mandate use of a metadata format that provides more detail about resources than the unqualified DCMES [4]. The problem facing the OAI technical committee, however, is that a metadata format suitable for cross-domain description that is less complex than the unqualified DCMES does not exist, while a format more complex will not provide for low-barrier cross-domain interoperability.

Lost within the debate over whether or not use of the unqualified DCMES should be mandated by the OAI technical committee is an examination of how the DCMES is currently used within DPs. Is the DCMES used to its fullest extent? An analysis of current usage

patterns is necessary to determine the appropriateness of the unqualified DCMES as a facilitator of interoperability between and resource discovery across OAI-PMH-compliant DPs. This paper presents the results of a larger study [7].

2. Methodology

We harvested metadata between 8 May and 12 October 2002 from the 100 OAI-PMH v. 1.1 DPs that were registered on the OAI web site by 28 July 2002. We used the Perl OAI Harvester v. 1.1 [6] as the Service Provider (SP) with which to harvest the metadata from the DPs, with the exception of arXiv. The harvester software ran on a Dell Precision 530, with dual 1.7 GHz Xeon processors, 2 GB RAM, and a U160 SCSI with 10,000 RPM drives. We harvested arXiv's metadata in November 2002 from an aggregator, Celestial [1], which harvested version 2.0 of the OAI-PMH. We harvested records from 82 of the 100 DPs, but only 76 of the 82 DPs could be harvested consistently.

In order to analyze usage of the DCMES, we wrote a Perl program to count the number of records harvested from each DP, and parsed the individual elements from the content of each record in order to count the number of times a record contained each of the 15 DC elements.

We determined the type of repository either by using a web browser to issue an Identify request or by exploring the DPs web site.

3. Results

3.1. Data Providers

Based on the information we gathered from reviewing the metadata records and/or web sites of the 82 DPs, we divided the DPs into three broad categories: STI (Scientific and Technical Information), Humanities, and Combo (both STI and Humanities). By number of repositories, 33 DPs fell into the Humanities category, 27 into STI, and 22 into Combo (STI-Humanities), not adjusting for duplicate domain names.

3.2. Metadata Records

The total number of records harvested from the 82 DPs was 910,919. The average number of records per DP was 11,109. When we divided the number of records as a percentage among the three types of repositories, Humanities repositories held 43% of the records, STI repositories 31%, and Combo (STI-Humanities), 26%.

3.3. Dublin Core Metadata Elements

As a ratio against the total number of records, there was an average of eight DC elements used per record, with an average of 91,785 DC elements in each DP (Table 1). The top five DC elements used, taken as a proportion of either the total number of DC elements or the total number of records, accounted for 71% of all element usage while the least-used five elements accounted for 6% of usage. The results showed that just over half of the 82 DPs used only the creator and identifier elements for approximately half of their overall usage.

Table 1. Percentage of DC Elements by All DC Elements and All Records

Records (General Summary)			
DC Element	Number of Elements per Record	Each Element as a % of the Total Number of Elements Used (n/7,526,331)	Each Element as a % of the Total Number of Records Across All DPs (n/910,919)
creator	1,617,910	21.5	177.6
identifier	1,292,707	17.2	141.9
title	860,488	11.4	94.5
date	834,949	11.1	91.7
type	802,538	10.7	88.1
subject	495,414	6.6	54.4
description	463,833	6.2	50.9
rights	312,403	4.2	34.3
publisher	235,759	3.1	25.9
coverage	202,936	2.7	22.3
language	146,579	1.9	16.1
format	136,501	1.8	15.0
relation	47,748	0.6	5.2
contributor	39,743	0.5	4.3
source	36,823	0.5	4.0
Total:	7,526,331	100	826.2

When the 15 DC elements were cross tabulated as a percentage within each DP (Table 2), the top five elements used remained the same, but the order, from most- to least-used, changed. When calculated as a percentage within each DP, almost 99% of DPs used the title element. Calculating the least-used DC elements as a percentage within a DP changed two of the five least-used elements, compared to the previous order.

STI, Humanities, and Combo (STI-Humanities) DPs each used creator, title, identifier, and type as their top four most-used elements. STI and Combo (STI-Humanities) DPs both used date to round out the top five, matching the trend across all DPs, while Humanities DPs used rights as the fifth most-used element. The creator and identifier elements accounted for more than half of the DC elements used by STI DPs, while in Humanities DPs, title, identifier, creator and type accounted for 48% of element usage. The creator, identifier and date elements accounted for almost 60% of the total number of DC elements used by Combo (STI-Humanities) DPs.

Table 2. Percentage of DC Elements by All DPs

DPs (Summary of Crosstabs Results)				
DC Element	Number of DPs That Never Used a Particular Element Out of 82 DPs		Number of DPs That Used a Particular Element Out of 82 DPs	
	No.	%	No.	%
title	1	1.2	81	98.8
creator	4	4.9	78	95.1
date	6	7.3	76	92.7
identifier	7	8.5	75	91.5
type	10	12.2	72	87.8
subject	14	17.1	68	82.9
description	23	28.0	59	72.0
language	39	47.6	43	52.4
publisher	41	50.0	41	50.0
format	43	52.4	39	47.6
rights	46	56.1	36	43.9
contributor	50	61.0	32	39.0
source	52	63.4	30	36.6
coverage	66	80.5	16	19.5
relation	66	80.5	16	19.5

We examined univariate statistics for all variables, but other than those reported above these were not particularly informative. The chi-square values test showed no significant difference in the observed versus the expected results for the eight most-used DC elements within a DP, but $p < .05$ for the 7 least-used DC elements. We ran three Independent Samples t-tests, and paired the 3 types of DPs as three sets against each of the 15 DC elements. The results did not produce $p < .05$ in any of the three tests.

4. Discussion

The results show that the unqualified DCMES is not used to the fullest extent possible within OAI-PMH-compliant DPs. We did not expect every author or cataloguer who submits metadata to use each element at least once, but neither did we expect that two elements out of fifteen would make up half the element usage in

over half of the DPs. The implication for the OAI is that building relevant cross-resource services based on the unqualified DCMES will be difficult at best, due to its underutilization. As well, the OAI technical committee may need to reconsider mandating the use of the unqualified DCMES.

Burnett, Ng, & Park [2] studied six metadata standards and found that title, author, and identifier are common to all the schemes, and that two others – place and date – are common to five of the six schemes. The top five elements used in OAI-PMH-compliant DPs are: title, creator, date, identifier, and type, whether viewed as a proportion of total elements, total records, or total DPs. Thus, the results correlate with the results of previous studies of metadata elements, but support the results at the system level, rather than the schema level.

The trend we see across all of the results is for a very small number, whether it is DPs or DC elements, to dominate. Out of 82 DPs, five (citebase, arXiv, dlpscoll, lcoal, and uiLib) hold 85% of the metadata records. Users have a choice of 15 DC elements, but five (creator, identifier, title, date, and type) are used 71% of the time.

The fact that approximately a quarter of the DPs could not be harvested, could not be harvested regularly, or did not provide any records reflects the version 1.1 experimental phase of the OAI-PMH. Many administrators adopted the OAI-PMH early in the protocol development; thus, either the implementation was problematic or the administrators registered their repositories before they had records in place to be harvested.

The high number of Humanities and “Combination” DPs supports the belief in the information community that the OAI has long since extended beyond its e-print roots.

5. Future work and conclusions

One area for future work would be an examination of why the DCMES is so underutilized. Is the source of the underutilization problem with the unqualified DCMES itself or with the users and information professionals who supply the metadata to the DPs, or both? Although some DPs are author self-archiving, others contain metadata prepared by information professionals. Until the source of the underutilization issue is determined and resolved, it will be difficult to build relevant cross-resource services on top of the OAI-PMH using the unqualified DCMES.

In conclusion, the unqualified DCMES is not used to its fullest extent within OAI-PMH-complaint DPs. Five of the 15 elements of the DCMES – creator, identifier, title, date and type – are used 71% of the time. The least-used five elements – language, format, relation,

contributor, and source – account for 6% of usage. Just over half of the 82 DPs used only the creator and identifier elements for approximately half of their overall usage. While the reasons for the underutilization of the DCMES need to be determined, the implication of this study is that the unqualified DCMES may not be the most appropriate metadata format for the OAI technical committee to mandate for the OAI-PMH.

6. Acknowledgements

We would like to thank Gregory B. Newby for his advice and editorial feedback. We would also like to thank Michael L. Nelson for suggesting the initial area that evolved into the actual research topic as well as his advice and feedback.

7. References

- [1] Brody, T. (2002), “Celestial Open Archives Gateway”. Retrieved November 9, 2002 from <http://celestial.eprints.org/>.
- [2] Burnett, K., Ng, K., & Park, S. (1999). A Comparison of the Two Traditions of Metadata Development. *Journal of the American Society for Information Science*, 50(13), 1209-1217.
- [3] Dublin Core Metadata Initiative. (1999). Dublin Core Metadata Element Set, Version 1.1: Reference Description. Retrieved November 24, 2002, from <http://www.dublincore.org/documents/dces/>.
- [4] Lagoze, C. (2001, January). Keeping Dublin Core Simple: Cross-Domain Discovery or Resource Description? *D-Lib Magazine*, 7(1). Retrieved January 27, 2001 from <http://www.dlib.org/dlib/january01/lagoze/01lagoze.html>.
- [5] Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2002). The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-6-14, document version 2002/09/13T11:34:00Z. Retrieved July 12, 2002, available from http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html.
- [6] Suleman, H., & Fox, E. (2001, December). A Framework for Building Open Digital Libraries. *D-Lib Magazine*, 7(12). Retrieved November 26, 2002, from <http://www.dlib.org/dlib/december01/suleman/12suleman.html>.
- [7] Ward, J. (2002). A Quantitative Analysis of Dublin Core Metadata Element Set (DCMES) Usage in Data Providers Registered with the Open Archives Initiative (OAI). Unpublished master’s paper, the University of North Carolina at Chapel Hill.