Learning from Artifacts: Metadata Utilization Analysis

Dr. William E. Moen
Texas Center for Digital Knowledge
University of North Texas
Denton, Texas
940-565-3563
wemoen@unt.edu

Dr. Shawne D. Miksa
Texas Center for Digital Knowledge
University of North Texas
Denton, Texas
940-565-2445
smiksa@unt.edu

Amy Eklund
Serhiy Polyakov
Gregory Snyder
Texas Center for Digital Knowledge
University of North Texas

ABSTRACT
Describes the MARC Content Designation Utilization Project, which is examining a very large set of metadata records as artifacts of the library cataloging enterprise. This is the first large-scale examination of descriptive metadata utilization. Presents an overview of study activities and suggests the study's significance to the broader use of metadata in digital libraries.

Categories and Subject Descriptors
H.3.6 [**Information Storage and Retrieval**]: Library Automation—MARC ; E.1 [**Data Structures**]: Records—MARC

General Terms
Design, Human Factors, Management, Measurement

Keywords
library cataloging, MARC records, metadata utilization

1. INTRODUCTION
The creation of metadata records that represent information objects results in artifacts. These artifacts reflect decisions by metadata creators (human or machine). Definitions of the term *artifact* commonly describe it as something created by humans for some practical or utilitarian purpose. We assert that metadata records can be characterized as artifact of a metadata creation process. We further assert that investigating these artifacts is important in understanding significant features of the metadata creation enterprise. One important feature of this enterprise is the utilization of available elements in a metadata scheme by metadata creators. Since metadata plays many different roles and supports various activities in digital libraries (e.g., resource description, resource management, rights management, etc.), determining optimal metadata utilization is desirable: are the right number of the appropriate metadata elements being used to support applications, users, and processes in digital libraries? This paper describes a current project that is examining over 56 million metadata records in the form of Machine-Readable Catalog (MARC) bibliographic records (i.e., descriptive metadata)..

JCDL'06, June 11–15, 2006, Chapel Hill, North Carolina, USA.

2. THE RESEARCH PROJECT
The Institute of Museum and Library Services (IMLS) awarded a National Leadership Grant to the Texas Center for Digital Knowledge at the University of North Texas to investigate metadata utilization in a very large set of MARC records. The project, *Examining Present Practices to Inform Future Metadata Use: An Empirical Analysis of MARC Content Designation Utilization* (hereafter the MCDU Project) is investigating the extent of use of MARC's rich and granular metadata elements. A primary goal of this research is to develop methodologies and procedures for accurately assessing the utilization of elements in a particular metadata scheme within a specific community of practice, namely library cataloging. The cataloging enterprise includes various staff with different levels of training and knowledge, as well as different responsibilities in producing metadata records. This large-scale research project was motivated by findings in a previous analysis of a small set of MARC records in which we found that the cataloging enterprise produced metadata records using only a small set of the available elements available in the MARC descriptive metadata format. [1]. Details on the current project research questions and objectives are described in documents available on the project website [2].

Descriptive metadata schemes that have a rich set of elements to describe resources may be needed. However, there has been little investigation into the use of such a rich metadata scheme to provide empirical evidence of the utility or cost/benefits of such rich encoding schemes.

Library catalogers have a critical responsibility to organize materials and prepare them for access and use by end users. No standard has been more central to their efforts than MARC [3]. Since its initial development more than 30 years ago as a structured record for communicating bibliographic data between systems, the MARC record has evolved into a complex metadata scheme providing rich content designation (in the form of fields, subfields, and indicators, or more generally *elements*). Current MARC specifications (i.e., MARC 21) define nearly 2,000 elements available to store resource description data related to a resource [3]. This rich metadata scheme offers library catalogers a wide range of very granular elements for recording bibliographic data they create through cataloging rules. A key question is the extent to which library catalogers use the available elements when creating resource descriptions of information objects.

3. STUDY METHODOLOGY
The MCDU Project's investigation into the artifacts of the library cataloging enterprise has a number of components, including:

- Dataset of MARC 21 records
- Procedures to prepare records for analysis

- Database system to store records
- Analysis procedures

OCLC provided to the project a full set of MARC bibliographic records contained in its WorldCat database (approximately 56 million records as of Spring 2005). Although sampling from the population was considered, statistical confidence of inferences about occurrences of elements in the MARC records depends on the actual number of occurrences, or proportions, of those elements in the population. Fairly small sample sizes are required for analysis of frequently occurring elements. The lower the occurrence of a particular element, the larger the sample must be in order to guarantee an acceptable relative error of measurement. Extremely large samples (>500,000) would be required to make an inference about any content designation structures of extremely low occurrence (< 0.1%). The previous analysis of MARC records revealed that the occurrences of individual elements vary widely [2].

A key objective in the research was to conduct frequency counts of the occurrences of each and all content designation used in each record. This implied that the MARC records had to be "decomposed" at an appropriate level of granularity to enable such analysis. Any decomposition procedures had to take into consideration the size of the dataset and the available technology. Recognizing that some of our analyses required frequency counts on records that describe several fundamentally different types of materials (books, cartographic materials, sound recordings, electronic resources, etc.), the entire MCDU dataset was divided into separate subsets based on material types. The team developed a comprehensive set of questions to guide the initial analysis. *Format Content Designation Analysis:Set Profiling and Analysis Queries* details the rationale and lists the analytical questions [4].

At this point in the research (Spring 2006), we have completed the first set of analyses that includes::

- Providing summary statistics that characterize key aspects of the entire dataset and individual subsets of the dataset
- Providing summary statistics on frequency of occurrences of elements in the 20 subsets of the dataset.

Reports available on the project website contain the results [2].

4. ASSESSING METADATA UTILIZATION

The results from our initial analysis form the basis for the current stage of utilization investigations. Of particular interest is the delineation of commonly used elements. A number of initiatives, such as the Program for Cooperative Cataloging, have tried to identify best practices for library cataloging of resources and recommended the elements to constitute Core Records. We are comparing actual practice against these best practices.

Even more interesting, however, is to analyze the elements used in the dataset within the context of the *Functional Requirements for Bibliographic Records* [5] which describes data required in bibliographic records to support four user tasks: Finding, Identifying, Selecting, and Accessing or obtaining. Delsey [6] has mapped the data requirements for these user tasks to MARC elements, and we are currently comparing the occurrences of elements in our dataset with the individual elements that support one or more user tasks. We may discover that catalogers, even while apparently using only a relatively small percentage of available elements available, are frequently supplying needed data to support the user tasks. Or, we may find the opposite. Either way, the user tasks are a useful framework for assessing metadata utilization.

5. CONCLUSION

The MCDU Project is undertaking what may be the first large-scale examination of metadata utilization through an analysis of more than 56 million metadata records. The tools and procedures being used provide a reliable methodology for this analysis. Through the end of the project (August 2006), we will continue to develop an understanding of catalogers' decisions and the results of the library cataloging enterprise as reflected artifacts of that enterprise.

There is little doubt that human and machine generated metadata (of various types) is crucial to digital libraries. However, there is little in the way of agreement on what constitutes sufficient and appropriate metadata to support various activities and functions such as resource description, resource management. The MCDU Project is examining metadata utilization of descriptive metadata, and assessing that utilization in the context of commonly used elements, user tasks, and metadata quality. The methodology and tools developed by the project will be adaptable to a wide-range of metadata types.

6. ACKNOWLEDGMENTS

Support for this research has been provided by a National Leadership Grant from the Institute of Museum and Library Services.

7. REFERENCES

[1] Benardino, Penelope, and Moen, W.E. Assessing Metadata Utilization: An Analysis of MARC Content Designation Use. In *2003 Dublin Core Conference: Supporting Communities of Discourse and Practice – Metadata Research and Application* (Seattle, WA, September 28-October 2, 2003). Seattle, Information School of the University of Washington, 2003. <http://www.siderean.com/dc2003/502_Paper58.pdf>

[2] The MARC Content Designation Utilization Project [website]. 2005. <http://www.mcdu.unt.edu/>

[3] Library of Congress. *MARC 21 Format for Bibliographic Data.* Washington D.C., Cataloging Distribution Service, 1999.

[4] Moen, W.E., et al. Format Content Designation Analysis: Data Report. Frequency Counts for Books, Pamphlets, and Printed Sheets Records Created by OCLC Member Libraries (Set 01_B_nonLC). 2005. <http://www.mcdu.unt.edu/wp-content/FANDRFreqCount01BnonLCwemFinal20Dec2005.pdf>

[5] IFLA Study Group on the Functional Requirements for Bibliographic Records, International Federation of Library Associations. Functional requirements for bibliographic records: final report. 1998. <http://www.ifla.org/VII/s13/wgfrbr/finalreport.htm>

[6] Delsey, Thomas J. Functional Analysis of the MARC 21 Bibliographic and Holdings Formats. Commissioned by the Network Development and MARC Standards Office at the Library of Congress. 2001. <http://www.loc.gov/marc/marc-functional-analysis/functional-analysis.html>