

A Proposal for a National Leadership Grant

Examining Present Practices to Inform Future Metadata Use: An Empirical Analysis of MARC Content Designation Utilization

Submitted by

Dr. William E. Moen
Dr. Shawne D. Miksa

School of Library and Information Sciences
Texas Center for Digital Knowledge
University of North Texas

Table of Contents

Abstract

Narrative

Introduction

1. National Impact
2. Adaptability
3. Design
4. Management Plan
5. Budget
6. Contributions
7. Personnel
8. Project Evaluation
9. Dissemination
10. Sustainability

This Project and IMLS Priorities

References

Appendixes

- A. Schedule of Completion
- B. Project Budget (Detailed, Summary, and Budget Justification) [not included]
- C. Current, Federally Negotiated Rate for Indirect Costs [not included]
- D. Proof of Non-profit Status [not included]
- E. Applicant Organizational Profile [not included]
- F. Signed Assurance Form [not included]
- G. Resumes of Key Project Personnel [not included]
- H. Letters in Support of the Proposal [not included]
- I. Paper Reporting Preliminary Research on MARC Content Designation Use [not included]
- J. Selection Criteria for MARC Records Samples [not included]

Examining Present Practices to Inform Future Metadata Use: An Empirical Analysis of MARC Content Designation Utilization

ABSTRACT

Libraries have a critical responsibility to organize materials and prepare them for access and use by end users. No group has been more central to this endeavor than library catalogers. No standard has been more central than the MARC record. Since its development more than 30 years ago, the MARC record has evolved into a complex encoding scheme providing rich content designation (in the form of fields, subfields, and indicators). Current MARC 21 specifications define nearly 2,000 fields/subfields available to catalogers to encode bibliographic data. The resulting MARC records reflect decisions by catalogers following standards, rules, and guidelines, and these records serve as artifacts of the entire cataloging enterprise. Surprisingly, there has been no programmatic effort to analyze MARC content designation utilization, and until recently, no current publicly available data on its utilization.

The results from a recent analysis of 400,000 MARC records conducted as part of an IMLS National Leadership Grant to establish a Z39.50 interoperability testbed indicated less than 50% of nearly 2,000 MARC 21 fields/subfields occurred even once in the records, and that only 36 of the fields/subfields accounted for approximately 80% of all use. These preliminary results have sparked interest by catalogers, managers of cataloging operations, standards developers, people involved in machine generation of metadata, and others. We are proposing a research project that builds upon the initial analysis to carry out a systematic analysis of MARC content designation use in large random samples of format-specific MARC 21 bibliographic records.

The overarching research question addressed by this project is: What is the extent of catalogers' use of content designation available in MARC 21? The project has three goals: 1) Provide empirical evidence to document MARC 21 content designation use; 2) Explore the evolution of MARC content designation for patterns of availability and adoption/use level; and 3) Investigate a methodological approach to understand the factors contributing to current levels of MARC content designation use and relationships with the cataloging enterprise.

The project has the following objectives:

- Develop and implement systematic methods, procedures, and software tools to produce reliable and valid analysis of MARC 21 content designation use
- Identify core elements of bibliographic records based on the analysis of format-specific samples and comparisons with existing recommendations for core records
- Document the evolution of MARC content designation to assess the availability of content designation at specific intervals of time and subsequent rates of adoption/use
- Develop a methodological approach to identify and understand factors contributing to catalogers' use of MARC content designation.

The project will result a number of high-impact deliverables of use and benefit to the library and other metadata communities:

- Documented methods and procedures, and open source software tools to conduct reliable and valid analyses of MARC 21 content designation, which can be used by individual libraries and adapted by other metadata communities
- Information on catalogers' use of MARC 21 content designation
- Identification of "core" elements in bibliographic records based on occurrence in the large samples of records and examination of initiatives recommending core bibliographic records
- A database application containing MARC 21 content designation specifications in structured, machine-processible form, for analyzing trends and patterns
- A methodology for identifying and understanding factors influencing the use of MARC 21 content designation.

An active dissemination effort will enable us to broadly share the results of this research. The research results will provide empirically based information to assist decision makers, including those assessing cataloging policies and practices, those involved with MARC maintenance and its evolution, and those developing metadata or other schemes for bibliographic data.

Examining Present Practices to Inform Future Metadata Use: An Empirical Analysis of MARC Content Designation Utilization

Introduction

Libraries have a critical responsibility to organize materials and prepare them for access and use. The library catalog occupies a unique and special place in every library, allowing users to explore the holdings of a library and the relationships those items have to one another within a particular collection and, in some cases, across many collections. No group has been more central to this endeavor than library catalogers. No standard has been more central than the Machine-Readable Catalog (MARC) record. The bibliographic records—created by catalogers—are fundamental to users' success in finding information in a library. Library cataloging is governed by a policies and practices that have evolved in the past 100 years. Using agreed upon rules and standards, catalogers have produced millions of bibliographic records. The MARC records that result from catalogers' work can be viewed as an artifact of the entire cataloging enterprise.

The emergence of the MARC record in the late 1960s enabled a new era for sharing bibliographic data. MARC provides a standard record structure for encoding and exchanging bibliographic data. The MARC record itself has evolved in the past 30 years into a complex encoding scheme with very rich content designation (e.g., fields, subfields, and indicators) to accommodate bibliographic data for many different formats of materials. The MARC 21 Format for Bibliographic Data (Library of Congress, Network Development and MARC Standards Office, 2003b) currently defines approximately 2,000 content designation structures to encode bibliographic data. Very little research, though, has been conducted on the extent to which catalogers use the MARC record for encoding bibliographic data. An examination of MARC content designation utilization (i.e., use of the fields/subfields defined in MARC) can reveal vital information about catalogers' use of MARC and can provide insights into the cataloging enterprise.

The results from a recent analysis of 400,000 MARC bibliographic records from OCLC's WorldCat database conducted as part of an IMLS National Leadership Grant to establish a Z39.50 interoperability testbed (the Z-Interop Project; see Moen, 2003) indicated less than 50% of the fields/subfields currently defined for MARC 21 occurred even once in the dataset. Even more startling, the analysis found that a mere 36 of the fields/subfields accounted for 80% of all content designation use. Appendix I contains a copy of a paper reporting the results of this analysis; the paper was presented at the 2003 Dublin Core Conference. These initial findings have generated substantial interest in various sectors of the library community for additional analysis on catalogers' use of MARC content designation.

Multiple user groups engage with bibliographic records. First are the catalogers who compile bibliographic data and encode them into MARC records. Second are the systems developers and vendors who design computer programs for the online catalog to process and manipulate the data. Finally, end users interact with the records as they go about their tasks of finding, identifying, selecting, and accessing or obtaining resources. In this project, we are focusing on the first user group, the catalogers, who use the MARC record structure to encode bibliographic data. It is their use of the structures in the MARC record that affect both the perceived performance of the vendors' systems and end users' success in searching, selecting, and retrieving information.

We propose a research project, building upon the initial analysis, to carry out a systematic examination of MARC content designation use through a quantitative analysis of large random

samples of format-specific MARC 21 bibliographic records. The overarching research question for this project is: What is the extent of catalogers' use of content designation available in MARC 21?

1. National Impact

Managers of cataloging departments and individual catalogers face the challenge of bibliographic control for increasing numbers of information resources (in print and digital formats). The entire library cataloging enterprise—including AACR2, subject cataloging, classification, MARC encoding—strains to keep up with these challenges. The dynamic nature of information formats in the digital age and the push for making information available to users regardless of physical location combine to create seemingly impossible demands upon those responsible for bibliographic control. Budgetary constraints are affecting the resources available for cataloging activities, and managers are seeking ways to stretch reduced budgets while efficiently providing high-quality bibliographic data.

Many in the library community have been suggesting ways to simplify cataloging, and by extension, simplify the use of the MARC record. For example, the Program for Cooperative Cataloging has defined a subset of MARC content designation for bibliographic core records (Program for Cooperative Cataloging, 2003). Others in the library community see the Extended Markup Language (XML) as a replacement for the MARC record (Tennant, 2002). The move to encode bibliographic data in XML is leading to reassessments of the entire set of MARC 21 content designation. One example of this is the Library of Congress's Network Development and MARC Standards Office's (2003) effort to develop the Metadata Object Description Schema. Others are rethinking the requirements for bibliographic data given the opportunity presented by new conceptual approaches such as those suggested by the Functional Requirements for Bibliographic Records (FRBR) (International Federation of Library Associations, 1998). FRBR concepts can be used for examining and critically assessing bibliographic data to assist end users' tasks for finding, identifying, selecting, and accessing or obtaining relevant information resources. Stanford University's Lane Medical Library XML Organic Bibliographic Information System project, informed by FRBR, is restructuring and reorganizing bibliographic data (Miller, 2000). In addition, a number of researchers are investigating machine-generation of metadata and metadata transformation services (e.g., Liddy, 2003; Godby, et al. 2003).

The proposed research is needed to provide empirical data about MARC content designation use and produce information of use and interest to these initiatives, as well as catalogers and cataloging managers. Except for one study in the late 1980s (Crawford, et al., 1986), no publicly available analyses of the use of MARC content designation are available. A review of the literature and discussions with experts indicated that there have been no programmatic efforts to analyze MARC content designation utilization to inform new initiatives or to understand the factors at play that affect its utilization. This project will develop systematic methods, procedures, and tools to analyze the use of MARC content designation.

The initiatives recommending subsets of MARC 21 for "simplified" or "core" records have not had empirical data about the actual use of MARC on which to base decisions for core records. This research can provide useful information to initiatives aimed at streamlining the complexity of MARC 21 content designation. Similarly, projects pursuing XML encodings of bibliographic data will find our research results useful as they identify and select elements to include in XML schemas for bibliographic data. Without an empirical basis for selection, these projects may be developing schemas that contain elements unlikely to be utilized. "Just-in-case" metadata schemes (i.e., those

that attempt to include all elements that may possibly be used at least once) may end up as costly endeavors. An empirically based approach to choosing high-value, and critical, elements is vital.

Bibliographic records are artifacts resulting from the overall cataloging process, of which MARC is only one part. An examination of the entire cataloging enterprise, including the interaction of standards (e.g., ISBD, Paris Principles), rules (e.g., AACR2r), and format (e.g., MARC 21), is outside the scope of the proposed research. However, we believe that MARC records can serve as artifacts of the cataloging enterprise and reflect policies and practices of the enterprise. A rigorous and valid analysis of bibliographic records will provide valuable insight into the utilization patterns of MARC and, indirectly, to the whole cataloging enterprise. The proposed research will develop and test a methodology to identify factors that influence utilization of MARC content designation. An understanding of the factors can point decision makers to focal areas of the cataloging enterprise for assessment. In addition, this understanding can, in turn, inform cataloging education and future catalogers. We think the research results can lead to more effective education of future cataloging librarians and metadata specialists. Students could be introduced to systematic methods and tools for assessing metadata utilization within and beyond library catalogs.

We consider catalogers, cataloging managers, MARC standards developers, librarians involved with XML-related bibliographic data initiatives, and researchers focusing on metadata generation and metadata transformation as key audiences for the deliverables of this research. Our research will provide benefits to these audiences through a number of high-impact project deliverables:

- Documented methods and procedures, and open source software tools to conduct reliable and valid analyses of MARC 21 content designation, which can be used by individual libraries and adapted by other metadata communities
- Information on catalogers' use of MARC 21 content designation through a rigorous analysis of large samples of bibliographic records
- Identification of "core" elements in bibliographic records based on occurrence in the large samples of records and examination of initiatives recommending core bibliographic records
- A database application containing MARC 21 content designation specifications in structured, machine-processible form, for analyzing trends and patterns
- A methodology for identifying and understanding factors influencing the use of MARC 21 content designation.

An active dissemination effort will enable us to broadly share the results of this research.

2. Adaptability

This project will provide empirically grounded data of benefit to a range of decision makers, including those assessing cataloging policies and practices, those involved with MARC maintenance and its evolution, and those developing metadata or other schemes for bibliographic data. In addition, a number of the project deliverables (e.g., methods, procedures, and tools) can be adapted to enable local libraries to carry out similar analyses on local collections of bibliographic records. The software tools we develop will use common programming, scripting languages, and open source platform and applications, allowing easy portability of the software and its use by others. These deliverables will leverage the initial IMLS investment in this project and broaden the applicability and relevance of this research project.

Another project deliverable will be an empirically based list of core elements for MARC bibliographic records for different formats of materials. The list of core elements will be informed by

a comparison with existing recommendations for core records, national minimum and full record guidelines, and recent research by Delsey (Library of Congress, Network Development and MARC Standards Office, 2003a) and others using the FRBR framework of user tasks and bibliographic data to support those tasks. Catalogers and managers can use such a list of core elements to determine possible changes to local cataloging policies. Certainly the notion of “core elements” takes on new meaning given the influence of FRBR and implementations using FRBR concepts by the Research Libraries Group (2004) and OCLC (see O’Neill, 2002). A set of core elements can also affect interoperability across systems of bibliographic records. The identification of core elements in MARC 21 bibliographic records can be used by other metadata communities or those involved in automatic metadata generation (Liddy, et al., 2003) or metadata transformation (Godby, et al., 2003).

There are other metadata communities beyond the library who are already asking questions similar to those posed in this research project. For example, a recent paper by Ward (2003) examined the use of unqualified Dublin Core records. The proposed research focuses on MARC 21 records, but our methods and procedures can be adapted by other metadata communities to analyze community-specific metadata records. In addition to adapting the methodological approach, we believe that our analysis of MARC evolution may be instructive to other metadata communities as they try to balance requests for more complex metadata structures with the costs of developing and implementing those structures.

3. Design

This project has three broad goals that build upon the core research activity, which is a quantitative analysis of MARC 21 content designation use:

- Provide empirical evidence to document MARC 21 content designation use
- Explore the evolution of MARC content designation for patterns of availability and effects on its use
- Investigate a methodological approach to understand the factors contributing to current levels of MARC content designation use and relationships with the cataloging enterprise.

We have formulated a set of research questions in support of the project’s goals:

What is the extent of catalogers’ use MARC 21 content designation as indicated by analyses of large random samples of MARC records?

1. What does the empirical evidence of MARC 21 content designation use suggest about a set of common or core elements in bibliographic records per format or type of material
2. What is the relationship between the availability of new MARC content designation and its subsequent adoption and use?
3. What methodology is appropriate to identify and understand factors contributing to cataloger’s utilization of available content designation and the interplay between MARC and the entire cataloging enterprise?

Answering these questions and achieving the project goals will be accomplished through the following objectives:

- Develop and implement systematic methods, procedures, and software tools to produce reliable and valid analysis of MARC 21 content designation use

- Identify core elements of bibliographic records based on the analysis of format-specific samples and comparisons with existing recommendations for core records, national minimum and full record guidelines, and research related to FRBR's conceptual framework, user tasks, and bibliographic data to support those tasks
- Document the evolution of MARC content designation to assess the availability of content designation at specific intervals of time and subsequent rates of adoption/use
- Develop a methodological approach to identify and understand factors contributing to catalogers' use of MARC content designation.

The project's scope is limited to analyses of format-specific MARC bibliographic record samples. The project does not directly address the "quality" of the bibliographic data in the records, but focuses on the catalogers' use of MARC 21 content designation in records. We recognize that choices of content designation are based on the interplay of data, the cataloging rules governing those choices, and possible contexts of the data within the record. We consider these some of the factors affecting catalogers' choices.

Project activities coalesce into a set of project work areas, and the following table identifies and briefly describes these. See Appendix A for a preliminary listing of key activities in each work area.

Work Area	Description
Project Management	Addresses activities and tasks to ensure the successful and timely completion of the project and its evaluation.
Research Methodology and Procedures	Addresses the development of reliable and valid procedures for the research. Includes procedures for selection and preparation of samples of MARC records, development of automatic processes for analysis, and identification of appropriate statistical methods for use in the analyses. Experience from the preliminary analysis conducted as part of the Z-Interop Project will inform these methods and procedures.
Analysis of MARC 21 Content Designation	Addresses the implementation of analysis procedures on multiple samples and reporting the findings. See below for description of the anticipated samples to be used in the analysis.
Analysis of Evolution of MARC Content Designation	Addresses an examination of historical documents at the Library of Congress, Network Development and MARC Standards Office, and other sources to identify changes in available content designation from 1972 to present. Includes development of a MARC Content Designation Database (see below)
Methodology for Identifying Factors Affecting Content Designation Utilization	Addresses a range of activities to develop and test an appropriate methodology to identify and analyze factors that can affect content designation use (e.g., MARC maintenance processes, historical development, local library cataloging policies, cataloging rules and guidelines, cataloging education, etc.).
Identification of Core Elements in Bibliographic Records	Addresses activities to identify a set of core elements in format-specific records. The analysis will compare results from our analysis of samples of bibliographic records with recommendations for core records, national minimum and full record guidelines, etc.
Project Evaluation	Addresses activities related to monitoring the progress and assessing success of the research project.

Record Samples: A key ingredient for the success of this project is a set of appropriate samples of MARC 21 records on which to conduct the analyses. We have secured a commitment from OCLC (see Section 6. Contributions) to provide 16 samples of MARC 21 records randomly selected from the WorldCat database. The results from analyzing these samples form the foundation for other project activities. The total number of records available for analysis is estimated between 800,000 and 1,050,000. Extraction of the samples will be based on random sampling and criteria related either to the format of material the records represent or the date of original cataloging. Appendix J contains a description of proposed selection criteria. Each sample will contain between 50,000 and 75,000 records. The following table lists the 16 samples and the records that will be contained in each:

Sample #	Sample Type	Description of MARC 21 Records in Sample
1	Format Specific	Books, Pamphlets, and Printed Sheets
2	Format Specific	Cartographic Materials
3	Format Specific	Electronic Resources
4	Format Specific	Continuing Resources
5	Format Specific	Manuscripts (including manuscript collections)
6	Format Specific	Music (Notated and manuscript music)
7	Format Specific	Sound Recordings (musical and non-musical)
8	Format Specific	Projected Media (including digital and non-digital)
9	Format Specific	Graphic Materials (includes mixed materials, with or without archival control)
10	Format Specific	Three Dimensional Artifacts and Realia
11	Date of Creation	Records created between 1999-2003 and never revised
12	Date of Creation	Records created between 1994-1998 and never revised
13	Date of Creation	Records created between 1989-1993 and never revised
14	Date of Creation	Records created between 1984-1988 and never revised
15	Date of Creation	Records created between 1979-1983 and never revised
16	Date of Creation	Records created between 1974-1978 and never revised

The format-specific samples will allow us to determine the frequency of content designation use among similar types of records. This is important because some MARC content designation is only used for certain formats. The substantial size of each of these samples will enable us to make reliable statements about the frequency of content designation use per format of material. The samples created using date of creation of the record (where the record has not subsequently been revised) will allow us to address the project objective: *Document the evolution of MARC content designation to assess the availability of content designation at specific intervals of time and subsequent rates of adoption/use*. This analysis intersects with project activities to document MARC 21 content designation change over time. That work will provide information about available content designation during these five-year periods. Frequency analysis on the records in these five-year samples can indicate the adoption rate and use of new content designation. Key to this analysis will be the development of a database, discussed next.

A MARC Content Designation Database: Although MARC 21 content designation documentation is available in electronic form (e.g., web pages), there is not a machine-processible form of the data. A database application that contains all MARC 21 tag numbers, names, subfield codes, subfield names, and descriptions of each field/subfield is needed to support our project's automated analysis and reporting activities. Discussions with the Library of Congress and OCLC indicate no such database exists. Therefore, as part of our project we will create a database application for our research. We will re-use and extend work carried out in the Z-Interop Project's preliminary analysis that began structuring MARC 21 specifications documentation for machine processing. In addition, the information compiled in the work related to the evolution of MARC content designation will be added to this database. For example, the database will list when a specific content designation was added, deleted, or changed. This will enable analysis of trends and patterns in MARC 21 evolution. Once completed, this MARC Content Designation Database will be offered to all libraries and other researchers for their use. We believe this can be a critical tool for current and future projects.

4. Management Plan

The goals, objectives, and research questions provide the overall framework for this 21-month project. Section 3. Design, indicated that the project encompasses a number of related but distinct research activities. Effective management will involve appropriate planning, oversight, and ongoing

monitoring to ensure milestones are met and deliverables are produced. The fiscal resources for direct costs of the project, amounting to approximately \$215,885, will be managed efficiently to produce a cost-effective project. Dr. Moen will serve as overall project manager and allocate staff resources to the work areas to ensure continuing progress on multiple areas concurrently. Drs. Moen and Miksa will be responsible for leading efforts in the 7 work areas as indicated in the table below, which also shows anticipated duration of each work area and deliverables. Appendix A contains the Schedule of Completion, including cost and key activities for each work area.

Dr. Moen has managed a number of large research and development projects (see Appendix G for resumes), including an IMLS National Leadership Grant to establish and operate a Z39.50 interoperability testbed, and a two-year project to design and develop a metasearch implementation for the Library of Texas, a statewide virtual library. As principal investigator, Dr. Moen was responsible for overall project design and management, management of project funds, and staffing. UNT and SLIS will provide accounting and billing services for the project. The project will follow all appropriate UNT and SLIS administrative procedures related to staffing, payment of salaries, travel, and other aspects of the project where expenses will be incurred.

Work Area	Lead	Duration	Deliverables (preliminary list)
Project Management	Moen	Months 1-21	<ul style="list-style-type: none"> A project plan with detailed tasking, scheduling, milestones, and anticipated deliverables An evaluation plan to assess the success of the project
Research Methodology and Procedures	Moen	Months 1-6	<ul style="list-style-type: none"> A detailed research design, methods, and procedures document Sampling procedures to ensure random samples of format-specific MARC 21 records Automatic processing scripts and programs for analyses
Analysis of MARC 21 Content Designation	Miksa	Months 6-16	<ul style="list-style-type: none"> Individual and summary reports on the frequency and other statistical analyses of content designation use in the MARC 21 samples.
Analysis of Evolution of MARC Content Designation	Miksa	Months 3-12	<ul style="list-style-type: none"> An analytical report summarizing the evolution of MARC content designation A database that contains information about all MARC content designation (see below for a description of this database)
Methodology for Identifying Factors Affecting Content Designation Utilization	Miksa	Months 12-21	<ul style="list-style-type: none"> A report describing a methodology for factor identification and analysis, along with preliminary results from testing the methodology.
Identification of Core Elements in Bibliographic Records	Moen	Months 12-21	<ul style="list-style-type: none"> An analytical report indicating possible sets of core content designation elements for bibliographic records for various formats of materials, for various uses and users of bibliographic data, and for the different types of data included in a MARC record
Project Evaluation	Moen	Months 6, 12, 18, 21	<ul style="list-style-type: none"> Interim reports to IMLS to indicate project progress Summary evaluation report to assess accomplishments and impact of project

To provide external oversight of the project, we will establish a project advisory group. The group will consist of 8 to 10 experts and practitioners in the field of library cataloging, MARC standards, and metadata. The advisory group will provide guidance by reviewing project plans, activities, findings, and deliverables from the project. Members of the advisory group will also provide a means of communication with the communities they represent.

5. Budget

We are requesting \$228,207 from IMLS to cover the costs (direct and indirect) for this research. As noted in Section 6. Contributions, we have secured cost sharing contributions in the amount of \$66,614. Although personnel expenses are the primary costs for the project, funding from IMLS will be used in five project cost categories:

- Salary support for three graduate research assistants' work on project activities
- Tuition support for three graduate research assistants while working on the project
- Course release and partial summer salary for the Principal and Co-Principal Investigators
- Travel for required IMLS meetings and research at the Library of Congress
- Documents and supplies.

Appendix B, Project Budget (Detailed, Summary, and Budget Justification) provides explanation and justification for anticipated expenses.

6. Contributions

The University of North Texas (UNT) and its School of Library and Information Sciences (SLIS) will provide cost sharing contributions. A reduced indirect cost (IDC) rate on the direct costs requested from IMLS amounts to approximately \$25,743. UNT will also provide \$5,000 to purchase three workstations and software for the project. SLIS will provide the investigators three course releases from regular duties to carry out project responsibilities. (See the letter from Dean Philip M. Turner in Appendix H). SLIS support amounts to approximately \$21,364. Total support from the UNT and SLIS is \$61,614. The Texas Center for Digital Knowledge will provide office space for the research assistants on the project at no additional cost. UNT will provide web hosting space for a project website. Dr. Moen will use components of the technical infrastructure developed for the IMLS-funded Z39.50 interoperability testbed project to carry out analyses on record samples. This technical infrastructure includes a Sun Solaris server and programs developed as part of the Z-Interop Project. This re-use of the technical infrastructure leverages previous funding by IMLS.

OCLC has agreed to contribute approximately 1,000,000 MARC 21 records from its WorldCat database; these are the samples described in Section 3. OCLC has assessed the value of the services to prepare the samples at \$5,000. (See the letter from Lorcan Dempsey, OCLC Vice President for Research, in Appendix H).

7. Personnel

The project staff will consist of Dr. William E. Moen, Principal Investigator; Dr. Shawne Miksa, Co-Principal Investigator; and three graduate students enrolled in either the masters or doctoral program at the School of Library and Information Sciences. Appendix G contains resumes for Moen and Miksa.

Dr. Moen is an Associate Professor in the School of Library and Information Sciences, and a Fellow in the Texas Center for Digital Knowledge. He will have overall responsibility for ensuring this research project achieves its goals and objectives. His primary activities on the project will include research design, project management, fiscal management, and oversight of the graduate research assistants. Dr. Miksa is an Assistant Professor in the School of Library and Information Sciences, and a Fellow in the Texas Center for Digital Knowledge. She has taught extensively in the areas of information organization, cataloging, classification, information representation, and subject analysis. Dr. Miksa will be lead researcher on several of the project work areas where her expertise in MARC and cataloging will be critical.

8. Project Evaluation

The results of this research project will deepen the understanding and broaden the knowledge base of the library community, which is a foundation for changes in skills, attitudes, knowledge, and behavior. Research results, however, must be disseminated before such changes are possible, and these changes may occur well after the 21-month project has concluded. Although it will be difficult to adequately assess the outcomes and impact of this research within the project timeline, we will have mechanisms for early feedback from the key groups mentioned above. We will set up a project discussion list that will be publicly available and allow us to actively disseminate the results of the project as they are available and circulate project deliverables for community review and response. This will begin the process of deepening the community knowledge base, which can then be manifested in longer-term changes in skills, attitudes, and behavior. Since all deliverables will be available on the project website, log analysis of website traffic can be an indicator of the extent to which the deliverables are being accessed. We will actively solicit responses from interested parties in the utility of two key deliverables: software tools for conducting MARC analysis in local libraries, and the MARC content designation database.

A tangible measure of the project's success that can be directly assessed during the project timeline is the extent to which the project achieves its objectives and answers the research questions. One measure of success is the completion of the anticipated deliverables that are associated with the various objectives. In addition, since a core activity of the project is a quantitative analysis of approximately one million MARC 21 records, the results can be assessed by the rigor in the statistical methods used. An assessment focusing on the project methods and deliverables is appropriate for this project.

9. Dissemination

Project deliverables will be the form of documents, software tools, and a database application. A critical success factor for this project will be to move the products of the research out of the academy and into the community. Disseminating research findings and sharing software tools and methods will be a high priority. A primary vehicle will be a project website; this mechanism was used successfully in the IMLS-funded Z-Interop Project <<http://www.unt.edu/zinterop>>. We will establish project website at UNT, which will be the repository for all project documents and deliverables. We will also establish a public project discussion list to post project news, as well as posting updates to relevant professional discussion lists to broaden the reach of our project.

Project staff will prepare and present papers at appropriate conferences such as those of American Society for Information Science and Technology, the American Library Association, and meetings that focus on cataloging, bibliographic control, and metadata. Articles reporting the findings of the research will be prepared for submission to library and information science scholarly and professional journals.

10. Sustainability

Sustainability for a research project can be cast in terms of broad access to project findings and the extent to which the project results can lead to lasting and/or systematic change to the field. In addition, sustainability can also refer to the extent to which project activities lay a foundation and serves as a catalyst for continuing research. We will make project deliverables broadly available through the website, and the website will be accessible for at least two years after the project ends. We believe the results of this groundbreaking research on catalogers' use of MARC content designation will have immediate and long-lasting effects on framing discussions for cataloging simplification, efficiency, and effectiveness. The project can assist in developing institutional

expertise for analysis and assessment of metadata utilization through the availability of project methods, procedures, and tools for use by individual libraries to assess local cataloging practices and policies. Project research assistants will gain valuable training and expertise that will prepare them for professional responsibilities for assessing metadata utilization. The proposed research is ambitious yet is likely to be the first step in a series of research activities related to assessing metadata utilization.

This Project and IMLS Priorities

This research project directly or indirectly addresses several of the IMLS priorities in the National Leadership Grants for Libraries Program. Earlier we identified several user groups of bibliographic records, and indicated that cataloger use of MARC records was our focus. Their choices and practices early in the life cycle of bibliographic records have downstream implications for the ultimate consumer of the records, namely library users.

- **Effective use of information resources** by individuals rests in part on the availability, consistency, and quality of metadata available. Determining core elements for bibliographic records, informed by our empirical analysis and comparison with FRBR' concepts related to user tasks, can lead to recommendations of best practices for appropriate data to be recorded in bibliographic records. This may lead to improvements in end users' abilities to discover and see relationships among resources, resulting in more effective use of those resources.
- A key library service is provided through its catalog. Although MARC is not a new technology, it is not likely to be replaced in the near future as a community standard for encoding bibliographic data. Given the lack of empirical data on catalogers' use of MARC content designation, our research has the potential to inform those responsible for the cataloging enterprise, potentially leading to improved consistency and use of MARC. This in turn can lead to **enhanced library service** offered through the catalog.
- Understanding **standards** to mean community agreements about technologies, practices, and services, our research will provide information to those involved in the maintenance and evolution of the MARC standard and its specifications. Their efforts have produced a rich encoding scheme to support bibliographic control, and our research can provide data for their deliberations on changes and improvement to MARC.
- **Data on library** catalogers' use of MARC has been lacking, which our research will redress. While we are not examining the impact on end users of the MARC record and the bibliographic data it contains, a focus on catalogers' use can be a first step in understanding if changes are needed in how the MARC record is used. Such changes in library practices can have an impact ultimately on end users.
- Metadata is a key tool for **knowledge integration** as well as digital preservation and other library responsibilities. Our research into the library community's metadata as represented in MARC records can be a first step to a deeper knowledge of community practices. Efforts that use FRBR concepts for showing relationships among objects represented by MARC records is one step in the direction of knowledge integration using tools libraries already implement.

We are confident that the results of the proposed research will lead the library community to take a closer look not only at MARC use but ultimately at cataloging practices. Current activities surrounding FRBR, initiatives such as the Program for Cooperative Cataloging, and the newly drafted "Berlin Principles" put forth by the International Federation of Library Associations (2003)

all indicate we are entering a new era of bibliographic control. The proposed research will contribute to the ongoing discussions by providing new and critical data about one of the key standards for bibliographic control. We must continuously assess what we do and how we do it in order to meet the challenges of increasingly complex information objects, information environments, and most importantly, the critical information needs of our users.

References

Crawford, Walt, Stovel, Lennie, and Bales, Kathleen. (1986). *Bibliographic displays in the online catalog*. White Plains, NY: Knowledge Industry Publications, Inc.

Delsey, Tom. (1998). *The Logical structure of the Anglo-American Cataloguing Rules, Parts I and II*. Drafted for the Joint Steering Committee for Revision of AACR. Retrieved from <http://www.nlc-bnc.ca/jsc/docs.html#logical>

Godby, Carol Jean, Smith, Devon, and Childress, Eric. (2003). Two paths to interoperable metadata. 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. Retrieved from http://www.siderean.com/dc2003/103_paper-22.pdf

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

International Federation of Library Associations. (2003). *Statement of international cataloguing principles*. Draft approved by the IFLA Meeting of Experts on an International Cataloguing Code, 1st, Frankfurt, Germany, 2003. Retrieved from http://www.ddb.de/news/pdf/statement_draft.pdf

Library of Congress, Network Development and MARC Standards Office (2003a). *Functional Analysis of the MARC 21 Bibliographic and Holdings Formats*. Retrieved from <http://www.loc.gov/marc/marc-functional-analysis/functional-analysis.html>

Library of Congress, Network Development and MARC Standards Office (2003b). *MARC Standards*. Retrieved from <http://www.loc.gov/marc/>

Library of Congress, Network Development and MARC Standards Office (2003c). *Metadata object description schema*. Retrieved from <http://www.loc.gov/standards/mods/>

Liddy, Elizabeth D., et al. (2003). *MetaTest: Evaluation of metadata from generation to use*. In *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 398 – 398). Houston, TX, June 2003. Washington, DC: IEEE Computer Society.

Miller, Dick. (2000). "Bibliographic access management at Lane Medical Library: fin de millennium experimentation and bruised-edge innovation." *Cataloging and Classification Quarterly*, 30 (2/3), 139-166.

Moen, William E. (2000). Realizing the vision of networked access to library resources: an applied research and demonstration project to establish and operate a Z39.50 interoperability testbed. Retrieved from <http://www.unt.edu/zinterop>

Moen, William E. and Benardino, Penelope. (2003). 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. Retrieved from http://www.siderean.com/dc2003/502_Paper58.pdf

O'Neill, Edward T. (2002). FRBR: Functional requirements for bibliographic records; Application of the entity-relationship model to *Humphry Clinker*. *Library Resources & Technical Services* 46,4 (October). Retrieved from http://www.oclc.org/research/publications/archive/2002/oneill_frbr22.pdf

Program for Cooperative Cataloging. (2003). Introduction to the Program for Cooperative Cataloging BIBCO core record standards. Retrieved from <http://lcweb.loc.gov/catdir/pcc/bibco/coreintro.html>

Research Libraries Group. (2003). Revolutionizing the catalog: RLG's RedLightGreen project. Retrieved from <http://www.rlg.org/redlightgreen/index.html>

Tennant, Roy. (2002, October). MARC must die. *Library Journal*, 127(17), 26-28.

Ward, Jewel. (2003). A quantitative analysis of unqualified Dublin Core metadata element set usage within data providers registered with the open archives initiative. In *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 315 – 317). Houston, TX, June 2003. Washington, DC: IEEE Computer Society.

Appendix A: Schedule of Completion

This appendix contains a schedule of completion. We are proposing a 21-month research project. The projected start date is December 1, 2004; completion date is August 31, 2006. Pre-project activities will commence upon award of grant (anticipated as mid-September 2004). Pre-project activities include recruitment of graduate students for the project, acquisition of computers, securing space for project, etc.

MARC Content Designation Utilization

Total direct costs requested from IMLS are \$184,251. The following table indicates how these funds will be allocated across the project work areas. The method of computation for each activity was to take average monthly costs as reflected in the preliminary budget. The monthly costs were allocated as a percentage to each of the activities occurring in a month. A total cost for each activity is based on number of months each activity occurs. The costs reported are those that will be supported by IMLS funds; cost sharing amounts are not included in the costs reported for the work areas.

Project Work Area	Key Activities	Duration	Cost
1. Project Management	<ul style="list-style-type: none"> • Develop detailed project plan • Develop evaluation plan • Establish project advisory board • Establish project website and project discussion list 	Months 1-21	
2. Research Methodology and Procedures	<ul style="list-style-type: none"> • Develop complete research plan, including statistical methods, procedures, and software tool requirements • Develop sampling procedures for OCLC • Design, create, and test automatic processing scripts and programs for data preparation and analyses, and produce user guide for using the software tools • Design reporting forms to present findings of analyses • Disseminate work area deliverables through project website 	Months 1-6	
3. Analysis of MARC 21 Content Designation	<ul style="list-style-type: none"> • Apply methods, procedures, and software tools to samples of records • Prepare individual and summary reports from quantitative analyses of all 16 samples of records • Disseminate work area deliverables through project website 	Months 6-16	
4. Analysis of Evolution of MARC Content Designation	<ul style="list-style-type: none"> • Identify all relevant sources of information related to evolution of MARC content designation • Do onsite examination of primary materials at Library of Congress • Specify requirements for the MARC Content Designation database; solicit review of requirements and design from Library of Congress, OCLC, Research Libraries Group, and others; revise requirements and design based on feedback • Develop and test database; Create user interfaces to the database • Populate the database • Prepare analytical report summarizing the evolution of MARC content designation • Prepare user guide for database maintenance and use • Disseminate work area deliverables through project website 	Months 3-12	
5. Methodology for Identifying Factors Affecting Content Designation Utilization	<ul style="list-style-type: none"> • Develop preliminary methodology for identifying and analyzing factors and how they affect catalogers' use of MARC content designation • Test methodology on selected set of identified factors • Prepare final methodology with report of test results • Disseminate work area deliverables through project website 	Months 12-21	

MARC Content Designation Utilization

Project Work Area	Key Activities	Duration	Cost
6. Identification of Core Elements in Bibliographic Records	<ul style="list-style-type: none"> Examine results from Work Area 3 and develop metrics for identifying core elements per format of material Prepare list of core elements based on empirical analysis using project-developed metrics Examine current recommendations and specifications for the Program for Cooperative Cataloging core records and compare with project identified core elements Examine Library of Congress recommendations for National Level Records (Full Level & Minimal Level) and compare with project identified core elements Examine reports of analyses related to the Functional Requirements for Bibliographic Records to inform choice of core elements related to FRBR's four users' tasks Prepare analytical report indicating sets of core elements expressed in MARC content designation for each of the 10 formats of bibliographic records analyzed in the project Disseminate work area deliverables through project website 	Months 12-21	\$30,425
7. Project Evaluation	<ul style="list-style-type: none"> Prepare 6-month interim reports Solicit feedback on findings from Work Area 3 Solicit feedback on software tools from Work Area 2 Solicit feedback on MARC content designation database from Work Area 4 Solicit feedback on sets of core elements from Work Area 6 Prepare project evaluation report 	Months 6, 12, 18, 21	\$10,421
Total Direct Costs from IMLS			\$184,521

This project requires a parallel effort in multiple areas. Duration for some of the activities is the entire project (e.g., project management), while others are periodic (e.g., project evaluation). The following table summarizes duration of all work areas.

Work Area	Dec 04	Jan 05	Feb 05	Mar 05	Apr 05	May 05	Jun 05	Jul 05	Aug 05	Sep 05	Oct 05	Nov 05	Dec 05	Jan 06	Feb 06	Mar 06	Apr 06	May 06	Jun 06	Jul 06	Aug 06
1	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■															
3						■	■	■	■	■	■	■	■	■	■	■					
4			■	■	■	■	■	■	■	■	■										
5												■	■	■	■	■	■	■	■	■	■
6												■	■	■	■	■	■	■	■	■	■
7						■						■	■	■	■	■	■	■	■	■	■