

Shawne D. Miksa, William E. Moen, Gregory Snyder, Serhiy Polyakov, Amy Eklund
Texas Center for Digital Knowledge, University of North Texas
Denton, Texas, U.S.A.

Metadata Assistance of the Functional Requirements for Bibliographic Records' Four User Tasks: a report on the MARC Content Designation Utilization (MCDU) Project

Abstract:

This paper describes the work of the MARC Content Designation Utilization (MCDU) Project, funded by a National Leadership Grant from the U.S. federal Institute of Museum and Library Services (IMLS). The MCDU Project is analyzing approximately 56 million MARC 21 Format for Bibliographic Data records from OCLC's WorldCat database to identify actual use of the content designation available in the MARC bibliographic record. We consider bibliographic records as artifacts resulting from the overall cataloging enterprise, of which the encoding of the bibliographic data into MARC is only one part. Concepts from the Functional Requirements for Bibliographic Records (FRBR) can be used to examine and critically assess the availability of bibliographic data in these records, data meant to assist end users in finding, identifying, selecting, and obtaining relevant information resources. Overall, the MCDU Project will provide empirical data reflecting the actual use of MARC content designation structures in this set of records. Specifically, the data can be used to demonstrate how catalogers' coding of bibliographic data may or may not assist end users in these four tasks. The project is using the mapping by Delsey of MARC data elements to FRBR user tasks in this analysis. These data are crucial for making decisions about the future of MARC and may inform current work on bibliographic rules reflected in the development of the next version of cataloging rules (i.e., Resource Description and Access) by the Joint Steering Committee for the Revision of the Anglo-American Cataloguing Rules.

1. Introduction

The successful use of any knowledge organization system by end users is profoundly influenced by the information professionals' effective use of the metadata schema underlying such a system. Information professionals have long held the belief that proper organization of humankind's recorded knowledge is key to the access of that knowledge. The library catalog in particular plays a unique role in libraries as it allows users to explore the holdings and the relationships those items have to one another within a particular collection and, in many cases, in collections across the globe. There is very little empirical evidence demonstrating the extent of utilization of a major metadata encoding scheme by information professionals, especially catalogers, when creating bibliographic records. The MARC Content Designation Utilization (MCDU) Project (<http://www.mcd�.unt.edu>) seeks to address that lack of evidence. This analysis of the actual use of MARC content designation structures (CDS) can reflect on policies and practices of the whole enterprise, especially as it relates to the rethinking of the requirements for bibliographic data such as the new conceptual approaches suggested by the Functional Requirements for Bibliographic Records, or FRBR (IFLA, 1998). FRBR concepts can be used to examine and critically assess bibliographic data to assist end users' tasks of finding, identifying, selecting, and obtaining relevant information resources. MCDU will show how catalogers' coding of bibliographic data may or may not assist end users in these four tasks by expanding on Delsey's mapping of FRBR user tasks to MARC

data elements (Delsey, 2003) and by providing empirical data on the actual use of the elements in the entire OCLC WorldCat database.

2. MARC Content Designation Utilization Project

The MCDU Project, funded by a National Leadership Grant from the U.S. federal Institute of Museum and Library Services, is a systematic examination of MARC content designation use through a quantitative analysis of over 56 million bibliographic records from OCLC's WorldCat database. The overarching research question for this project is: What is the extent of catalogers' use of content designation available in MARC 21? In addition, a set of research questions more specifically guide the project:

- What does the empirical evidence of MARC 21 content designation use suggest about a set of common or frequently occurring elements in bibliographic records per format or type of material?
- What is the relationship between the availability of new MARC content designation and its subsequent adoption and use?
- 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?

OCLC provided a dump of the entire WorldCat database in spring 2005, at which time there were 56,177,383 records. This comprises the MCDU Project dataset. These records have been decomposed to facilitate analysis. The content designation structures were parsed and the resulting data were stored in a large relational database. These content designation structures are the basic units of analysis. Data preparation and management, software tools, and systematic methods and procedures developed for the project will ensure reliable and valid analyses of MARC 21 content designation use. The central component of this process was the development of parsing scripts for decomposing each MARC record's content designation structures (CDS)—fields, subfields, indicators, etc.—for storage in a database that allows the structures to be retrieved and analyzed.

The entire MCDU dataset was then divided into separate subsets based on ten material types. Moreover, in the interest of analyzing the practices of Library of Congress catalogers separately from those of other institutions and cataloging enterprises, the records were further segregated according to the agency responsible for creating them—Library of Congress or other OCLC member libraries (i.e., nonLC records). These efforts resulted in the creation of twenty separate databases, which in turn were populated according to each record's creating agency (i.e., LC or non-LC) and the type of material described (books, cartographic materials, electronic resources, etc.). This data preparation was guided by the anticipated types of analyses and frequency counts the study team carried out.

3. Relationship of MCDU Project Frequency Counts to the FRBR Model

One of the planned deliverables of the MCDU project is a list of frequently used MARC elements for bibliographic records representing different material formats, as indicated by the empirical evidence resulting from our analyses. The set of frequently used elements will be based in part on Delsey's functional analysis of the MARC 21 bibliographic format (Delsey, 2003). Ultimately, identifying frequently used elements can have practical applications for catalogers, managers, and others involved in the cataloging enterprise by informing their decisions regarding potential changes in local cataloging policies and practices. In many ways this corresponds to the FRBR study group's recommendation that the

entity-relationship analysis resulting from their study “...might also serve as a useful conceptual framework for a re-examination of the structures used to store, display, and communicate bibliographic data” (IFLA, 1998, 6). Moreover, the identification of frequently used elements in MARC 21 bibliographic records may be used to inform the practices of and research within other metadata communities, as well as those involved in automatic metadata generation, metadata harvesting, or metadata transformation.

The FRBR study group began their investigation by making no “*a priori* assumptions about the bibliographic record itself, either in terms of content or structure” (IFLA, 1998, 3). The MCDU Project, on the other hand, directly addresses the content designation structures (CDS) of MARC 21, and with the data gathered from OCLC’s WorldCat database it can test this model of basic functionality and make some conclusions as to whether or to what extent catalogers are utilizing MARC to support the model.

To identify which CDS support the four user tasks—find, identify, select, and obtain—we rely on the results of Delsey’s functional analysis of the relationship between MARC 21 format data and FRBR’s user tasks, commissioned by the Library of Congress, Network Development and MARC Standards Office (Delsey, 2003). This research provides a detailed mapping of data elements (fields, subfields, character positions in fixed fields, and indicator positions) specified in the MARC bibliographic and holdings formats to FRBR user tasks. While Delsey’s analysis examines three categories of user tasks (resource discovery tasks, resource use tasks, and data management tasks), our interests focus on the category of resource discovery, which corresponds to FRBR’s four user tasks. Detailed definitions of the four resource discovery tasks are provided in Delsey’s functional analysis of MARC 21 shown here in Figure 1.

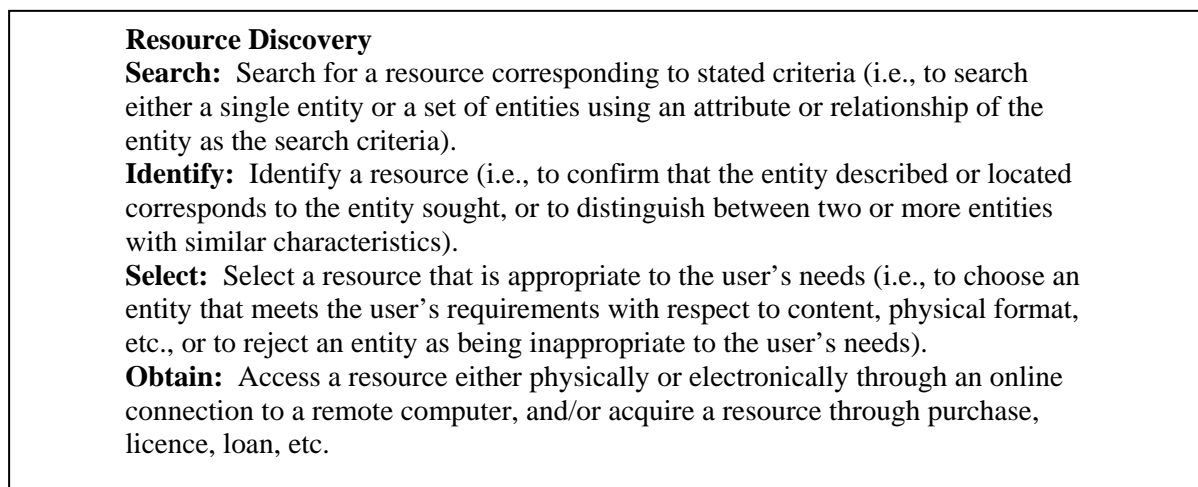


Figure 1: The Four User Tasks in Resource Discovery (Delsey, 2003, 10)

4. Methodology, Threshold Identification and Matching to FRBR User Tasks

It should be noted that the task correlating to FRBR’s find task is referred to in Delsey’s analysis as “search”; as the data elements and resource discovery task correlations we use are drawn from Delsey’s data, we retain this convention.

In addition to the identification of these user tasks, Delsey maps MARC data elements to their corresponding FRBR entities, along with the associated attributes and relationships. In our analyses, we provide frequency count data for these same elements, grouping them according to entity and user task to show how catalogers have used the CDS related to them. By providing the actual utilization of content designation for the MARC elements associated

with the user tasks and entities of the FRBR model, this project’s analyses reveal the extent to which data in the bibliographic records are available to support end users’ activities when using library catalogs for resource discovery.

In the process of developing a methodology for examining the correlation between catalogers’ use of MARC content designation and the FRBR model, there are some fundamental issues regarding the nature and structure of the project’s analyses and the data provided in Delsey’s functional analysis of MARC 21 that need to be addressed. The intent of the Delsey study was “to link MARC 21 format data with models identified in major studies that have recently been developed in the area of bibliographic control,” including the FRBR model (Delsey, 2003, 3). As that research encompasses both the *MARC 21 Format for Bibliographic Data* and the *MARC 21 Format for Holdings Data*, the data resulting from the analysis understandably includes elements from both formats. However, the MCDU project’s analyses are concerned only with catalogers’ creation of bibliographic records, and therefore do not include any of the elements defined only in the MARC holdings format. However, where there is a redundancy between the two (e.g. 022 \$a -- International Standard Serial Number, or 852 \$a Location, which are defined in both formats), those elements are included in our data.

Another issue involves the nature of the sets and subsets on which our analyses focus. The large number of bibliographic records in our dataset—more than 56 million—required the creation of twenty smaller subsets, based on the creating agency and type of material, or format, described in the records, in order to optimize processing and querying functions. To minimize discontinuities between MARC categories and the parameters of our own format-specific sets, we have chosen to focus the analyses of the elements that support the four FRBR user tasks by providing frequency counts only for the variable data fields and related subfields, as the structure of these elements is common both to all MARC material types as well as all of our format sets.

Finally, Delsey’s extensive analysis covers all of the content designators specified in the MARC formats, including indicators, the two character positions in the variable data fields whose values interpret or supplement the data found in the field (Library of Congress, 2001). In considering the efficiency of obtaining frequency counts for indicator positions from all of the 56 million records in our dataset, a comparison was made between the total numbers of elements designated by Delsey as supporting each of the four user tasks and the quantity of indicator positions that support the tasks. Judging from the small number of instances in which an indicator position supports a given user task (as shown in Table 1), the project team concluded that frequency counts for these specific content designators would not contribute significantly to the general understanding of catalogers’ utilization of MARC. Frequency counts for indicator positions are, therefore, not included in our data.

MARC 21 Bibliographic Format	FRBR Resource Discovery Tasks			
	Find (Search)	Identify	Select	Obtain
Total no. of elements (fields, subfields, fixed field positions, and indicator positions) that support a given task	454	972	375	468
Total no. of indicator positions that support a given task	0	3	7	6

Table 1. Number of Elements Supporting User Tasks, as of February 2006

An especially valuable product of Delsey's functional analysis of MARC 21 is the detailed mapping of data elements specified in the MARC bibliographic format to the four user tasks of the FRBR model, including correspondences to the FRBR entities. All of the tabular data from the efforts of this analysis are provided in an Access 2000 database that has been updated by the Network Development and MARC Standards Office to reflect recent updates or additions to the MARC format. Providing the data in this format allows us to reorder and filter the data for our needs, as well as transfer relevant data elements into electronic tables to intersect with the results of our own analyses. The variable field data elements from the MARC bibliographic format (fields 010 through 999) and their related subfields that were mapped by Delsey to the four user tasks were extracted from the Access 2000 database and combined with frequency count data showing the utilization of each CDS and separated by each of the formats specified by the project (e.g., Books, Pamphlets, and Printed Sheets; Cartographic Materials; Electronic Resources; etc.).

In order to highlight the most frequently occurring CDS, and because there are approximately 2,000 fields and subfields defined in *MARC 21 Format for Bibliographic Data*, we found it necessary to determine the threshold at which elements occur at least once in a record. The threshold is based on a statistical calculation explained in a separate document available on the project website. Briefly, the threshold calculations are based on the frequencies of use of each CDS expressed as number of records where a particular CDS is used. CDS were presented as a list ordered by the descending frequencies.

5. Partial Results of the Analysis

Focusing on an analysis of only the dataset of non-LC Books records (all records where Leader 06 value is "a" and where Leader 07 value is "a", "c", "d", or "m", and where 008/23 is not value "s"), some results of the mapping of the variable field content designation structures (CDS) within the MCDU project thresholds that support the four user tasks (find (search), identify, select, and obtain) in the non-LC dataset for Books are given in the tables below. This series of cumulative tables will show data for all four user tasks as well as the associated FRBR entities. With this data we can test the FRBR model of basic level of functionality and make some conclusions as to whether or to what extent catalogers are utilizing MARC to support the model.

Tables 2-5 show both the total number of variable field CDS, and the percentage of threshold variable field CDS within those totals, that support the user tasks as designated by Delsey's analysis. The entities included in these tables include both the primary entities as described in the FRBR model and the additional, or secondary, entities that Delsey (2003) defined as relating to *work* and *item*. For example, a *work* can result from the performance of a *task* which is part of a *project* which in turn is part of a *program*. The *project* and *program* can be funded through a *contract* or *grant* which is funded by a *corporate body* (Delsey, 2003, 58). The abbreviation "C/O/E/P" represents *concept*, *object*, *event*, and *person* which can be subjects of a work as defined in FRBR (IFLA, 1998, 15).

Entity	Find (Search)		
	No. of Variable Field CDS that are Threshold Elements	Total No. of Variable Field CDS Used in Set	% of Variable Field CDS That Are Threshold Elements
Action	0	3	0.0%
Any	2	3	66.7%
C/O/E/P	3	8	37.5%
Concept	4	12	33.3%
Corp. Body	9	36	25.0%
Curriculum	0	3	0.0%
Event	2	13	15.4%
Expression	1	28	3.6%
Item	5	17	29.4%
Manifestation	11	64	17.2%
Person	10	22	45.5%
Place	3	22	13.6%
Work	11	151	7.3%
Total:	61	382	16.0%

Table 2: Threshold Percentages That Support the Find (Search) User Task and Associated Entities in CDS Found Within Book Records Created by OCLC Member Libraries (nonLC).

The Find (Search) task (Table 2) is supported by 61 field/subfields above the calculated threshold. Stated another way, only 16% of the 382 variable fields utilized by catalogers support this user task within Book records created by OCLC member libraries (non-LC). Furthermore, taking all the entities that are specifically associated with the Find (Search) task and the secondary items associated with those (as designated in Delsey’s 2003 analysis) we can say that 12% of the 382 variable fields utilized by catalogers support this user task within Book records created by OCLC member libraries (non-LC). This set of Book records accounts for 34.5 million of the 56 million records in OCLC’s WorldCat database.

Entity	Identify		
	No. of Variable Field CDS that are Threshold Elements	Total No. of Variable Field CDS Used in Set	% of Variable Field CDS That Are Threshold Elements
Action	0	3	0.0%
Any	2	3	66.7%
C/O/E/P	2	6	33.3%
Concept	4	12	33.3%
Contract	0	1	0.0%
Corp. Body	10	50	20.0%
Curriculum	0	3	0.0%
Event	2	15	13.3%
Expression	3	47	6.4%
Grant	0	1	0.0%
Item	5	23	21.7%
Manifestation	29	281	10.3%
Person	10	22	45.5%
Person/Corp.	0	2	0.0%
Place	2	16	12.5%
Program	0	1	0.0%
Project	0	1	0.0%
Study program	0	1	0.0%
Task	0	1	0.0%
Work	13	194	6.7%
Total:	82	683	12.0%

Table 3: Threshold Percentages That Support the Identify User Task and Associated Entities in CDS Found Within Book Records Created by OCLC Member Libraries (nonLC).

The Identify user task is supported by 683 variable field CDS, of which only 82 (12%) CDS are above the calculated threshold. If we compare this with the FRBR model's basic level of functionality we can see that *work*, *expression*, and *manifestation* are supported by only 6.6%, a little more than half of the 12% of CDS above the threshold. The secondary entities associated with *work*—*contract*, *curriculum*, *grant*, *program*, *project*, *study program*, *task* do not have significant numbers of CDS to include in this percentage.

Entity	Select		
	No. of Variable Field CDS that are Threshold Elements	Total No. of Variable Field CDS Used in Set	% of Variable Field CDS That Are Threshold Elements
C/O/E/P	1	2	50.0%
Corp. Body	1	1	100.0%
Event	0	5	0.0%
Expression	6	28	21.4%
Item	0	1	0.0%
Manifestation	16	104	15.4%
Place	1	7	14.3%
Study program	0	1	0.0%
Work	3	23	13.0%
Total:	28	172	16.3%

Table 4: Threshold Percentages That Support the Select User Task and Associated Entities in CDS found Within Book Records Created by OCLC Member Libraries (nonLC).

The Select user task is supported by a total of 172 variable field CDS, of which only 28 (16.3%) CDS are above the calculated threshold. This user task is also associated with *work*, *expression*, and *manifestation* as outlined in the FRBR model's basic level of functionality, and taken together these account for the majority of the 16.3% of above-threshold CDS.

Entity	Obtain		
	No. of Variable Field CDS that are Threshold Elements for the Set	Total No. of Variable Field CDS Used in Set	% of Variable Field CDS That Are Threshold Elements
Action	0	1	0
Any	2	3	66.7%
Corp. Body	1	9	11.1%
Expression	2	7	28.6%
Item	5	18	27.8%
Manifestation	24	250	9.6%
Work	1	7	14.3%
Total:	37	295	11.9%

Table 5: Threshold Percentages That Support the Obtain User Task and Associated Entities in CDS Found Within Book Records Created by OCLC Member Libraries (nonLC).

Finally, in Table 5 we can see that in this set of Book records the Obtain user task is supported by a total of 295 variable field CDS, of which only 11.9% are above threshold. The

concentration of this percentage centers on fields that are associated with the *manifestation* entity, which corresponds to the relationship between the Obtain task and *manifestation* as designated in the FRBR model's basic level of functionality.

6. Conclusions and Summary

By pairing the MARC content designation structures associated with the four user tasks with the frequency count data from the MCDU Project's analysis, we are able to add another layer to Delsey's functional analysis of MARC 21, showing the correspondence of its actual utilization to the FRBR model. However, this raises the question of what these levels actually mean in the overall picture of cataloger utilization of MARC 21. For instance, do we know how many content designation structures are needed to support a user task? Does a higher percentage of CDS used in a record necessarily mean there is stronger support for a task? Further study is needed to explore these types of questions.

This paper has endeavored to show how catalogers' coding of bibliographic data may or may not assist end users' tasks for finding, identifying, selecting, and obtaining relevant information resources. The results from the current research are important contributions to discussions about the future of MARC and bibliographic rules such as the current work on Resource Description and Access (i.e., AACR3) by the Joint Steering Committee for the Revision of the Anglo-American Cataloguing Rules. The MCDU Project team has developed a methodology to identify factors that influence utilization of MARC content designation and query software that allows for deeper levels of analysis (e.g., the correlation between descriptive cataloging form, encoding levels and format; the distinction between Library of Congress created records and non-LC created records within formats, etc.). An understanding of the factors can point decision makers to focal areas of the cataloging enterprise for assessment, especially as it relates the FRBR. This understanding can, in turn, inform cataloging education and future catalogers, both nationally and internationally.

References

- Delsey, Tom. (2003). *Functional analysis of the MARC 21 bibliographic and holdings formats, Second revision*. Prepared for the Network Development and MARC Standards Office, Library of Congress. Retrieved February 15, 2006, from <http://www.loc.gov/marc/marc-functional-analysis/functional-analysis.html>
- International Federation of Library Associations, IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). *Functional requirements for bibliographic records: final report*. Retrieved February 15, 2006, from <http://www.ifla.org/VII/s13/frbr/frbr.pdf>
- Library of Congress, Network Development and MARC Standards Office. (2001). *MARC 21 Format for Bibliographic Data, Update no. 2*. Washington D.C.: Library of Congress Cataloging Distribution Service.
- Library of Congress, Network Development and MARC Standards Office. (2006). Access 2000 database filename: FRBR_Web_Copy.mdb, updated 07 February 2006 [Data file]. Retrieved February 27, 2006, from <http://www.loc.gov/marc/marc-functional-analysis/functional-analysis.html>

Dr. Shawne D. Miksa
School of Library and Information Sciences
University of North Texas
P.O. Box 311068
Denton, TX 76203-1068 U.S.A.
Submitted February 28, 2006