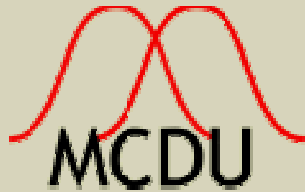


MARC Content Designation and Utilization

Inquiry and Analysis



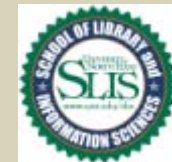
Catalogers' Use of MARC: Learning from Artifacts through Metadata Utilization Analysis

William E. Moen

<wemoen@unt.edu>

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

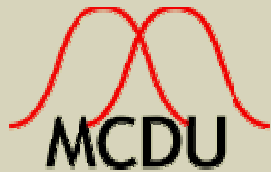
Research funded by a National Leadership Grant from the Institute for Museum and Library Services. Additional support provided by the University of North Texas School of Library and Information Sciences and the Texas Center for Digital Knowledge.





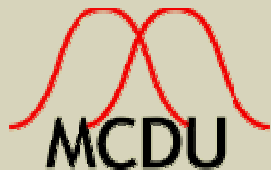
Metadata – Assumptions

- Essential in library applications
- Variety of metadata schemes
- Variety of functions and services supported
- Increasing use of machine-generated metadata
- Role of handcrafted metadata needs continuing review and assessment
- Research on use of metadata schemes can provide empirical data for decisions



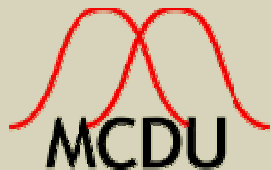
Metadata record as artifact

- Metadata creation as process
- Resulting metadata records as artifacts of the process
- Artifact reflects decisions, policies...
- Artifact can be investigated to understand metadata utilization decisions
 - Decisions to use or not use available metadata elements



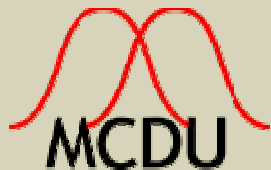
The MCDU Project

- **MARC Content Designation Utilization**
 - Provide empirical evidence of catalogers' use of MARC content designation
 - Identify commonly used elements of bibliographic records
 - Contribute to community discussion about core elements in MARC bibliographic records
 - Explore the evolution of MARC content designation
 - Develop research approach to understand the factors influencing levels of MARC content designation use



Project deliverables

- Reports containing results of analysis of utilization
- Reports addressing commonly used elements
 - Across formats
 - In context of national recommendations (e.g., BIBCO)
 - In context of FRBR user tasks
- HistoriMARC
 - Database of MARC historical information about evolution of fields/subfields, etc.
 - Enable analysis of patterns of adoption and utilization
- A methodology to understand factors influencing catalogers' use of MARC
- Software tools and methods for others to use



Why study MARC utilization?

- Standard record structure for exchange of descriptive and other types of metadata
- Evolved since late 1960s as key mechanism for sharing metadata among libraries
- Metadata record with approximately 2,000 elements available
 - Approximately 200 fields
 - Approximately 1800 subfields or other structures
- To what extent is the richness/complexity exploited and to what purpose?
 - See Goldsmith and Knudson regarding Los Alamos Research Library choice of a metadata scheme

Although often disparaged or dismissed in the library community, the MARC standard, notably the MARCXML standard, provides surprising flexibility and robustness for mapping disparate metadata to a vendor-neutral format for storage, exchange, and downstream use.



Richness of MARC

MARC 21 Field Groups	Currently Defined (MARC 21 or OCLC MARC Bib.)	MARC 1972
00x	6	3
0xx	311	28
1xx	76	40
2xx	176	15
3xx	155	4
4xx	45	37
5xx	344	8
6xx	235	66
7xx	477	41
8xx	249	36
9xx	16	
TOTAL	2074	278



Data set and preparation

- 56,177,383 MARC 21 bibliographic records from OCLC WorldCat
- Decomposed the records to store in MySQL
 - [Parsing tool](#)
 - 82 hours to process and load records
 - 295 GB final database size (with indexing)
- Structuring of decomposed records align with analytical questions
- [Sample decomposed record](#) (data fields only)
- All documented on project website:
<http://www.mcd�.unt.edu>

	Number	%	Number	%	Total
MCDU Project Dataset	56,177,383	100			
	LC-Created Records		Non-LC-Created Records		
MCDU Project Dataset by LC/nonLC	8,713,665	15.5	47,463,718	84.5	56,177,383
Books Records	7,595,887	13.5	34,546,200	61.5	42,142,087
Cartographic Materials	242,132	0.4	596,642	1.1	838,774
Electronic Resources	39,879	0.1	871,881	1.6	911,760
Continuing Resources	388,332	0.7	2,193,009	3.9	2,581,341
Manuscripts	11,471	0.02	4,390,970	7.8	4,402,441
Music	109,249	0.2	1,167,654	2.1	1,276,903
Sound Recordings	241,940	0.4	1,702,342	3.0	1,944,282
Projected Media	22,088	0.04	1,415,606	2.5	1,437,694
Graphic Materials	62,625	0.1	506,401	0.9	569,026
Three-Dimensional Objects and Realia	62	0.0001	73,013	0.1	73,075



Analytical questions

- What is the average length of the records?
- What is the average length of the records?
- What is the frequency of types of records?
- What is the status of the record?
- What is the frequency of encoding levels?
- What is the frequency of the descriptive cataloging forms?
- What are the total occurrences of all control and data fields?
- What are the total occurrences of each control and data field?
- What are the total occurrences of all subfields?
- What are the total occurrences of each subfield?
- What percentage of records contains at least one occurrence of each control and data field?
- What percentage of records contains at least one occurrence of each subfield?



Categories of questions

- General profile of the dataset (e.g.):
 - What is the distribution of records by Type of Record?
 - What is the distribution of records by Encoding Level?

- Occurrences of content designation structures:
 - What is the number of total occurrences of all control and data fields and how many unique field tags are used?
 - In how many and in what percentage of records is each unique field/subfield combination used at least once?



Example results

- 7,595,887 LC-created records in dataset
- Type of Record: Book, Pamphlets, and Printed Sheets
- Total number of unique fields occurring: 167
- Number of fields accounting for 80% of occurrences: 14 fields (8.3%)
- Number of fields accounting for 90% of occurrences: 21 fields (12.6%)
- Approximately 110 fields (66%) occur in less than 1% of all records

[Note: Fields are cataloger-supplied, not system-supplied]



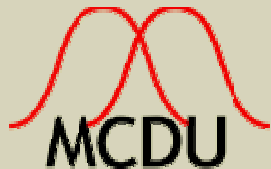
Project implications

- Empirical basis for decisions about core elements in a metadata scheme
- Profiling repositories of metadata for aggregators
- Reuse of methodologies and tools to analyze local utilization levels
- Contributions to changes in cataloging rules, practices, policies, and standards



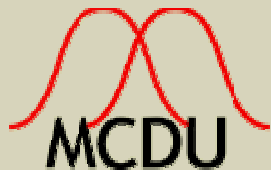
Making sense of the numbers

- Frequency counts provide raw but informative data
- Determining commonly occurring elements
 - A concept of ***threshold***
 - Average based on total number of occurrences and number of CDS
 - Comparing to recommended core records
 - ***MARC 21 Format for Bibliographic Data: National Level Record – Bibliographic Full Level & Minimal Level***
<<http://www.loc.gov/marc/bibliographic/nlr/>>
 - Comparing to recommendations for national level records
 - Comparing the FRBR user tasks data
- See handout



Element use and FRBR tasks

- FRBR describes four user tasks
 - Find
 - Identify
 - Select
 - Obtain
- Are library catalogers providing data to support FRBR tasks?
- Delsey mapped these tasks to MARC CDS for FRBR entities



FRBR user task: Find (search)

- Approximately 460 fields/subfields can support this task
- In MCDU dataset, only 59 (13%) of these occur at or above the threshold of use (i.e., commonly occurring) in OCLC book records



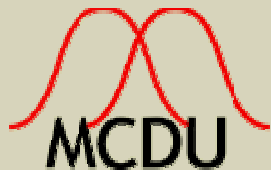
Questions for consideration?

- Can MCDU results inform your local practices?
- What about the 62% of all fields used in less than 1% of the records?
- What is needed in a bibliographic record?
 - Support for the four user tasks?
 - Management of information resources?
 - How do your systems use the infrequently used data?



Questions for consideration?

- Can you argue persuasively for the cost/benefit of your existing practice?
- Should the focus be on high-value, high-impact, high-quality data in a few fields/subfields?
 - Can you identify these few fields/subfields?
 - What would it mean for costs of cataloging?
 - What would this mean for training?



Confluence for change

- Within library community...
 - Influence of FRBR concepts and model for metadata
 - Resource Description and Access (RDA)
 - Next generation “MARC”
 - Re-examination of library catalog and its position within the landscape of resource discovery tools
 - Development of a bibliographic metadata element set



References

- MARC Content Designation Utilization Project
 - <http://www.mcd�.unt.edu/>
- Assessing Metadata Utilization: An Analysis of MARC Content Designation Use
 - http://www.unt.edu/wmoen/publications/MARCPaper_Final2003.pdf.pdf
- Goldsmith and Knudson. 2006. Looking back, Looking forward: A metadata standard for LANL's aDORe repository
 - <http://doi.acm.org/10.1145/1141753.1141814>