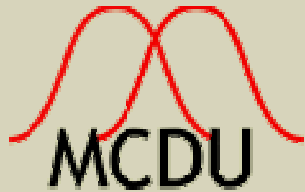


# MARC Content Designation and Utilization

## *Inquiry and Analysis*

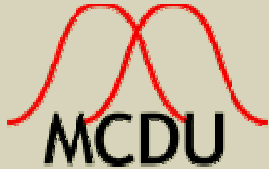


# Informing the Future of MARC

## *An Empirical Approach*

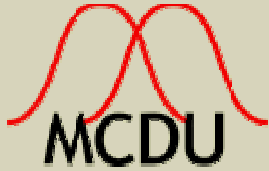
Research funded by a National Leadership Grant from the Institute for Museum and Library Services. Additional support provided by the University of North Texas School of Library and Information Sciences and the Texas Center for Digital Knowledge.





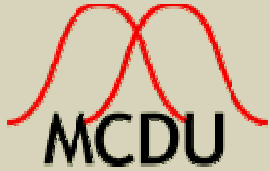
# Presentation introduction

- A presentation in three parts
  - Overview of research project and findings
    - William E. Moen, *Assistant Professor, Texas Center for Digital Knowledge, University of North Texas*
  - Interpreting findings and the community's response
    - Shawne D. Miksa, *Assistant Professor, Texas Center for Digital Knowledge, University of North Texas*
  - MARC Futures
    - Sally H. McCallum, *Chief, Network Development and MARC Standards Office, Library of Congress*
- Thanks to our sponsors
  - Backstage Library Works
  - Association for Library Collections & Technical Services (ALCTS)
  - MARBI
- Special thanks to UNT SLIS Research Assistants



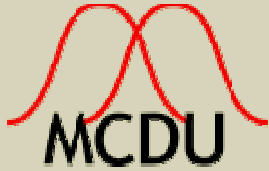
## To start...

- Discussion of the future of MARC is only partially about MARC
  - The broader digital information landscape
  - Technologies
  - Cataloging practices
  - The diminishing market share of:
    - Libraries in the information marketplace
    - Library catalogs as a resource discovery tool



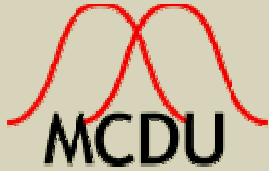
# Confluence of changes

- 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
    - Designing the Future -- Library Systems and Data Formats
    - Next Generation Catalogs for Libraries
- Outside the library community...
  - The Web and its standards
  - Metadata landscape for resource description and discovery
  - Metadata interoperability, transformation, and reuse



# When we say MARC?

- Record format
  - Defined by ISO 2709/ANSI Z39.2
  - Structural elements of the format
- Metadata scheme
  - Defined by MARC 21
  - Fields, subfields, indicators and their semantics as metadata elements



# MARC's richness

- Metadata record with approximately 2,000 elements available
  - Approximately 200 fields
  - Approximately 1800 subfields or other structures
  - See Handout Table 1 for evolution of richness
- To what extent is the richness/complexity exploited and to what purpose?

*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.*

Goldsmith and Knudson. 2006



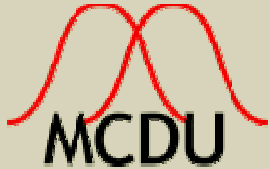
# Why study MARC?

- IMLS-funded Z39.50 Interoperability Testbed Project (2001-2003)
  - Assessing Metadata Utilization: An Analysis of MARC Content Designation Use
    - 4% of all fields/subfields account for 80% of all occurrences
    - 96% of all fields/subfields account for 20% of all occurrences
- The community's key record structure/metadata scheme for recording/exchanging descriptive and other types of bibliographic metadata
- No large scale publicly available analysis of MARC use by catalogers



# IMLS National Leadership Grant

- *Examining Present Practices to Inform Future Metadata Use: An Empirical Analysis of MARC Content Designation Utilization*
- Metadata record as artifact of the cataloging enterprise
  - Artifact reflects decisions, policies...
  - Artifact can be investigated to understand metadata utilization decisions
- Understanding current practice to inform:
  - Changes to standards
  - Changes to practice
- Focus on structural use – not data quality nor end-user benefits

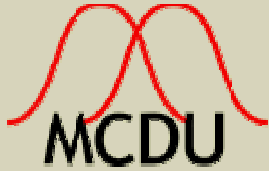


# The MCDU Project

## ■ **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

**For more information, go to: <http://www.mcd�.unt.edu>**



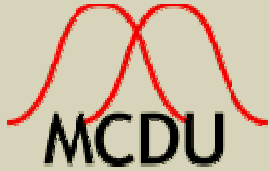
# Project deliverables

## ■ ***Reports***

- Results of frequency analysis of utilization
- Addressing commonly used elements across formats
  - In context of national recommendations (e.g., BIBCO)
  - In context of FRBR user tasks

## ■ ***HistoriMARC***

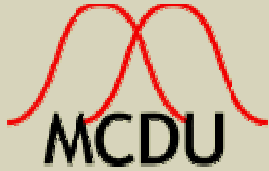
- Database of historical information about evolution of fields/subfields, etc.
- Enable analysis of patterns of adoption and utilization
- Software tools and methods for others to use
- A methodology to understand factors influencing catalogers' use of MARC



# Data set and preparation

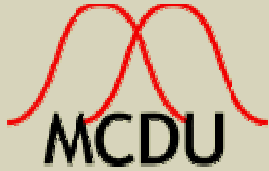
- 56,177,383 MARC 21 bibliographic records from OCLC WorldCat (as of May 2005)
- Decomposed the records to store in MySQL
  - See Handout Figure 1
  - Structured decomposed records to align with analytical questions
- Partitioned decomposed data into 20 databases
  - Type of record (ten)
    - See Handout Table 2
  - Source of cataloging (LC or non-LC)
    - See Handout Table 3

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



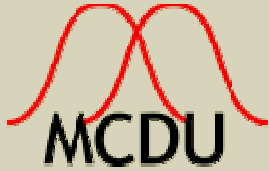
# Categories of questions

- 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?
  - What is the distribution of records by Descriptive Cataloging Form?
    - See Handout Tables 4 & 5
- Occurrences of content designation structures:
  - In how many and in what percentage of records is each unique field/subfield combination used at least once?
    - See Handout Tables 6 & 7
  - What is the number of total occurrences of all control and data fields and how many unique field tags are used?



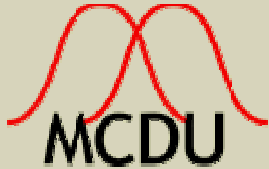
## Example results

- 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%)** [cataloger-supplied, not system-supplied]
- Number of fields accounting for **90%** of occurrences: **21 fields (12.6%)** [cataloger-supplied, not system-supplied]
- Approximately 110 fields (66%) occur in less than 1% of all records
  - See Handout Tables 8, 9, & 10 for results from analysis of OCLC member created records for **Book, Pamphlets, and Printed Sheets**



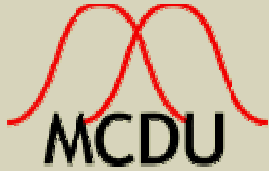
# A threshold of use

- Frequency counts show range of use:
  - 650 field occurs 11,778,732 time in 5,387,282 out of 7,595,887 records
  - 656 field occurs 1 time in 1 out of 7,595,887 records
- Determining a threshold of use
  - 80/20 rule (see previous slide)
  - Calculating a threshold that would be unambiguously determined, described, and universally applicable to our datasets
- Calculated threshold
  - Based on average occurrences (total occurrences/number of fields or field/subfields)
  - Has a meaning as a relative rather than an absolute value and can be used as a benchmark for different sets.
  - Fields/subfields within at/above the threshold have higher than average contribution toward total occurrences.



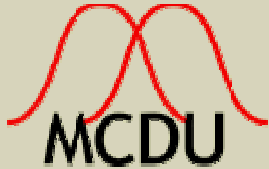
# Commonly occurring elements

- Second level analysis: Commonly occurring elements
  - Calculated threshold demarcates commonly occurring
- Commonly occurring across all Type of Records
  - Constitutes a base record
  - LC-created: Commonly occurring fields: **7**; Commonly occurring subfields: **10**
  - Non-LC-created: Commonly occurring fields: **6**; Commonly occurring subfields: **20**
  - See Handout Tables 11 & 12
- Commonly occurring elements in specific Type of Records
  - Additional elements to base record
  - Sample results: **Books, Pamphlets, and Printed Sheets Records**
  - LC-created: Commonly occurring fields:**16**; Commonly occurring subfields:**70**
  - Non-LC-created: Commonly occurring fields:**25**; Commonly occurring subfields: **107**
  - See Handout Tables 13 & 14
- More second level analysis and results... over to Dr. Miksa!



# Comparison with other standards

- Program for Cooperative's Cataloging (PCC) BIBCO Core Record Standards
- CONSER Record Requirements for Full, Minimal and Core Level Records for Serials
- National and Minimal Level Bibliographic Records Requirements
- Questions:
  - What are the sets of commonly used elements per format, and how do these compare with the elements prescribed in current national, core, and minimal level recommendations or guidelines for cataloging?
  - Conversely, are there elements which are frequently used by catalogers but are not prescribed in current national, core, and minimal level recommendations or guidelines for cataloging?

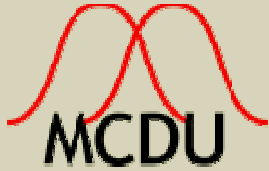


# Mandatory field requirements

- In a given PCC core level standard (CONSER or BIBCO)
  - Approximate count of mandatory elements (fields or subfields) is generally between **5-8**
  - Approximate count of mandatory if applicable elements (fields, subfields, or field blocks) is generally between **20-30**

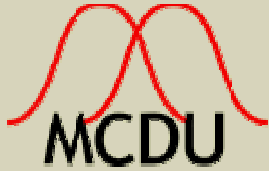
## Number and Percent of Records by 042 code

042 code	# OCLC	% OCLC	# LC	% LC
pcc	387,878	<1	538,644	6.2
lcd	265,503	<1	74,165	<1
msc	151,709	<1	44,769	<1



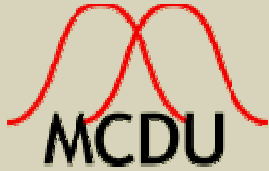
# PCC BIBCO & MCDU datasets

- Records in the datasets have disparate levels of fullness in terms of encoding level (full, minimal, etc.)
- BIBCO depends heavily upon the use of other standards (MARC, AACR2, etc.); closely tied to the cataloging rules of AACR2
- BIBCO standards provide bibliographic detail at approximately the second or third level of description, where all categories specified in the first and most or all categories in the second level are taken into account, while providing additional elements at the third level which are specific to the material being described
- MCDU format-specific datasets do not directly map to PCC BIBCO material types



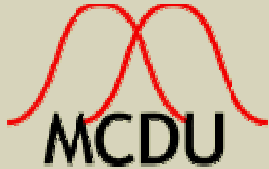
# Summary thoughts on analysis

- These guidelines are not grounded in empirical evidence, but rather in the interest of cost savings (in the case of PCC BIBCO Core Level Record Standards) and in universal usability and standardization (in the case of National and Minimal Level Record Requirements).
- MCDU presents empirical evidence that **catalogers' utilization of MARC fields/subfields is not directly aligned with the fields/subfields prescribed in these standards.**
- These results provide information for standards developers and the cataloging community to continue the dialog of standards development armed with empirical evidence to improve decision making, and also can aid in the improvement of the efficiency and effectiveness of cataloging practices.



## Example: nonLC Books

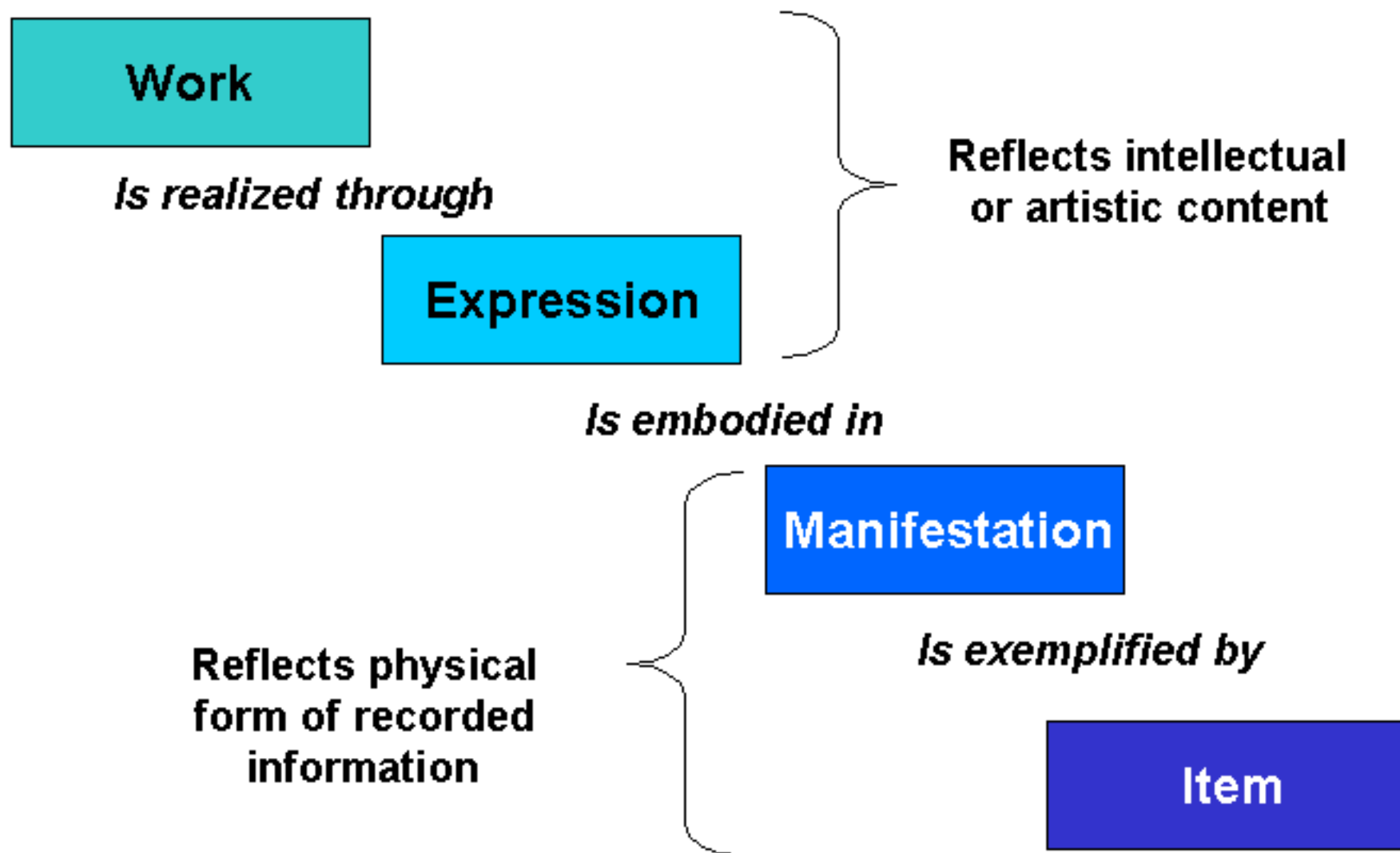
- 34,546,200 records (where Leader/06=a and Leader/07=m and where 008/23=s)
- 31 cataloger-supplied fields that account for 92.2% of all cataloger-supplied field utilization; 127 subfields account for 95.4% of all cataloger-supplied field utilization (see Tables 8 & 9)
- 25 fields and 107 subfields commonly occurring, not including 6 fields and 20 subfields commonly occurring in all nonLC records



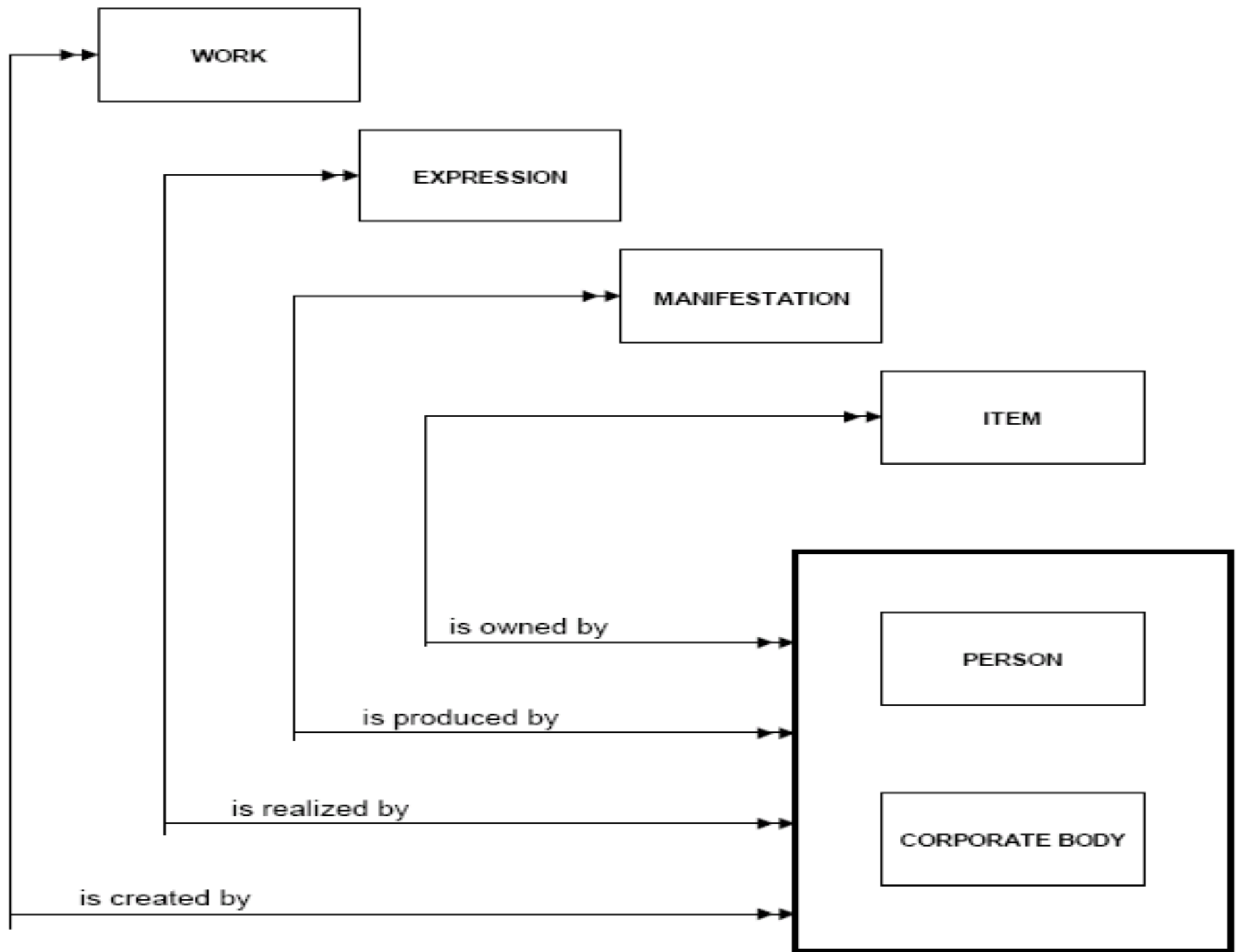
## Relationship between MCDU and FRBR

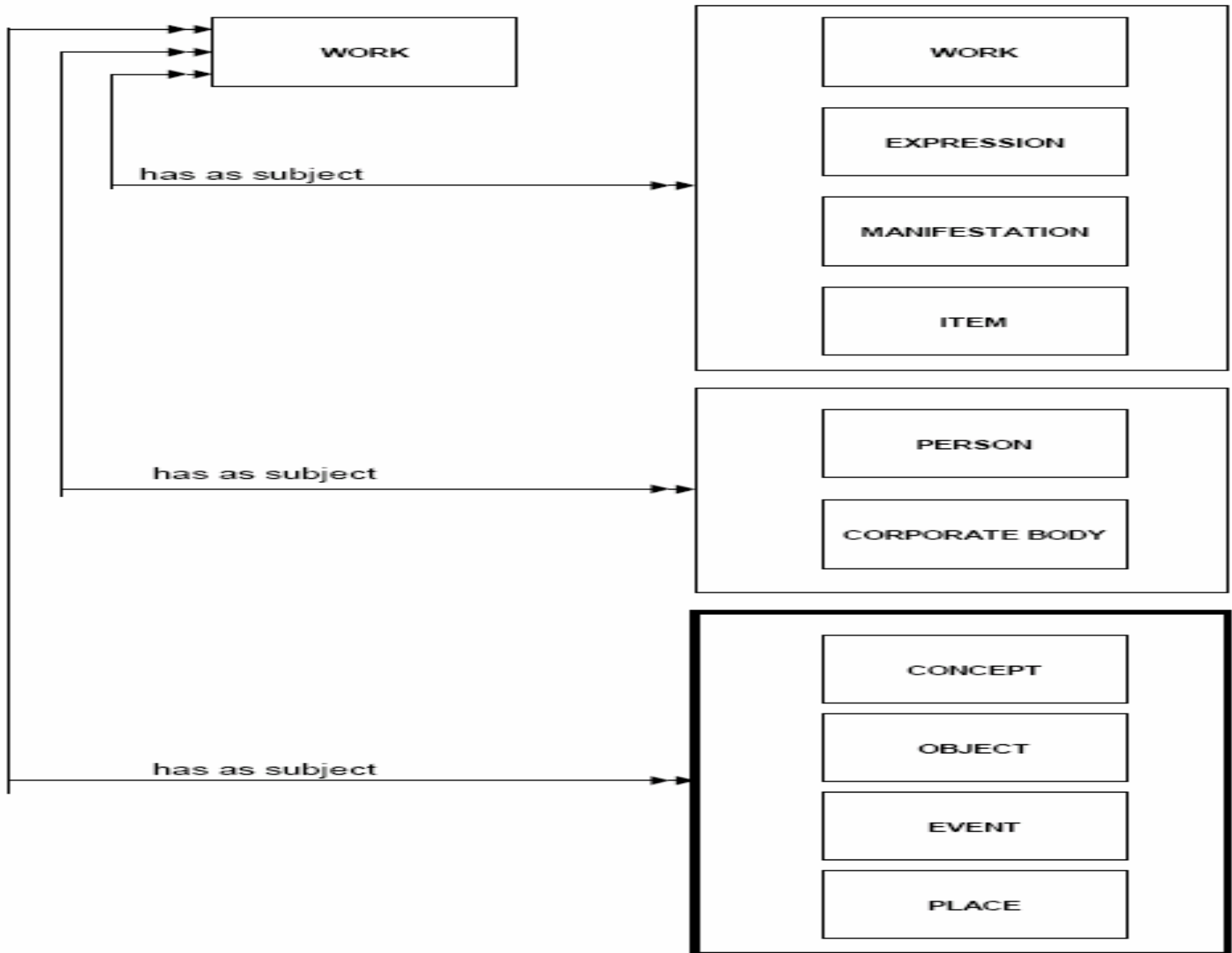
- Identifying commonly occurring elements in MARC 21 records can be used to inform the **practices of and research within** other metadata communities, as well as those involved in automatic metadata generation, metadata harvesting, or metadata transformation.
- FRBR makes no “*a priori*” assumptions about the bibliographic record itself, either in terms of content or structure”
- MCDU Project directly addresses the content designation structures (CDS) of MARC 21 with our data
- We can test FRBR’s entity-relationship model of basic functionality and make some conclusions as to whether or to what extent catalogers are utilizing MARC to support the model
- Identify commonly occurring elements that support the four FRBR user tasks—find, identify, select, and obtain

# Entities and relationships



Note: one-to-many relationships as you move down

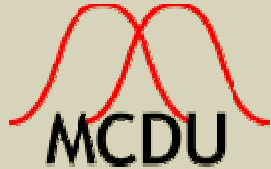






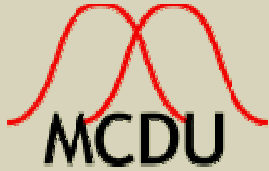
# FRBR model

- Based on Entity-Relationship modeling
  - Entity – something that can be described
  - Attributes –features of the entity that characterize it
  - Relationships between entities
- Three groups of entities in model
  - Group 1: Products of intellectual or artistic endeavor
  - Group 2: Entities responsible for the intellectual or artistic content, the physical production, etc.
  - Group 3: Entities that serve as the subjects of intellectual or artistic endeavor



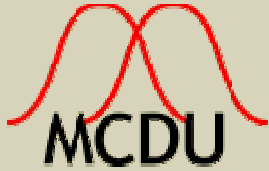
# FRBR's four user tasks

- **Find (Search):** Discovering if something exists by searching one or more attributes
- **Identify:** Examine retrieved records to determine the items that met user's search request
- **Select:** Examine retrieved records for those that meet other user needs/requirements
- **Obtain:** Using data in retrieved records to gain physical access to the described object
  
- Delsey mapped these tasks to MARC fields/subfield for FRBR entities
- Analysis of MCDU data for each of the four tasks adds another layer to Delsey's work by illustrating how bibliographic records support end users' activities when using library catalogs for resource discovery



# FRBR basic level of functionality

- **Find** all *manifestations* embodying:
  - the *works* for which a given *person* or *corporate body* is responsible
  - the various *expressions* of a given *work*
  - *works* on a given subject
  - *works* in a given series
- **Find** a particular *manifestation*:
  - when the name(s) of the *person(s)* and/or *corporate body(ies)* responsible for the *work(s)* embodied in the *manifestation* is (are) known
  - when the title of the *manifestation* is known
  - when the *manifestation* identifier is known
- **Identify** a *work*
- **Identify** an *expression* of a *work*
- **Identify** a *manifestation*
- **Select** a *work*
- **Select** an *expression*
- **Select** a *manifestation*
- **Obtain** a *manifestation*



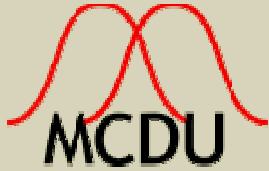
# Methodology for user task analysis

- Relied on Delsey's functional analysis of relation between MARC 21 and the four user tasks outlined in FRBR
- Obtained Delsey's mapping of tasks to MARC data elements (via Access Database, 2003 version)
- Paired with MCDU Frequency Counts (i.e., the actual utilization of the content designation ) which are at the level of variable data fields and related subfields
- Using data set of non-LC created Book 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"
- Non-LC Books records account for 34.5 million of the 56 million records in OCLC WorldCat database



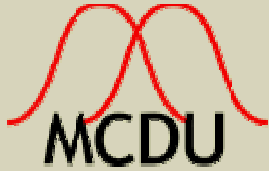
# Tasks and MARC elements

MARC 21 Bibliographic Format	FRBR Resource Discovery User Tasks			
	Find (Search)	Identify	Select	Obtain
Total number of elements (fields, subfields, fixed field positions, and indicator positions) that support a given task	454	972	375	468
Total number of indicator positions that support a given task	0	3	7	6



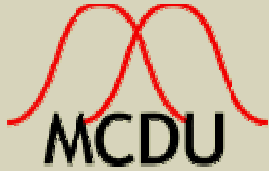
# Relationship between MCDU and FRBR

- Tables 15-18 show both the total number of variable fields/subfields occurring at least once in non-LC-created Book records (34,546,200 records) that support the user tasks as designated by Delsey's analysis.
  - **Find (Search):** 382 variable fields/subfields used, 61 are above the threshold; the **Find** task is supported by 16% of commonly occurring fields/subfields within this set.
  - **Identify:** 683 variable fields/subfields used, 82 are above the threshold; the **Identify** task is supported by 12% of commonly occurring fields/subfields within this set.
  - **Select:** 173 variable fields/subfields, 29 are above the threshold; the **Select** task is supported by 16.3% of commonly occurring fields/subfields within this set.
  - **Obtain:** 295 variable fields/subfields, 37 are above the threshold; the **Obtain** task is supported by 11.9% of commonly occurring fields/subfields within this set.



# Primary and secondary entities

- In the tables, entities 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, p.58). In the tables, 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).



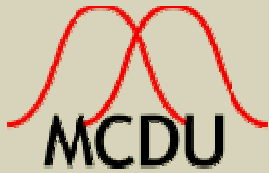
# Supporting user tasks

- MCDU has added another layer to Delsey's functional analysis of MARC 21—correspondence of actual utilization to the FRBR model.
- However this raises more questions of what these levels actually mean in the overall picture. For example:
  - *Do we know how many content designation structures are needed to support a user task?*
  - *Does a higher % of CDS used in a record necessarily mean stronger support for a user task?*



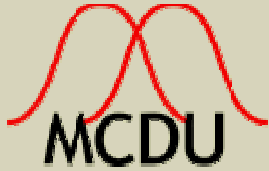
# Thoughts on more in-depth analysis of MCDU data

- Analysis of MCDU set is currently at the level of occurrence
- What could we discover if we look at levels of correlation of content designation structures between records; patterns of occurrence?
- Will this aid in identifying how often CDS is used and how they are utilized within record sets and what this says about the catalogers' utilization? For example:
  - What is the level of representation of CDS at or above threshold in the set of 56 million records?
  - What is the distinction of threshold CDS between record sets by format?
  - How does encoding level (Leader/17) and descriptive cataloging form (Leader/06) correspond to those threshold CDS?
  - What is the distribution of library types in OCLC participating libraries (040 \$a, \$c, \$d) and how does this help to characterize the records containing threshold CDS?
- These type of questions will require following the parsed fields/subfield CDS back to the non-parsed, original 56 million record set via OCLC record #, etc.



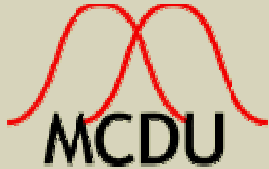
# Example: encoding levels and type of record

- Table 20 shows the correspondence of encoding levels with type of record which provides some interesting correlations that can lend itself for informing future discussions on encoding level
- All materials:
  - 1.2% of entire 56 million encoded at Core level (4)
- Language materials (a):
  - 16% encoded *Full* (#)
  - 16% encoded *Less-than full input by OCLC participants* (K)
  - 16% encoded *Less-than-full added from batch process* (M);  
2% encoded *Full level input added from batch process* (L)
  - 41% encoded at *Full level input by OCLC participants* (I)
- Encoding level 2, E, consistently low across type of material; does low occurrence mean these levels not useful?



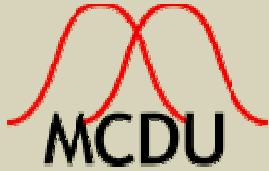
## Community's response

- What does low occurrence of fields/subfields suggest to you, the cataloging community?
- What would be the anticipated results of a more in-depth survey of CDS, beyond occurrence?
- What might we find? How would it make a difference?



## 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?



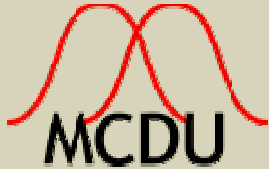
## Questions for consideration?

- 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?



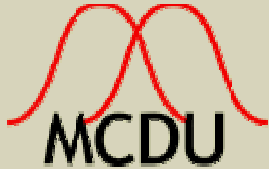
# References

- **MARC Content Designation Utilization Project**
  - <http://www.mcd�.unt.edu/>
- Goldsmith and Knudson. 2006. **Looking back, Looking forward: A metadata standard for LANL's aDORe repository**
  - <http://doi.acm.org/10.1145/1141753.1141814>
- Moen and Benardino. 2003. **Assessing Metadata Utilization: An Analysis of MARC Content Designation Use**
  - [http://www.unt.edu/wmoen/publications/MARCPaper\\_Final2003.pdf.pdf](http://www.unt.edu/wmoen/publications/MARCPaper_Final2003.pdf.pdf)
- **Designing the Future -- Library Systems and Data Formats**
  - <http://futurelib.pbwiki.com/>



# References

- **Next Generation Catalogs for Libraries**
  - [NGC4LIB@listserv.nd.edu](mailto:NGC4LIB@listserv.nd.edu)
- Library of Congress, NDMSO. 2001. **MARC 21 Format for Bibliographic Data, Update no. 2.**
- OCLC. 2003. **OCLC Bibliographic Formats and Standards (3<sup>rd</sup> edition).**
  - <http://www.oclc.org/bibformats/>
- Program for Coop Cataloging (PCC). 2005. **Introduction to the Program for Cooperative Cataloging BIBCO Core Record Standards**
  - <http://www.loc.gov/catdir/pcc/bibco/coreintro.htm>
- CONSER. 2005. **Record Requirements for Full, Core, and Minimal Level Records.**
  - <http://www.loc.gov/acq/conser/recordreq.html>



# References

- International Federation of Library Associations. 1998. **Functional requirements for bibliographic records: final report.**
  - <http://www.ifla.org/VII/s13/frbr/frbr.pdf>
- Delsey, Tom. 2003. ***Functional analysis of the MARC 21 bibliographic and holdings formats, Second revision.***
  - <http://www.loc.gov/marc/marc-functional-analysis/functional-analysis.html>
- Library of Congress, Network Development and MARC Standards Office. (2006). **Access 2000 database filename: FRBR\_Web\_Copy.mdb**, [Data file].
  - <http://www.loc.gov/marc/marc-functional-analysis/functional-analysis.html>



## References

- Miksa, S., Moen, W., Snyder, G., Eklund, A., & Polyakov, S. (2006). **Metadata Assistance of the Functional Requirements for Bibliographic Records' Four User Tasks: a report on the MARC Content Designation Utilization (MCDU) Project.** In Knowledge Organization for a Global Learning Society: Proceedings of the 9th International Conference for Knowledge Organization. International Society for Knowledge Organization 9th International Conference, Vienna, Austria. July 5-7, 2006. G. Budin, C. Swertz, & K. Mitgutsch (Eds.), *Advances in Knowledge Organization*, vol 10, pp. 41-49. Würzburg: Ergon.