Assessing Metadata Utilization: An Analysis of MARC Content Designation Use

William E. Moen, Penelope Benardino
School of Library and Information Sciences, Texas Center for Digital Knowledge
University of North Texas, USA
wemoen@unt.edu, pen.ben@attbi.com

Abstract

Metadata schemes emerge to meet community and user requirements, and they evolve over time to meet changing requirements. This paper reports results of an analysis of a large sample of MARC 21 bibliographic records. MARC 21 is an encoding scheme related closely to metadata elements occurring in library bibliographic records. The records were analyzed for the utilization of content designation available in MARC 21. Results indicate that less that 5% of available content designation accounts for over 80% of occurrences .The implications of these findings affect indexing policies, system design, and can inform setting requirements for extending a metadata scheme based on a threshold of community requirements.

Keywords: *Metadata Utilization, MARC 21, Cataloging Practices, Indexing Policies, Interoperability*

1. Introduction

Communities develop and evolve metadata schemes to serve their current and emerging needs. In its first incarnation, the Dublin Core Metadata Element Set comprised thirteen elements to assist in resource discovery. Subsequently two additional elements were added. Over the past six years, the metadata scheme has evolved to provide more specific encoding through the use of qualifiers, and the extensibility of Dublin Core has been exercised by a number of communities (as reflected in the application profiles created by several communities) [1]. Two significant questions emerge: When is a need significant enough to warrant additional capability in the metadata scheme? To what extent will the additional refinements and enrichment of the metadata scheme be utilized?

The Machine Readable Catalog record (MARC) provides a structure for content designation used in resource description, typically in the context of library materials [2]. Its development since the late 1960s reflects capability for content designation. The availability for rich encoding and content designation does not necessarily imply utilization of that richness. This paper reports preliminary findings from an analysis of approximately 400,000 MARC 21 records from OCLC's WorldCat database. This analysis was carried out for a specific purpose as part of the Z39.50 Interoperability Testbed Project. The examination of the dataset revealed the extent to which various fields and subfields are actually used in practice.

2. Background for the Analysis

The Z39.50 Interoperability Testbed (Z-Interop) Project is an applied research and demonstration project funded by the U.S. federal Institute of Museum and Library Services through a National Leadership Grant awarded the School of Library and Information Sciences and the Texas Center for Digital Knowledge at University of North Texas [3]. The goal of Z–Interop is to improve Z39.50 semantic interoperability among libraries for information access and resource sharing. The mission of Z–Interop is to:

- Provide a trusted testing environment for vendors and consumers of Z39.50 products to demonstrate and evaluate those products
- Develop rigorous methodologies, test scenarios, and procedures to measure and assess interoperability
- Demonstrate and operate a Z39.50 interoperability testbed.

A critical component of the Z-Interop Project is a test dataset of 419,657 MARC 21 bibliographic records (hereafter referred to as the Z-Interop dataset). OCLC, a Z-Interop Project collaborator, provided these records from its WorldCat bibliographic database. At the time of extraction from the WorldCat database, the Z-Interop dataset comprised approximately a one percent sample of WorldCat records. The extraction algorithm used to select the sample was based on the number of holdings indicated for a single bibliographic item. Although the resulting sample was neither a random nor stratified sample, it comprised a relatively representative sample of bibliographic records based on frequency of holdings of OCLC member libraries.

A key area of consideration when addressing Z39.50 interoperability is the indexing policies in effect in different online catalog systems. These indexing policies prescribe which fields/subfields in a MARC 21 record are included to populate an individual index. The Z-Interop Project developed indexing guidelines to use in the reference implementation of an online catalog system and Z39.50 server [4]. Sirsi, another collaborator on the Z-Interop Project, contributed its Unicorn system to serve as an online catalog and Z39.50 server reference implementation. Z-

Interop Project staff had complete control over indexing decisions for the Unicorn system.

To develop the indexing guidelines for selected keyword indexes, the MARC 21 bibliographic format was examined and all fields/subfields that hold author, title, or subject data were identified as candidates for indexing. The number of fields/subfields identified in the indexing guidelines for several keyword indexes are:

- Author-related data: 119 fields/subfields
- Author- and title-related data: 21 fields/subfields
- Title-related data: 253 fields/subfields
- Subject-related data: 144 fields/subfields

Table 1 summarizes these fields/subfields in the various MARC 21 tag groups. MARC is a very rich format for content designation, and local system implementations choose which fields/subfields will be used for the various indexes established. One approach is to simply index each field/subfield that contains author-, title-, or subject-related data. Establishing and setting up indexing policies, however, can be a time consuming task; for the Z-Interop Project's online catalog reference implementation, setting up the indexing policies for author, title, and subject keyword indexes took approximately forty person-hours. More importantly from the user's perspective is whether such extensive indexing has meaningful consequences for search and retrieval. These questions motivated the analysis of the actual occurrence of the MARC 21 fields/subfields in the Z-Interop dataset.

Table 1. Fields/Subfields Identified for Indexing in Z-Interop Indexing Guidelines

MARC 21 Field Groups	Currently Defined	Fields/Subfields Unlikely To Be Used	Total
00x	0	0	0
0xx	0	0	0
1xx	54	2	55
2xx	65	1	66
3xx	0	0	0
4xx	5	39	44
5xx	8	0	8
6xx	136	4	140
7xx	145	4	149
8xx	73	2	75
Total	**486**	**52**	**537**

2.1. Brief Discussion of MARC

The Machine-Readable Catalog Record (MARC) was developed at the Library of Congress in the 1960s. A major requirement for the MARC structure was to accommodate bibliographic information contained on library catalog entries while making the information available for computer processing. Originally referred to as the MARC Communication Format, it was intended to provide a standard structure for exchanging bibliographic records among library automation systems. MARC originated as a means to communicate bibliographic data about printed texts, but has evolved to communicate data about books, computer files, maps, serials, music, visual materials and archival materials.

The structure of the record is specified by national and international standards, ANSI/NISO Z39.2 and ISO 2709 respectively [5,6]. The specifications for the record structure do not provide semantics for the content designation (i.e., the semantics of the field tags, subfield codes, etc.) and additional technical specifications have been developed to provide semantics and procedures for encoding bibliographic data into the record structure. The MARC 21 format is the latest iteration of MARC content designation. The content of the bibliographic records is governed by other rules and sources, typically cataloging rules in the form of the Anglo-American Cataloguing Rules [7], authority lists, and controlled vocabularies.

The MARC 21 Format for Bibliographic Data is a very rich encoding and content designation scheme with 1908 fields/subfields available [8,9]. Table 2 shows a breakout by MARC 21 tag groups for the fields/subfields included in the MARC 21 Format for Bibliographic Data. The extent to which this metadata structural richness is utilized and how to assess utilization of a metadata scheme and its encoding are the focus of this paper.

Table 2. Fields/Subfields in MARC 21 Bibliographic Format

MARC 21 Field Groups	Currently Defined	Obsolete *	Total
00x	6	1	7
0xx	238	7	245
1xx	66	1	67
2xx	137	32	169
3xx	109	32	141
4xx	69	0	69
5xx	323	38	361
6xx	184	5	189
7xx	452	47	499
8xx	141	20	161
Total	**1725**	**183**	**1908**

*Obsolete content designators are not to be used in new records but they may appear in records created prior to the time a content designator was defined as obsolete.

2.2. Methodology

As part of the Z-Interop Project, the original MARC 21 records were decomposed into multiple subrecords based on individual words in each field/subfield. For information describing the decomposition, see [10]. Each MARC 21 record was decomposed into separate subrecords that included: OCLC Number, Field Tag, First Indicator Value, Second Indicator Value, Subfield Value, Field Position in Record, Subfield Position in Record, Word Position in

Field/Subfield, and Specific Character String (i.e. the "word"). Table 3 provides a sample of the

Table 3. Components of a Z-Interop Dataset Subrecord

OCLC#	Tag	1st Indicator	2nd Indicator	Subfield	Field Position	Subfield Position	Word Position	Word
3	110	2		a	11	1	1	national
3	110	2		a	11	1	2	study
3	110	2		a	11	1	3	service
3	245	1	0	a	12	1	1	illegitimacy
3	245	1	0	a	12	1	2	and
3	245	1	0	a	12	1	3	adoption
3	245	1	0	b	12	2	1	report

decomposed records. Each row in the table represents a "subrecord" for the parent MARC 21 record. The data comprising the subrecords were loaded into a MySQL database for processing. The decomposed records were analyzed to produce a frequency count of occurrences of fields/subfields contained in the 419,657 MARC 21 records. The output was a sorted list of occurrences of individual fields/subfields. Table 4 contains a sample of the resulting frequency count data. Included in the sample list is an instance of the MARC 21 field 650 $a to demonstrate a repeatable field/subfield provided in MARC 21. A number of fields/subfields can occur multiple times in a single record, and therefore the occurrence of a field/subfield can be greater than the total number of records (e.g., 602,362 occurrences is greater than the 419,657 number of records). The focus of the analysis was on number of total occurrences in the dataset rather than number of records in which the field/subfield occurred. Certain fields are required to be in every record (e.g., the 001), and there is a one-to-one match between occurrences of these fields/subfields in the dataset and the total number of records.

Table 4. Sample Frequency Count Data

MARC 21 Field	MARC Subfield	Occurrence
001		419,657
003		419,657
005		419,657
006		652
007		30,556
008		419,657
010	a	305,407
010	b	2
010	z	6,627
650	2	15,361
650	6	9
650	a	602,362
650	b	28
650	c	4
650	d	16
650	f	1
650	k	2
650	v	83,607
650	x	326,867
650	y	32,728
650	z	231,459

The frequency count data were imported into a spreadsheet for subsequent analysis. Using the MARC 21 Concise Bibliographic Format, field/subfield names and semantics were added [11]. OCLC's Bibliographic Formats was consulted to account for MARC fields/subfields that could not be found in the MARC 21 documentation [12]. Linking fields were noted according to whether the field had a $6 (Linkage) field/subfield. For the fields/subfields whose definitions were taken from OCLC, linking information was not available. Repeatability of fields/subfields was noted. The repeatability indication was based on the repeatability of the subfield within the field rather than the repeatability of the field within the record. For example, field $650 (Subject Added Entry-Topical Term) is repeatable within a record, however within field $650, subfield $a is not repeatable; subfield $650a will show to be a non-repeating subfield in the analysis, even though it can occur as many times in a record as the cataloger deems necessary to adequately describe the entity. Because the occurrences in the frequency count list are broken down to the subfield level, the repeatability indication was based on the subfield's repeatability within the field. In addition, the review of MARC documentation showed 102 fields/subfields occurring in the Z-Interop dataset as "Obsolete", "LC use only", "OCLC use only", "Do not use", or "Unlikely to be used". Also, one field is used in specific cataloging software, and sixteen fields/subfields were assumed to be cataloging mistakes since there was no description for them in MARC 21 or in OCLC's MARC documentation (these fields/subfields occurred at the most 3 times).

Three sets of fields/subfields from the Z-Iinterop Indexing Guidelines (those candidate fields/subfields for author-, title-, and subject-keyword indexes) were also imported into a spreadsheet [4]. This spreadsheet was cross-referenced with the frequency count spreadsheet. The fields/subfields in the frequency count spreadsheet identified as candidates for author-, title-, and subject-indexing were indicated. The result was a set of spreadsheets with each field/subfield identified by name, semantics, source of information, as a linked field, its repeatability, miscellaneous information (e.g., not likely to

be used for various reasons), and whether it was a candidate for keyword indexing for author, title and/or subject.

Two other spreadsheets clustered the frequency count fields/subfields into MARC tag groups (i.e., 0xx, 1xx, 2xx, etc.) and clustered candidate fields/subfields for indexing into MARC tag groups.

3. General Characteristics of the Bibliographic Records

The Z-Interop dataset contains 419,657 MARC 21 bibliographic records. These records describe bibliographic items in various formats. Table 5 lists approximate percentages of each format represented in the dataset.

Table 5. MARC 21 Formats Represented in Dataset

MARC 21 Format	Approximate Percent of Total
Book	91%
Cartographic Material	less than 1%
Visual Materials	1%
Sound Recording or Printed or Manuscript Music	4%
Electronic Resources	less than 1%
Archival/Mixed Materials	less than 1%
Serial	3%

In the 419,657 records, 926 fields/subfields occur at least once. As noted above, a total of 1908 fields/subfields are defined in the MARC 21 Format for Bibliographic Data (including those currently available for use and obsolete) [9]. A first indication of utilization shows that less than 50% of the content designation available in MARC 21 appears to be used in this set of records from OCLC WorldCat. Table 6 summarizes the number of fields/subfields in the Z-Interop dataset for each MARC 21 tag group (compare with Table 2 above that shows all available content designation).

Table 6. Fields/Subfields Occurring in Z-Interop Dataset

MARC 21 Field Groups	Currently Defined	Obsolete	Fields /Subfields Unlikely To Be Used	Total
00x	6	0	0	6
0xx	96	1	33	130
1xx	49	0	2	51
2xx	81	0	19	100
3xx	23	6	0	29
4xx	10	0	30	40
5xx	128	1	3	132
6xx	104	1	7	112
7xx	205	0	5	210
8xx	105	3	8	116
Total	**807**	**12**	**107**	**926**

Each of the 926 fields/subfields was examined using MARC 21 and OCLC MARC documentation as references. This review revealed that 119 (13%) of the fields/subfields were labeled as "obsolete" or unlikely to be used. Since these records were taken from the OCLC WorldCat database, it is important to note that a number of fields/subfields are specific to the OCLC MARC records. Thirty-three of the fields/subfields are specific to OCLC MARC records and 1 subfield (69$a) is described as a Local Call Number field in Dynix catalogs. Frequency of occurrence of these 119 fields/subfields range from 1 to 419,657 times.

Certain fields/subfields are applicable only to certain formats (see Table 5). For example, the MARC 21 field 255, Cartographic Mathematical Data, has seven subfields defined. These subfields address data specific only to the Cartographic Materials format. When considering the analysis of occurrences of fields/subfields, the raw count has to be seen in the context of format-specific content designation options. If a specific format of material occurs in a small percentage of the 419,657 records, the count of format-specific fields/subfields may be relatively small overall, but can actually be highly used within those format-specific records. For example, the 255$a occurs 1,289 times in the dataset (less than 1% of total occurrences). However, there are less than 1% of all records in the dataset that are designated as Cartographic Materials (1,677). We may assume that the approximately 77% of the records describing Cartographic Materials contain a 255$a. Future analysis will examine more specifically the occurrences of format-specific content designation.

3.1. Analysis of Content Designation Use

In the dataset, 926 fields/subfields are present and the frequency of occurrence ranges from 1 to 602,362 times. Table 7 summarizes the occurrences in the MARC tag groups.

Table 7. Occurrences per MARC Tag Group in Dataset

MARC 21 Field Groups	Number of Fields/Subfield Used	Total Occurrences of Fields/Subfields
00x	6	1,709,836
0xx	130	4,393,134
1xx	51	577,856
2xx	100	2,438,275
3xx	29	1,086,239
4xx	40	200,424
5xx	132	707,316,
6xx	112	1,919,409
7xx	210	560,769
8xx	116	259,273
Total	**926**	**13,145,215**

One approach to assessing utilization of the content designation available in the MARC 21 format is to analyze

the number of occurrences of individual fields/subfields in the Z-Interop dataset. This analysis revealed that a very small number of fields/subfields account for the highest occurrences within the dataset. Table 8 summarizes the number of fields/subfields occurring in groups of approximately 100,000 occurrences. Total number of all content designation occurrences in the Z-Interop dataset is 13,840,499, and 36 of the most frequently occurring fields/subfields account for approximately 80% of occurrences of all fields/subfields. This means that only 4% of all fields/subfields present in the Z-Interop dataset account for 80% of the occurrences, or to state it another way, 96% of all fields/subfields account for less than 20% of occurrences.

Table 8. Number of Fields/Subfields by Range Frequency

Frequency	Number of MARC 21 Field/Subfields	Percent of All Occurrences
> 600,000	1	4.4%
500,000 – 599,999	0	0%
400,000 – 499,999	13	39.9%
300,000 – 399,999	6	14.3%
200,000 – 299,999	6	10.6%
100,000 – 199,999	10	10.3%
Total	**36**	**79.5%**

Table 9 provides a list of the 36 most frequently occurring fields/subfields in the dataset. Certain fields (e.g., 650) and certain subfields (e.g. 40 $d), can occur more than once in a single record. Four of these content designation structures are shown in Table 9; they are the only ones that occur at a frequency greater than the total number of records in the dataset. Several fields are mandatory and non-repeating in all MARC 21 records. These are listed in Table 9 with a frequency of 419,657. The frequency count for the occurrence of each of these fields is exactly the same as the total number of records in the test dataset.

The 36 most frequently occurring fields/subfields can be combined into their respective MARC 21 tag groups to represent the relative use of these field groups in the dataset (see Table 10). The table also provides the percent of all occurrences the 36 most frequently occurring fields/subfields account for.

Table 9. Top 36 Occurring Fields/Subfields in Z-Interop Dataset

Frequency	MARC 21 Field	Subfield	Field & Subfield Name
602,362	650	a	Subject Added Entry Topical Term Subfield a: Topical term or geographic name as entry element
454,451	40	d	Cataloging Source Subfield d: Modifying agency
451,808	260	a	Publication, Distribution, etc. (Imprint) Subfield a: Place of publication, distribution, etc
435,783	260	b	Publication, Distribution, etc. (Imprint) Subfield b: Name of publisher, distributor, etc
419,657	001		Control Number
419,657	003		Control Number Identifier
419,657	005		Date and Time of Latest Transaction.
419,657	008		Fixed-Length Data Elements
419,657	040	c	Cataloging Source Subfield c: Transcribing agency
419,657	049	a	Local Holdings Subfield a: Holding Library
419,641	245	a	Title Statement Subfield a: Title
416,908	300	a	Physical Description Subfield a: Extent
415,423	040	a	Cataloging Source Subfield a: Original cataloging agency
410,790	260	c	Publication, Distribution, etc. (Imprint) Subfield c: Date of publication, distribution, etc.
391,899	300	c	Physical Description Subfield c: Dimensions
329,796	245	c	Title Statement Subfield c: Statement of responsibility, etc.
326,867	650	x	Subject Added Entry Topical Term Subfield x: General subdivision
318,692	100	a	Main Entry Personal Name Subfield a: Personal name
305,407	010	a	Library of Congress Control Number. Subfield a: LC control number
300,385	050	a	Library of Congress Call Number

Frequency	MARC 21 Field	Subfield	Field & Subfield Name
			Subfield a: Classification number
285,578	050	b	Library of Congress Call Number Subfield b: Item number
274,313	082	a	Dewey Decimal Call Number Subfield a: Classification number
235,864	300	b	Physical Description Subfield b: Other physical details
231,459	650	z	Subject Added Entry Topical Term Subfield z: Geographic subdivision
228,173	020	a	International Standard Book Number Subfield a: International Standard Book Number
210,250	500	a	General Note Subfield a: General note
186,997	504	a	Bibliography, Etc. Note Subfield a: Bibliography, etc. note
176,916	700	a	Added Entry Personal Name Subfield a: Personal name
169,178	245	b	Title Statement Subfield b: Remainder of title
149,540	100	d	Main Entry Personal Name Subfield d: Dates associated with a name
144,261	082	2	Dewey Decimal Call Number Subfield 2: Edition number
141,409	043	a	Geographic Area Code Subfield a: Geographic area code
118,647	651	x	Subject Added Entry Geographic Name Subfield x: General subdivision
113,050	651	a	Subject Added Entry Geographic Name Subfield a: Geographic name
112,156	019	a	OCLC Control Number OCLC use only
110,257	850	a	Holding Institution Subfield a: Holding institution

Table 10. 36 Fields/Subfields Compared to All Occurrences in MARC 21 Field Groups

MARC 21 Field Group	Occurrences in Top 36 Fields/ Subfields	Occurrences of All Fields/ Subfields	Percent of All Accounted for by Top 36 Fields/ Subfields
00X	1,678,628	1,709,836	98%
0XX	3,500,870	4,393,134	80%
1XX	468,232	577,856	81%
2XX	2,216,996	2,438,275	91%
3XX	1,044,671	1,086,239	96%
4XX	0	200,424	0%
5XX	397,247	707,316	56%
6XX	1,392,385	1,919,049	73%
7XX	176,916	549,097	32%
8XX	110,257	259,273	43%
Total	**10,986,202**	**13,840,499**	**79%**

3.2. MARC 21 Content Designation and Indexing Analysis

The initial motivation for this examination of field/subfield occurrence was to assess indexing policies for the Z-Interop Testbed. The Z-Interop indexing guidelines identified a total of 537 author-, title-, or subject-related fields/subfields that could be candidates for indexing [4]. Only 381 of those fields/subfields actually occurred in the Z-Interop dataset (see Table 11).

The analysis also looked at the frequency of occurrences of the 381 fields/subfields in the dataset. Total occurrences of the 381 fields was 4,397,712. Nineteen of the most frequently occurring fields/subfields account for approximately 80% of this total. These nineteen fields/subfields occur a total of 3,489,198 times in the dataset. This means that approximately 5% of the fields/subfields identified as candidates for indexing account for 80% of all occurrences, or stated another way, 95% of the candidate fields/subfields account for only 20% of all occurrences. Table 12 lists the nineteen fields/subfields.

Table 11. Summary of Fields/Subfields in Indexing Guidelines

Keyword Index Guidelines	Fields/ Subfields in Indexing Guidelines	Indexing Guidelines Fields/Subfields Occurring in Dataset	Percent Occurring
Author Only	119	86	72%
Author and Title	21	16	76%
Subject Only	144	101	70%
Title Only	253	178	70%

Total	**537**	**381**	71%

Table 12. Summary of Fields/Subfields in Indexing Guidelines

# Occurrences	Marc 21 Field	Subfield	Description	Index
602,362	650	a	Subject added entry Topical Term Subfield a = Topical term or geographic name as entry element	Subject
419,641	245	a	Title Statement Subfield a = Title	Title
329,796	245	c	Title Statement Subfield c = statement of responsibility	Author
326,867	650	x	Subject added entry Topical Term Subfield x = General subdivision	Subject
318,692	100	a	Main entry Personal Name Subfield a = personal name	Author
231,459	650	z	Subject added entry Topical Term Subfield z = Geographic subdivision	Subject
176,916	700	a	Added entry Personal Name Subfield a = personal name	Author
169,178	245	b	Title Statement Subfield b = Remainder of title	Title
149,540	100	d	Main entry Personal Name Subfield d = dates associated with a name	Author
118,647	651	x	Subject added entry Geographic Name Subfield x = General subdivision	Subject
113,050	651	a	Subject added entry Geographic Name Subfield a = Geographic name	Subject
83,607	650	v	Subject added entry Topical Term Subfield v = Form subdivision	Subject
74,606	700	d	Added entry Personal Name Subfield d = dates associated with a name	Author
69,636	600	a	Subject added entry Personal Name Subfield a = personal name	Subject
66,375	710	a	Added entry Corporate Name Subfield a = corporate name or jurisdiction name	Author
64,433	440	a	Series Statement Added Entry Title Subfield a = title	Title
62,853	490	a	Series Statement Subfield a = Series statement	Title
56,229	600	d	Subject added entry Personal Name Subfield d = dates associated with a name	Subject
55,311	653	a	Index Term Uncontrolled Subfield a = the term	Subject

4. Discussion

This analysis has provided a description of the use of a metadata and content designation scheme. MARC 21 is a rich encoding scheme with nearly 2,000 discrete structures for content designation. Less than 50% of these structures actually occurred in a large dataset of these records, but more interesting is that only 4% of the occurring fields/subfields account for nearly 80% of all occurrences. Should this be of concern?

One might suggest that the rich encoding structure provides a capability in case we need it. In case there is a specific datum that needs to be recorded with a discrete MARC 21 content designation, the format has it available. From the vantage point of a system designer, whether or not these content designations are ever used, the system must be programmed to be ready in case one of the structures occurs in a record. There is a potential resource impact at the level of system design and implementation, with associated costs in the final product.

From the perspective of the Z-Interop Testbed Project, where semantic interoperability depends in part on common indexing practices, accounting for over 500 fields/subfields in the indexing policies has a resource impact on setting up the indexing policies.

Furthermore, as metadata schemes such as Dublin Core or Metadata Object Description Schema (MODS) [13] are developed and evolve, there will always be requirements to extend the capability of the metadata scheme to accommodate new requirements of communities and users. MARC has developed over thirty years, and the approximately 2,000 structures for content designation reflect a response to those community and user requirements. While it may only be possible after a scheme has been implemented for some time to analyze the extent

to which the content designation is actually utilized, there may be lessons from the evolution of MARC that point to the need for policies that identify "thresholds of needs" before additional content designation capability is introduced. A balanced approach that allows a metadata scheme to be responsive to evolving needs while minimizing increasing capability that ends up being under-utilized would be most desirable.

5. Additional Analyses and Future Research

The analysis of one sample of MARC 21 records illustrates an approach to assessing and preliminary results of utilization of available content designation. Further analysis will be carried out to refine the results including:

- Investigating the encoding levels of the records since all records may not be full-level cataloging and this may affect use of content designation
- Identifying utilization of format specific content designation
- Examining the occurrence of the content designation at a record level rather than frequency counts of total occurrences in the dataset.

It will also be important to carry out this analysis on other collections of MARC bibliographic records. Using collections of bibliographic records from library catalogs of a university library and a large public library would allow a comparison of findings from the current analysis.

In addition to refining the analysis and conducting similar analyses on other collections of records, utilization analysis results can be linked to other investigations. The following are some planned next steps and questions in this stream of research.

5.1. Use of Content Designation Related to Cataloging Rules

The MARC record's content is created using a variety of rules and guidelines, particularly the Anglo-American Cataloguing Rules, Library of Congress Subject Cataloging Manual, and other associated tools. An analysis needs to be carried out that looks at infrequently occurring fields and subfields, and the cataloging rules and MARC input rules associated with these fields. Are there particular issues about these rules (very specialized, too obscure, etc.) that result in the minimal utilization of the associated MARC content designation?

5.2. National and Minimal Level Cataloging Guidelines

The Network Development and MARC Standards Office at the Library of Congress publishes MARC 21 Format for Bibliographic Data: National Level Record---Bibliographic Full Level & Minimal Level [14]. This document identifies specific fields/subfields that must occur (M), must occur if applicable (A), or are optional (O) in catalog records. Using the 36 most frequently occurring fields, Table 13 indicates how these are designated in the National Level Record document. A similar analysis could be carried out on additional fields/subfields to see the relationship of their occurrence and the national level record guidance published by the Library of Congress. Do the guidelines include requirements for fields/subfields that in practice are seldom used?

5.3. Analysis of MODS

The Metadata Object Description Schema (MODS) that is being developed is a subset of MARC 21 content designation [13]. It would be appropriate to examine the MODS structure from the perspective of the analysis done on the Z-Interop dataset. Data structures included in MODS that relate to seldom used MARC 21 content designation could be examined and reconsidered in light of actual use of these MARC 21 fields/subfields.

5.4. Functional Analysis of the MARC 21 Bibliographic and Holdings Formats

The Network Development and MARC Standards Office commissioned a study to analyze the MARC 21 format from the following perspectives:

- The Functional Requirements for Bibliographic Records (FRBR) model
- The Anglo-American Cataloging Rules model
- A set of user tasks that the format might logically support

The findings from the report provide the basis for another comparison between what exists in actual records and recommendations for bibliographic data to support user tasks and other activities [15]. The study mapped the attributes in the FRBR model to the MARC data elements, identified MARC data elements that fall outside the FRBR model, and analyzed the data content of the MARC format as it corresponds to the user tasks outlined in the FRBR model. It is interesting to note that the study found that approximately 50% of the MARC data elements corresponds to the FRBR and AACR models. This finding is similar to the results found in our comparison of actually occurring content designation in the Z-Interop dataset with all available MARC 21 content designation. Is this just a coincidence?

5.5. Impacts on Information Retrieval

The Z-Interop Testbed Project, for which the analysis reported here was initially carried out, will experiment with indexing policies based on the findings from this analysis. Currently, indexing policies for author-, title-, and subject-keyword searching address all 537 fields identified in the indexing guidelines. The testbed has defined test searches

with known results to be returned based on the current indexing policies. The testbed will implement indexing policies only using the 19 fields/subfields that are most frequently occurring. Test searches can be issued and comparison in search results can be used to determine if information retrieval has suffered because of using a very small number of fields/subfields in the indexing policies. With this information, local library implementations of online catalogs can be in a better position to determine the extent of fields/subfields that must be included in their indexing policies for appropriate levels of retrieval.

6. Conclusion

This study presents a preliminary approach for assessing utilization of metadata schemes by examining actual records that implement the scheme. In the Z-Interop dataset, less the 4% of available MARC 21 content designation accounts for 80% of all occurrences of the content designation. MARC has evolved over thirty years, an evolution that responded to community and user needs. New content designation was added to the MARC format in response to those needs. The results of this analysis of actual use of the content designation provides a point of departure for discussions about when and to what extent should a metadata scheme's content designation capability be extended. As Dublin Core and schemes such as MODS evolve, the question of extensions and expansion needs to be addressed. Policies that address increasing content designation capability should be considered as well as mechanisms to review actual utilization of the content designation. The methodology of metadata utilization assessment presented in this paper provides a first step in developing robust and rigorous utilization assessments for a variety of metadata schemes.

Table 13. Top 36 Fields/Subfields and National and Minimal Level Cataloging Requirements

Frequency	MARC 21 Field	Subfield	National Level Cataloging	Minimal Level Cataloging		Frequency	MARC 21 Field	Subfield	National Level Cataloging	Minimal Level Cataloging
419,657	001		M	M		169,178	245	b	A	O
419,657	003		M	M		329,796	245	c	A	O
419,657	005		M	M			260		A	A
419,657	008		M	M		451,808	260	a	A	O
	010		A	A		435,783	260	b	A	A
305,407	010	a	A	A		410,790	260	c	A	A
112,156	019	a	[OCLC defined field]				300		M	M
	020		A	A		416,908	300	a	M	M
228,173	020	a	A	A		235,864	300	b	A	O
	040		M	M		391,899	300	c	M	O
415,423	040	a	A	A			500		O	O
419,657	040	c	M	M		210,250	500	a	M	O
454,451	040	d	A	A			504		O	O
	043		A	O		186,997	504	a	M	O
141,409	043	a	M	O			650		A	O
419,657	049	a	[OCLC defined field]			602,362	650	a	M	O
	050		O	O		326,867	650	x	A	O
300,385	050	a	M	O		231,459	650	z	A	O
285,578	050	b	A				651		A	O
	082		O	O		113,050	651	a	M	O
144,261	082	2	M	O		118,647	651	x	A	O
274,313	082	a	M	O			700		A	O
	100		A	A		176,916	700	a	M	O
318,692	100	a	M	M			850		O	O
149,540	100	d	A	A		110,257	850	a	M	O

Frequency	MARC 21 Field	Subfield	National Level Cataloging	Minimal Level Cataloging		Frequency	MARC 21 Field	Subfield	National Level Cataloging	Minimal Level Cataloging
	245		M	M						
419,641	245	a	M	M						

Acknowledgements

Support for this research has been provided by a National Leadership Grant from the Institute of Museum and Library Services. The following research assistants contributed to this analysis: Ed Kim, Jung Won Yoon, and Patrick Yeh.

References

[1] Dublin Core Metadata Initiative Website (2003). Available URL: http://dublincore.org/

[2] Network Development and MARC Standards Office. (2003). MARC Standards Website. Available URL: http://lcweb.loc.gov/marc/

[3] Z39.50 Interoperability Testbed Project Website. (2003). Available URL: http://www.unt.edu/zinterop

[4] Moen, William E. (2002, February). Indexing guidelines to support Z39.50 profile searches: Bath and U.S. national profiles, Functional Area A, Level 0 searches. Denton, TX: Texas Center for Digital Knowledge, University of North Texas. Available URL: http://www.unt.edu/zinterop/Documents/IndexingGuidelines1Feb2002.pdf

[5] National Information Standards Organization. (1994). ANSI/NISO Z39.2 - 1994 (R2001) information interchange format. Bethesda, MD: National Information Standards Organization. Available URL: http://www.niso.org/standards/index.html

[6] International Organization for Standardization. (1996). ISO 2709:1996 Information and documentation -- Format for information exchange. Geneva: International Organization for Standardization.

[7] Anglo-American Cataloguing Rules, Second Edition, 2002 Revision. (2002). Chicago: American Library Association.

[8] Network Development and MARC Standards Office. (2002, October). MARC 21 concise format for bibliographic data (Update No. 3). Washington, DC: Library of Congress, Network Development and MARC Standards Office. Available URL: http://www.loc.gov/marc/bibliographic/ecbdhome.html

[9] Network Development and MARC Standards Office. (2002, October). MARC 21 format for bibliographic data: Field list (Update No. 3). Washington, DC: Library of Congress, Network Development and MARC Standards Office. Available URL: http://www.loc.gov/marc/bibliographic/ecbdlist.html

[10] Moen, William E. (2002, January). Decomposing MARC 21 records for analysis. Denton, TX: Texas Center for Digital Knowledge, University of North Texas. Available URL: http://www.unt.edu/zinterop/Documents/DecomposingMARCRecordsFinalJan2002.pdf

[11] Network Development and MARC Standards Office. (2002, October). MARC 21 concise format for bibliographic data (Update No. 3). Washington, DC: Library of Congress, Network Development and MARC Standards Office. Available URL: http://www.loc.gov/marc/bibliographic/ecbdhome.html

[12] OCLC Online Computer Library Center, Inc. (2003). Bibliographic formats and standards. Dublin, OH: OCLC Online Computer Library Center, Inc. Available URL: http://www.oclc.org/bibformats/

[13] Network Development and MARC Standards Office. (2003, April). Metadata object description schema official web site. Washington, DC: Library of Congress, Network Development and MARC Standards Office. Available URL: http://www.loc.gov/standards/mods/

[14] Network Development and MARC Standards Office. (2002, October). MARC 21 format for bibliographic data: National level record--- bibliographic full level and minimal level. Washington, DC: Library of Congress, Network Development and MARC Standards Office. Available URL: http://lcweb.loc.gov/marc/bibliographic/nlr/nlr.html#intro

[15] Delsey, Tom. (2002). Functional Analysis of the MARC 21 bibliographic and holdings formats. Library of Congress, Network Development and MARC Standards Office. Available URL: http://www.loc.gov/marc/marc-functional-analysis/home.html