Exploration of Subject Representation and Support of Linked Data in Recently Created Library Metadata

Examination of Most Widely Held WorldCat Bibliographic Records

Vyacheslav Zavalin, Oksana L. Zavalina, and Shawne D. Miksa

This paper presents results of the examination of subject representation in the most recently created library metadata records. The bibliographic records were collected from the WorldCat global database. The records were created in 2020 according to the latest version of Resource Description and Access (RDA) and MARC 21Format for Bibliographic Data. A purposive sample of the records with the widest reach—as expressed in the highest number of holdings and the highest level of editing made by multiple institutions—was selected for in-depth content analysis. The level and patterns of application were analyzed for all subject representation data elements (record fields and subfields), specifically for those that were Linked-Data-enabling. The study examined the level and patterns of application of subject controlled vocabularies. Co-occurrences between various subject representation data elements and between subject controlled vocabularies within the records were explored.

Vyacheslav Zavalin (vzavalin@twu.edu) is an Assistant Professor in the School of Library & Information Studies at Texas Woman's University. **Oksana L. Zavalina** (oksana.zavalina@unt.edu) is an Associate Professor in the Department of Information Science at the University of North Texas. **Shawne D. Miksa** (shawne.miksa@unt.edu) is an Associate Professor in the Department of Information Science at the University of North Texas.

Manuscript submitted March 27, 2021; returned to authors for minor revision June 22, 2021; revised manuscript submitted July 18, 2021; accepted for publication August 4, 2021.

Helping users to satisfy their information needs and obtain needed information resources is the top priority in the field of library and information science. The representation of information objects through metadata is a key activity of libraries, archives and museums that is necessary to provide access to recorded knowledge held by those institutions. Several types of metadata records are used by these communities. Metadata records that represent information objects are commonly referred to as bibliographic records. The most common data traditionally included in bibliographic records are titles and subjects of works, plus the names of their creators.

In the current information environment, the amount of generated data and published information continues to rapidly increase and is often referred to as an information explosion, resulting in information overload.[1] As resource discovery by title or creator of an information object is seriously limited by this information

overload, resource discovery by subject becomes even more important.[2] This places an increasing emphasis on the functionality of subject metadata—the parts of bibliographic records that represent the aboutness of information objects.[3] The creation of subject metadata relies not only on analysis of aboutness, but also on examination of relationships among topics, form, and genre in the context of the intended audience and possible uses of information objects.[4]

MARC 21 Format for Bibliographic Data (MARC 21) is currently the dominant cataloging encoding format for description of information objects and the exchange of metadata among databases. Development of Linked Data potentially improves discoverability of information through metadata records, including subject access through subject metadata. The new BIBFRAME standard is developed with Linked Data functionality support in mind, and will eventually replace MARC 21. Millions of existing MARC 21 records that collectively represent and provide access to the vast body of recorded knowledge will need to be reformatted or converted from MARC 21 to BIBFRAME. Due to the sheer volume of that conversion task, the reformatting will need to be automated. As the output quality in automatic conversion processes relies greatly on the input quality, to ensure the conversion produces meaningful and functional results, the input metadata (data values in the fields of MARC 21 bibliographic records) needs to support that functionality. However, it is unclear as to what extent the Linked Data functionalities can be realized when the records are converted automatically from MARC 21 to BIBFRAME. This paper reports selected results of the exploratory study that sought answers to this question, with a focus on the subject representation in library bibliographic metadata records.

Literature Review

Bibliographic records are created according to several types of standards. Currently, the most widely used data content standard in the library community is Resource Description and Access (RDA).[5] The prevailing data encoding and transmission standards are the well-established MARC format, and the more recent alternative, Bibliographic Framework Initiative (BIBFRAME)—both of which are metadata element sets.[6] The data value standards include controlled vocabularies (e.g., thesauri, lists of subject terms and codes, etc.) and classification schemes.

RDA is an international standard that began to be developed in 2005 and was officially implemented by the Library of Congress (LC) in 2013. It was initially planned as a third major revision to the Anglo-American Cataloguing Rules (AACR), and evolved into a new standard with a substantially different conceptual base. RDA continues to evolve to meet user needs. The recently completed 2020 revision of RDA (3R) has not yet been widely adopted by metadata practitioners due to usability issues, and is expected to be adopted in 2023; most catalogers currently rely on the April 2018 revision.[7] Development of RDA is informed by the ideas of the Semantic Web, which seeks to connect pieces of information in a logical way that is understandable and can be processed by machines to improve information retrieval.[8] This way of connecting information is called Linked Data. One of the most important steps in ensuring its validity is the inclusion of unique Uniform Resource Identifiers (URIs) that link to openly available information on the entity in question and related entities.

BIBFRAME builds upon application of Linked Data principles to bibliographic metadata and is projected to replace the MARC 21 standard. BIBFRAME metadata record creation tools are being developed and explored by the early adopters, and software companies are starting to incorporate them into the integrated library systems such as for example Ex Libris' Alma. Until these tools become mainstream (a process that will take years, if not decades), most newly created records will follow the MARC 21 standard. Currently, hundreds existing MARC 21 bibliographic records collectively provide access to the body of recorded knowledge, and MARC 21 maintains its importance as an encoding standard.[9] Furthermore, the MARC 21 bibliographic metadata element set constantly evolves as new fields and subfields are added to support the Linked-Data-related and other RDA requirements.

Since the beginning of RDA's development, several new subject representation data elements have been added to the MARC 21 bibliographic element set to improve functionality and to support Linked Data. As part of these revisions, MARC 21 has been expanded to include new subfields in variable fields that enable the inclusion of URIs into bibliographic records. The 650 (Subject Added Entry—Topical Term) is one of the fields for which the subfield $4 Relationship was initially added to MARC 21 Bibliographic Standard in 2007. This subfield was later renamed as Relator Code and redefined in 2017 to include URIs as data values. Additionally, in 2017 the subfield $1 Real World Object URI, was added to MARC 21 Bibliographic Format for this use in several fields. The subfield $0 Authority record control number that had been part of the standard since 2005 was emphasized after 2013 as the way to link bibliographic records to authority records. The library community is working to improve the Linked Data functionality of existing MARC 21 bibliographic records by enriching them with URIs, including subject metadata fields.[10] Recent publications include reports of converting the Program for Cooperative Cataloging (PCC) MARC 21 records to BIBFRAME, comparative evaluations of Linked Data ontologies and data models as they apply to MARC,

and discussion of the future of authority control in libraries in the Linked Data environment.[11] Zeng and Mayr shared results of their review of how existing knowledge organization systems used in library metadata (including subject controlled vocabularies) can transition to become Linked Open Data.[12]

The principles of building subject controlled vocabularies have been developed and refined in communities of information professionals for many years, with the first major controlled vocabularies appearing in the nineteenth century. In the twenty-first century, construction of subject controlled vocabularies is guided by the International Standard for Thesauri and Interoperability with Other Vocabularies, currently in version 1.4.[13] More than three hundred controlled vocabularies for subject representation have been developed and are maintained worldwide.[14] Some are multilingual, such as the French-English Répertoire de vedettes-matière (RVM), which is used for verbal subject representation in Canada. Subject controlled vocabularies are often developed by national libraries or archives such as the Gemeinsame Normdatei, which is developed by the German National Library or the US Library of Congress Subject Headings (LCSH). Faceted Application of Subject Terminology (FAST) is a derivative controlled vocabulary that relies on LCSH, and provides an added level of functionality by splitting LCSH subject strings into facets. Book Industry Study Group (BISAC) subject headings are another subject controlled vocabulary of general applicability intended for use by publishers and bookstores.[15] A number of subject controlled vocabularies focus on specific domains: one example is the Medical Subject Headings (MeSH) used to represent works originating in the biomedical knowledge domain.

Controlled vocabularies for subject representation include the broad classifications that cover the entirety of human knowledge, such as DDC, Library of Congress Classification (LCC), Universal Decimal Classification (UDC), and classification systems that focus on specific knowledge domains (e.g., National Agricultural Library classification for agricultural materials, Government of Canada Classification for government publications, etc.). Some subject controlled vocabularies combine verbal and nonverbal subject representation—one example is the Chinese Classified Thesaurus (Zhong guo fen lei zhu ti ci biao).[16]

Tools and technologies that enable automatic and semi-automatic generation of metadata from the full-text of textual information objects use indexing to help expand access to information. Despite its advantages, full-text indexing cannot provide the same level of access as subject representation with controlled vocabularies. Research shows deficiencies for information retrieval, for example, in representing foreign language materials.[17] They are also not useful for creating metadata for non-textual information objects (e.g., works of music, visual art, photographs, etc.).

Smith-Yoshimura et al. emphasized providing controlled-vocabulary subject access in the creation of MARC 21 records: "The number of full-text documents available on the Web will substantially increase over the next few years, and the need for surrogate 'descriptive metadata' will decrease. Focus instead on the authorized names, classifications, and controlled vocabularies that key word searching of full-text will not provide."[18] As of 2015, automated indexing tools were considered as not yet sufficiently developed for full-scale implementation by the library community in a meaningful way.[19] However, a promising new multilingual automated subject indexing tool Annif has been developed and used by early adopters in the international library community for digital collections in 2017-2020.[20]

Studies of MARC 21 library metadata typically draw datasets for analysis from large databases such as the LC's catalog or OCLC's WorldCat. The advantage of the WorldCat database compared to a single library catalog (even as large as the LC's) as a source of data for studying worldwide cataloging practices is its heterogeneity and global impact. Since 1998, WorldCat has been the major centralized shared database of bibliographic records created and edited collaboratively by the international library community. The WorldCat database is widely used in fulfilling interlibrary loan (ILL) requests.[21] It is also a major tool used in cooperative cataloging worldwide: when a library or other institutional member of WorldCat adds an item to its collection, it either

- submits a new bibliographic record to the database; or
- if the record is already in WorldCat, uses the existing record is used as is or edits the master record and updates the holdings information by indicating that it has the item in its collection.

In both copy and original cataloging, bibliographic records are also added to the institution's online catalog. As this paper was being finalized, a new record was added to the WorldCat database every second, and contained more than 516 million metadata records representing information objects in 483 languages.[22]

Several studies of MARC 21 bibliographic metadata have examined subject representation in library metadata records. Almost all were completed pre-RDA, meaning that they were conducted before new subject data elements (fields and subfields, including Linked-Data enabling ones) were added to the MARC 21 standard and applied to bibliographic records. Furthermore, most of the studies conducted since 2000 did not focus on subject metadata. Relevant findings from these studies are reviewed below.

Moen and colleagues found that the MARC 21 Field 600 Subject Added Entry—Personal Name is the one

most frequently occurring subject-related field in MARC 21 bibliographic records.[23] Eklund et al. observed that the MARC field 655 Index Term—Genre/Form was present in only 5 percent of records for sound recordings.[24] Mayernik observed that the MARC field 650 Subject Added Entry—Topical Term appeared in 66 percent of records and exhibited the largest average number of occurrences (1.84 per record), and that other subject representation fields–050 Library of Congress (LCC) Call Number, 043 Geographic Area Code, and 082 Dewey Decimal Classification (DCC) Number–were among the most frequently occurring MARC 21 fields.[25] Smith-Yoshimura et al. noted that four subject metadata fields were among the top twenty-two most frequently occurring MARC 21 fields: 650 (46 percent of records), 050 (20 percent), 043 (19 percent), and 082 (14 percent). Smith-Yoshimura's team also separately examined the application of fields recently added to the MARC 21 standard as of the time of their analysis and found subject metadata fields 648 Subject Added Entry-Chronological Term and 662 Subject Added Entry—Hierarchical Place Name to be used in under 0.1 percent of records.[26] Moen and Benardino examined 400,000 MARC 21 bibliographic records and observed 122 different MARC 21 subject metadata subfields (e.g., 650 $v Subject Added Entry Topical Term—form subdivision, 651 $y Subject Added Entry Geographic Term—chronological subdivision, etc.).[27]

Taylor and Simpson compared LC's Cataloging-In-Publication records with other bibliographic records, and found mistakes and omissions in subject headings, geographic area codes, DDC and LCC classification codes.[28] In her analysis of records from two databases, Intner discovered a lack of subject headings or classification numbers in records.[29] In the meta-analysis of subject search in online catalogs, Larson summarized improvements in subject representation in bibliographic records that had been proposed by researchers. These suggestions included assigning more LCSH headings per record, supplementing them with terms from specialized thesauri (e.g., MeSH), providing more specific class notations, assigning additional class numbers to represent multiple facets of a work, etc.[30] Hoffman examined the practice of facilitating subject access through the creation of individual bibliographic records with more specific subject headings for each work aggregated in a multi-work item instead of assigning more general subject headings in a single record describing the whole item.[31]

A more recent study by Zavalina, Shakeri, and Kizhakkethil examined RDA-based MARC 21 bibliographic records to determine the quantitative patterns of change between 2013 and 2015 in the application of subject metadata fields in video recording records. That study found a slight increase in the use of Linked-Data-enabling subfields, but reported low overall level of their application. The authors observed the overall trend toward an increase in the average number of subject headings per record.[32] The replication of these analyses in 2021 to compare the 2020 versions of the same records representing video recordings to their 2015 versions demonstrated an increase in the trend for addition of further subject fields and subfields to existing records. However, that study found some decrease in the level of application of Linked-Data-supporting subfield $0 Authority record control number or standard number in field and/or in the average number of instances of this subfield per record that includes the field in five subject metadata fields.[33] Zavalin examined the application of subject and genre controlled vocabularies in a sample of 688 WorldCat bibliographic records contributed by the LC's Children's and Young Adults' Cataloging (CYAC) Program between 2014 and 2020. The author observed the use of twenty genre and eighteen subject controlled vocabularies.[34]

None of the previous studies of subject representation in MARC 21 bibliographic records examined the use of the co-occurrences of data elements and controlled vocabularies within a record. Additionally, no published studies of MARC content designation focused on the examination of subject metadata in records created after the major revision of RDA in 2018 and following the latest addition of a new subject field to the MARC 21 Bibliographic format: the field 688 Subject Added Entry—Type of Entity Unspecified in 2019. This study seeks to be one of the first to analyze records following these and other recent changes to library metadata standards to develop an understanding of the current level and patterns in subject representation, including support for Linked Data, as evidenced in a set of MARC 21 bibliographic records that are included in a large number of library catalogs.

Method

This study used the content analysis of the recently created RDA-based MARC 21 bibliographic metadata records in the WorldCat database. A purposive sample of 100 MARC 21 records that were created in 2020, and thus were expected to follow the most recent versions of RDA and MARC 21 standards, was selected, based on several major criteria. First, records that represent information objects that were held in at least 500 library collections at the time of data collection in May 2020 were targeted. Records with such a high level of holdings have the greatest impact on access to information (including subject access) in library collections. Also, regardless of age, these widely held records are typically edited more than once by multiple institutions since their creation. Second, records were selected with the highest overall quality as indicated by the "full level of encoding" code in the ELvl subfield of the fixed field. Specifically,

the authors targeted those with the code blank as "the most complete MARC record[s] created from an inspection of the material" or those with the code "I," indicating the next most complete full-encoding level.[35]

The authors used MarcEdit, a metadata manipulation and editing software suite, to collect the data using the Z39.50 client-server protocol developed for searching and retrieving information from remote databases through Transmission Control Protocol/Internet Protocol (TCP/IP) supporting networks by LC's Maintenance Agency. The application of criteria discussed above and the deduplication of the list of records matching these criteria resulted in a set of one hundred unique metadata records.

An in-depth content analysis of these metadata records was performed. The study examined general characteristics such as types and languages of materials represented, types and locations of institutions that created records, language of cataloging, etc. The focus was on the levels of application of various subject fields and selected subfields, including Linked-Data-supporting data elements, co-occurrence between subject data elements intended for the same type of information within a record, and levels of application of subject controlled vocabularies and co-occurrence within records.

A common limitation of content analysis is researcher bias, which is normally alleviated by using detailed coding manuals, coding by multiple coders, and subsequent evaluation of the intercoder agreement. However, the design of this study bypassed researcher bias because only objective (i.e., mostly numeric and binary) characteristics and measures were assessed. No subjective evaluations (e.g., those regarding the accuracy of subject metadata) were included.

Findings

General Characteristics of Records in the Sample

Based on the holdings data attached to the records collected from WorldCat, the number of institutions that included the analyzed records in their catalogs at the time of data collection ranged between 577 and 1,514. The material types represented by the collected records were distributed as follows: books (83 percent, including regular print books, 76 percent, online books 3 percent and large print books 4 percent), visual materials (13 percent, including online materials 1 percent), sound recordings (3 percent), and continuing resources (1 percent). Forty percent of the records were created as part of LC's Copy Cataloging program (lccopycat), with headings "verified with the relevant authority file, except those subject headings not from LCSH."[36] An additional 8 percent were created under the Program for Cooperative Cataloging's (PCC) auspices, which means that "subject headings are checked for authorized forms and combinations supported by the relevant authority."[37]

The sampling approach did not limit data collection to any specific language of items represented by records or any specific language of cataloging. However, analysis demonstrated that all records in this sample of the most widely held WorldCat bibliographic records with the highest self-identified completeness represented only English-language materials. The records were created by thirty-one institutions from six countries, with English as the language of cataloging: Australia, Canada, Hong Kong, New Zealand, the United Kingdom, and the US. The records were contributed to WorldCat by academic libraries (e.g., University of Hong Kong's library), school libraries (e.g., Anchorage school district library in Alaska), public libraries (e.g., Winnipeg Public Library), state/national libraries (e.g., Libraries Australia), federal/national government agencies (e.g., US National Library of Medicine), associations/foundations (e.g., Libraries Horowhenua in New Zealand), and vendors and other corporate/business organizations (e.g., Baker & Taylor). The number of records contributed by each institution ranged from one to twenty-nine, with 3.225 on average.

Application of Subject Representation Data Elements

Table 1 shows the level of application of observed MARC 21 subject representation metadata fields. The dataset contained a total of eighteen MARC 21 bibliographic metadata fields for subject representation (see table 1). At the time of data collection, all but one of these subject metadata fields, field 043 Geographic Area Code, were repeatable, meaning that more than one instance of a field could be included in a bibliographic record. However, with the December 2020 publication of Update no. 31 to MARC 21 Format for Bibliographic Data, the 043 also became repeatable.[38]

As shown in table 1, only the 650 Subject Added Entry—Topical Term field was present in all records. The number of instances of this field varied between two and forty-six instances. Three other fields appeared in 98 percent of the records. This included two fields that provide classification data and one that represents genre: 050 Library of Congress Call Number, 082 Dewey Decimal Classification Number, and 655 Index Term—Genre/Form. The level of application of the remaining fourteen subject representation fields in MARC 21 bibliographic records ranged widely between 1 percent (fields 080 Universal Decimal Classification Number, 092 Locally Assigned Dewey Call Number, and 654 Subject Added Entry—Faceted Topical Terms) and 59 percent of records (field 651 Subject Added Entry—Geographic Name). The average number of

Table 1. Level of application of observed MARC 21 subject representation metadata fields

Field	% of Records with 1+ Instance	Ave. No. of Instances per Record	Median	Mode	Range	Variance	Standard Deviation
043 Geographic Area Code	53	1.0000	1	0	0-1	0.2516	0.5016
050 Library of Congress Call Number	98	1.0204	1	1	0-2	0.0404	0.2010
055 Classification Numbers Assigned in Canada	5	1.0000	0	0	0-1	0.0480	0.2190
060 National Library of Medicine Call Number	3	1.0000	0	0	0-1	0.0294	0.1714
072 Subject Category Code	2	1.5000	0	0	0-2	0.0496	0.2227
080 Universal Decimal Classification Number	1	1.0000	0	0	0-1	0.0100	0.1000
082 Dewey Decimal Classification Number	98	1.0204	1	1	0-2	0.0404	0.2010
084 Other Classification Number	12	1.0833	0	0	0-2	0.1344	0.3667
092 Locally Assigned Dewey Call Number	1	1.0000	0	0	0-1	0.0100	0.1000
600 Subject Added Entry-Personal Name	28	2.3571	0	0	0-6	1.5600	1.2490
610 Subject Added Entry-Corporate Name	7	2.0000	0	0	0-2	0.2630	0.5129
611 Subject Added Entry-Meeting Name	4	1.2500	0	0	0-2	0.0682	0.2611
647 Subject Added Entry-Named Event	6	1.1667	0	0	0-2	0.0860	0.2932
648 Subject Added Entry-Chronological Term	16	1.0000	0	0	0-1	0.1358	0.3685
650 Subject Added Entry-Topical Term	100	13.3500	12	12	2-46	56.3914	7.5094
651 Subject Added Entry-Geographic Name	59	2.5424	1	0	0-8	2.8586	1.6907
654 Subject Added Entry-Faceted Topical Terms	1	3.0000	0	0	0-3	0.0900	0.3000
655 Index Term-Genre/Form	98	6.9286	7	8	0-19	14.5110	3.8093

Table 2. Total number of various subject metadata fields and field instances per record

	Mean	Median	Mode	Range	Variance	Standard Deviation
No. of Different Subject Fields per Record	5.99	6	6	3-10	2.51505	1.585891
Total No. of Instances of All Subject Fields per Record	25.7	26	27	5-68	80.2929	8.960632

instances of fields 650 (13.35) and 655 (6.93) was the highest. Subject representation fields 651, 600 (Subject Added Entry—Personal Name), and 610 (Subject Added Entry—Corporate Name) appeared in two or more instances per record on average (2.54, 2.35, and 2.00). The highest level of variability as expressed in standard deviation of 7.51 was observed in field 650. Also, relatively high standard deviation between 1.25 and 3.81 was observed in three additional fields: 655 (Index Term—Genre/Form), 651, and 600. In the remaining fourteen fields, the standard deviation was below 0.6, which indicates consistent levels of application of a field across the records in the purposive sample.

Table 2 shows that on average, a total of six various subject fields appeared in records, with the range of three to ten. The total number of instances of all subject fields combined ranged much more substantially from five to sixty-eight per record. The central tendency measures—mean, median, and mode—for the number of instances of all subject fields combined per record were between 25.7 and 27. The analysis demonstrated high variability for the total number of subject field instances (variance of 80.29 and standard deviation of 8.96) and relatively moderate variability for the number of subject fields (variance of 2.52 and standard deviation of 1.59).

Ninety-eight percent of records in the sample included one or more instances of subfield $0 Authority Record Control Number or Standard Number, which is considered as the most important Linked-Data-enabling MARC 21 subfield.[39] A total of 778 instances of this subfield, as shown in figure 1, appeared in seven subject metadata fields: 600, 610, 611, 647, 650, 651, and 655. Almost 85 percent total of all instances of subfield $0 in subject representation fields occurred in the two most widely applied fields: 650 (48.2 percent) and 655 (36.89 percent). All instances of subfield $0 Authority Record Control Number or Standard Number observed in subject representation metadata fields included data values expressed as literals as opposed to URIs. The Linked-Data-enabling subfield $0 appeared only in the instances of 6XX

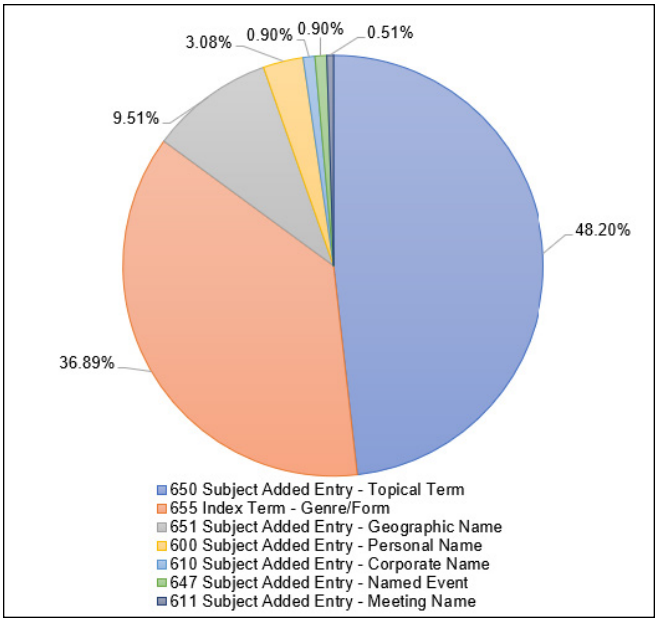

Figure 1. Distribution of Linked-Data-enabling subfield $0 Authority Record Control Number or Standard Number (% of all instances of subfield $0 observed in the sample)

fields that contained FAST headings (with exception of when FAST headings were used in field 648 Subject Added Entry-Chronological Term where it was not observed). The $0 subfield was not observed in any instances of 6XX fields that included terms from the other subject controlled vocabularies: LCSH, LC Children's Subject Headings (CSH), Medical Subject Headings (MeSH), Répertoire de Vedettes-Matière (RVM), BISAC Subject Headings, Guidelines on Subject Access to Individual Works of Fiction, Drama, Etc. (GSAFD), Library of Congress Genre and Form Thesaurus (LCGFT), Sears subject headings, Gemeinsame Normdatei (GND), and GOO-Trefwoorden Thesaurus by Koninklijke Bibliotheek in the Netherlands (GTT). No record in the sample included any of the two other Linked-Data-enabling subfields in subject metadata fields: $1 Real World Object URI and $4 Relationship.

The application of three additional subfields in subject representation fields—043 $a Geographic Area Code and subfields $z Geographic Subdivision and $y Chronological Subdivision in 6XX fields—was examined and compared to the application of other MARC 21 subject metadata elements intended for representing chronological and geographical aboutness of information objects. Table 3 presents the overall levels of application of these subfields. The largest number of instances was observed for $z Geographic Subdivision: it occurred seventy-two times in a total of thirty-three records in the sample. Subfield 043 $a Geographic Area Code occurred in a larger proportion of records (53 percent) but in a smaller overall number of instances (sixty-two). Subfield $y Chronological Subdivision was the least frequently used: sixteen instances total were observed in 9 percent of records. The average number of instances was the lowest for 043 $a (1.1698) and the highest for 6XX $z (2.1818). The mode number of instances was zero for all three subfields, and only the 043 $a exhibited a median number of instances above zero. The highest variability in the level of application was observed for 6XX $z.

Application of Controlled Vocabularies

LCSH was used most often to represent subject content of information resources represented by the records. It was observed in six MARC 21 subject representation metadata fields: 600, 610, 611, 650, 651, and 655. Table 4 provides the level of application of LCSH controlled vocabulary in these fields. The highest level of use of the LCSH occurred in fields 650, 600, and 651: an average of 4.05, 1.44, and 1.13 instances of the field respectively. However, the median and mode number of instances of a field with data values

Table 3. Level of application of three subject metadata subfields

Field	% of Records with 1+ Instance	Ave. No. of Instances per Record	Median	Mode	Range	Variance	Standard Deviation
043 $a	53%	1.169811	1	0	0-4	0.693112	0.480404
6XX $z	33%	2.181818	0	0	0-9	1.484465	2.203636
6XX $y	9%	1.777778	0	0	0-7	0.76171	0.580202

Table 4. Level of application of the LCSH controlled vocabulary

Subject Representation Field	% of Records with 1+ Instance	Ave. No. of Instances per Record	Median	Mode	Range	Variance	Standard Deviation
600	27%	1.4444	0	0	0-3	0.5635	0.7507
610	7%	1.0000	0	0	0-1	0.0658	0.2564
611	1%	1.0000	0	0	0-1	0.0100	0.1000
650	99%	4.0505	4	3	0-15	5.0403	2.2451
651	36%	1.1389	0	0	0-2	0.3453	0.5877
655	34	1.0294	0	0	0-2	0.25	0.5000

from LCSH, as well as variance and standard deviation, were under 0.76 for all except field 650.

Figure 2 shows the application of other subject controlled vocabularies in 6XX fields. A total of ten other subject controlled vocabularies was observed. The two most widely applied (i.e., found in over 90 percent of records) non-LCSH controlled vocabularies were the Faceted Application of Subject Terminology (FAST) and the Library of Congress Genre/Form Terms for Library and Archival Materials (LCGFT). Terms from four others—BISAC Subject Headings List (BISAC), Sears List of Subject Headings (SEARS), Guidelines on Subject Access to Individual Works of Fiction, Drama, Etc. (GSAFD), and Children's Subject Headings (CSH)—were found in between 37 percent and 72 percent of records. The Medical Subject Headings (MeSH), Gemeinsame Normdatei (GND), Répertoire de vedettes-matière (RVM), and GOO-Trefwoorden Thesaurus (GTT) were used much less often: between 1 percent to 4 percent of records. Finally, 12 percent of records included one or more instances of 6XX field(s) with the controlled vocabulary not specified.

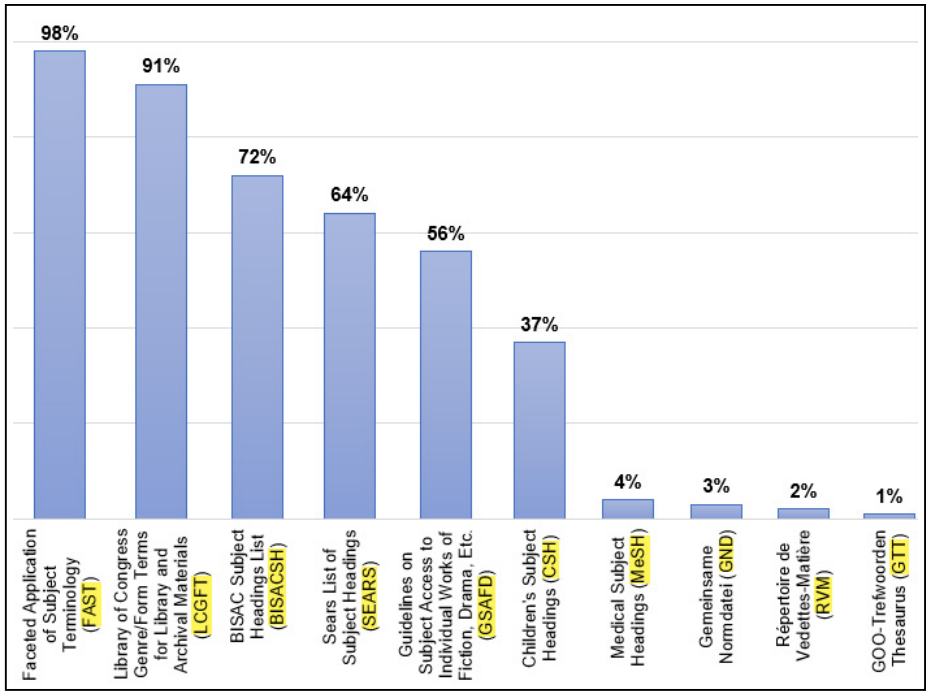

Figure 2. Level of application of other subject controlled vocabularies in 6XX fields

Two classification fields—072 Subject Category Code and 084 Other Classification Number—were found to contain source code "bisacsh" indicating the terms were drawn from BISAC Subject Headings List. A total of six instances of field 072 in three records (100 percent of records with that field in the sample) and a total of twelve instances of field 084 in eleven records (91.67 percent of records with that field in the sample) included BISAC terms. Between one and two instances of three other controlled vocabularies used in these fields were observed for: (1) Book Industry Communication Standard Subject Categories (indicated by code "bicssc"), (2) Nederlandse Basisclassificatie (Dutch Basic Classification Code; indicated by code "bcl"), and (3) Elizabeth M. Moys Classification and Thesaurus for Legal Materials (indicated by code "moys").

Co-occurrence of Subject Data Elements and Controlled Vocabularies

As previously shown in table 2, between three and ten different subject fields were observed in each record, with an average of six. Certain pairs of subject metadata elements (fields/subfields combinations) providing similar or related types of information often co-occurred. Most (94 percent) records included two classification fields: 050 Library of Congress Call Number and 082 Dewey Decimal Classification Number. The co-occurrence between these two fields was the second highest, after the pair of fields 650 Subject Added Entry-Topical Term and 655 Index Term-Genre/Form that occurred together in 98 percent of records. Analysis also revealed noticeable levels of co-occurrence of

- field 648 Subject Added Entry-Chronological Term and subfield $d Date of meeting or treaty signing;
- field 611 Subject Added Entry-Meeting Name (84 percent of records in the sample);
- field 648 and subfield $y Chronological subdivision in 6XX fields (83 percent);
- field 043 Geographic Area Code and subfield $z Geographic subdivision in 6XX fields (43 percent); and
- fields 043 Geographic Area Code and 651 Subject Added Entry-Geographic Name (39 percent).

Co-occurrences between other subject fields (e.g., 650 and 651, 600 and 610, etc.) and pairs of classification fields other than 050 and 082 was much lower.

Certain pairs of subject controlled vocabularies were used together in the same records. Figure 3 presents these findings for most frequently co-occurring controlled vocabularies. LCSH and FAST were found most often together

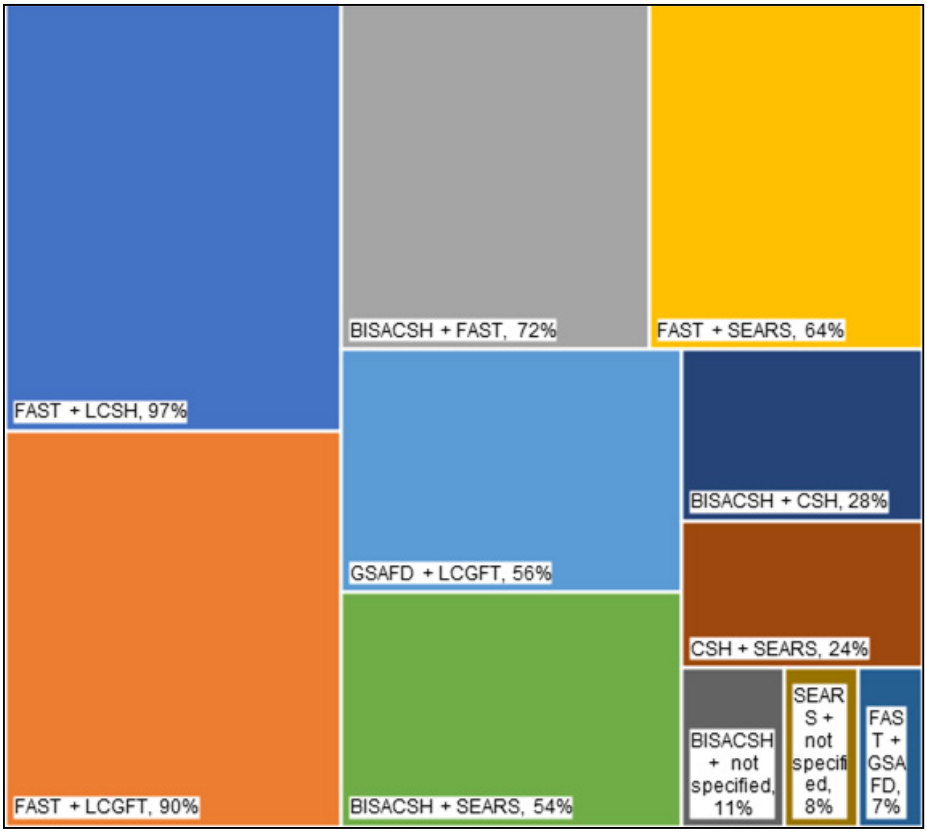

Figure 3. Co-occurrence of subject controlled vocabularies within the same records

(97 percent of records). Both FAST and LCGFT terms were included in 90 percent of records. Four additional pairs of controlled vocabularies co-occurred in more than 50 percent of records overall: FAST and BISAC subject headings (72 percent), FAST and Sears subject headings (64 percent), LCGFT and GSAFD genre headings (56 percent), and Sears and BISAC subject headings (54 percent). Although not shown in figure 2, it is worth noting that the lowest level of co-occurrence (1 percent of records in the sample) was observed for the terms from MeSH and BISAC subject headings, MeSH and Sears subject headings, Répertoire de Vedettes-Matière (RVM) and BISAC subject headings, and RVM and Sears subject headings. No co-occurrences were observed between MeSH and other controlled vocabularies beyond BISAC and Sears, or between RVM and other controlled vocabularies beyond BISAC and Sears.

Discussion

Patrick Wilson posited that

> the sense of a position in an organizational scheme is given by the rules of assignment and by what we can deduce from those rules. When position is assigned on the basis of identification of the subject and selection of the most closely fitting position, whatever sense we have of positions depends on what we know about how it is decided what the subject of a writing is, hence what it means to say of a writing that its subject is this or that.[40]

This study is the first to provide insight into the patterns of application of subject representation in the MARC 21 records created by libraries worldwide using the latest revisions of RDA and MARC 21 metadata element set to facilitate increased Linked-Data functionality of library metadata. The findings indicate that the available MARC 21 content designation intended to support this functionality is not currently being used to its full capacity. Only one of the three Linked-Data-enabling subfields was observed in the analyzed records, with URIs for the terms from just one of the controlled vocabularies (FAST). This omission means that when MARC 21 records are converted to BIBFRAME 2.0, URIs for most controlled-vocabulary terms would not be included, and for subject representation other than that with FAST (based on LCSH), records would mostly rely on literal data values (strings of characters) that have no Linked Data power.

Overall, the findings demonstrate that subject representation has substantially increased in extent and variety compared to MARC 21 metadata created earlier and analyzed in previous studies conducted between 2003 to 2010.[41] There is an especially noticeable increase in the level of application observed for fields 650 Subject Added Entry-Topical Term, 655 Index Term-Genre/Form, and 651 Subject Added Entry-Geographic Name. The practice of enriching records by adding non-LCSH subject terms from a variety of controlled vocabularies of topical terms observed in the records analyzed in this study significantly expands subject representation in records, and, if accompanied with Linked-Data-enabling metadata elements, greatly increases functionality of bibliographic records in supporting the Explore user task as defined in the *Library Reference Model* (LRM).[42]

Eighteen of thirty-seven subject metadata fields defined in the latest version of MARC 21 metadata element

set were found in the records analyzed for this study. Both the average number of various subject fields included in records (5.99), and the average number of instances of these subject fields (25.7) are substantially higher in RDA compliant records created in 2020 than was observed in the previous studies of pre-RDA MARC 21 bibliographic records. However, it is important to note that several new subject metadata fields have been added to the MARC 21 metadata element set since the time these pre-RDA studies were completed. These include fields 083, 085, 086, and 688 that were not observed in the present study, plus fields 084 and 647, which were observed. Also, surprisingly, even within this purposive sample of the most complete cataloging records based on the full-level cataloging standard followed, a substantial variability was observed (as measured through variance and standard deviation indicators) in the application of several key subject metadata fields, including the MARC 650 Subject Added Entry-Topical Term and MARC 655 Index Term-Genre/Form.

The findings of this in-depth content analysis suggest several recommendations to the library metadata community, including creators and editors of RDA-based MARC 21 records, LC's Linked Data Service, OCLC, and developers of controlled vocabularies. Implementation of these recommendations will result in a stronger support of Linked Data in bibliographic metadata. The first recommendation is to include subfield $0 with an authority record ID number for all instances of field 043, which contains terms from Geographic Area Code controlled vocabulary (currently available through LC's Linked Data Service portal), field 655, which contains LCGFT genre headings (also available through LC's Linked Data Service portal), and field 648, which contains FAST chronological facet terms.[43]

Second, the authors advocate for revising the algorithm for automatic generation of FAST headings from those available in LCSH subject strings in fields 650 and 651 to also include fields 600, 610, and 611, which have chronological terms in subfields $d and/or geographical terms in subfield $c. This will result in automatically generating field 648 with chronological facet terms from FAST controlled vocabulary and field 651 with geographic name facet term from FAST controlled vocabulary. Another recommendation is to consistently include field 084 or field 072 with BISAC subject codes from BISAC controlled vocabulary whenever BISAC subject headings are used in field 650.

Finally, the authors recommend adding the most frequently used non-LCSH-based lists of subject headings—BISAC and SEARS—to the LC Linked Data Service Portal in Linked Data form with unique record IDs. When these controlled vocabularies become available in Linked Data form, it would be possible to add subfields $0 in field 650 instances that contain SEARS and BISAC subject headings.

The recommended steps are based on issues observed during in-depth analysis of records in this study, and addressing these deficiencies is expected to substantially improve subject access in general and Linked Data functionality of subject representation in bibliographic metadata in particular. The authors of this paper realize that in practical terms, the recommended steps will increase the workload of the cataloging agencies and would require additional resources to implement. While the authors believe (and many of their colleagues would agree) that the projected gains in subject access and Linked Data functionality support for the users are worth additional efforts, discussion is needed on the most logistically sound and cost-effective ways to approach these tasks.

Conclusion

This exploratory study aimed to address existing gaps in research and practice related to subject representation and Linked Data support in bibliographic metadata. It used in-depth content analysis of widely held RDA-based MARC 21 records in the WorldCat database. The study provides insight into the patterns of application of subject representation in the MARC 21 records created by libraries worldwide using the latest revisions of RDA and MARC 21 metadata elements to facilitate increased Linked Data functionality of library metadata. The findings indicate that the available MARC 21 content designation intended to support this functionality is not currently used to full capacity, and specific practical recommendations for addressing this gap are provided in the Discussion section of this paper. However, despite the observed limitations, overall, results of this study demonstrate that subject representation has substantially increased in extent and variety. The questions still remain about the extent to which this increased subject representation and Linked Data functionality supports the evolving user needs and impacts access to information. Future user studies will need to explore these questions.

The content analysis study presented here has several limitations that need to be addressed in future research. It is worth noting that the records analyzed were created in the first half of 2020 and therefore might reflect the changes in cataloging practices brought to light by realities of the COVID-19 pandemic and resulting adjustments to cataloging workflows to accommodate remote work. It is possible that because many catalogers in 2020 (as well as in much of 2021) worked remotely without direct access to non-digital collections, the emphasis shifted to refining existing records as opposed to creating new ones. A study of records created in the previous years but revised in 2020–21 might shed light on these trends and their effect on the completeness and overall quality of bibliographic records, and on

the degree of their support for Linked Data functionality. Another possibility is that the COVID-19-related adjustments in workflows resulted in scaled back cataloging with the intent to revisit it once catalogers were back on-site. Examination of the records created after the world largely emerges from the pandemic and those created earlier but last updated in late 2022 and beyond would allow to assess the impact of those trends on record creation and enrichment activity levels.

This study relied on a purposive sample of the one hundred most widely held (with five hundred or more holding institutions) RDA-based MARC 21 bibliographic records created in 2020 with the highest level of completeness. The study demonstrated that this group of records did not include records for materials in languages other than English or records that were cataloged in non-English languages. This limitation did not enable the authors to comparatively evaluate subject metadata, including application of Linked-Data-enabling subfields, for different groups of records based on the language of cataloging or language of materials. Future studies will address this limitation by analyzing large diverse samples of records. Additionally, there is a need to monitor the trends in subject representation practices, so future studies will compare the records created in 2020 with the records created in 2021 and beyond, including those created with the latest revision of RDA(3R). Moreover, comparing records analyzed in this study with the revised versions of the same records will help to trace changes in subject representation and, ideally, find more and stronger subject representation overall in our library information systems.

References

1. Ray R. Larson, "Between Scylla and Charybdis: Subject Searching in Online Catalogs," *Advances in Librarianship* 15, no. 1 (1991): 175–36.
2. Marcia J. Bates, "After the Dot-bomb: Getting Web Information Retrieval Right this Time," *First Monday* 7, no. 7 (2002): 1–2, https://firstmonday.org/ojs/index.php/fm/article/view/971/892.
3. Patrick Wilson, *Two Kinds of Power: An Essay on Bibliographic Control* (Berkeley: University of California Press, 1968), 69.
4. Daniel N. Joudrey, Arlene G. Taylor, and David P. Miller, "Chapter 11: Subject Access," in *Introduction to Cataloging and Classification*, 11th ed. (Santa Barbara, CA: Libraries Unlimited, 2015): 521–34.
5. RDA Steering Committee, *Resource Description and Access* (Chicago: American Library Association; Ottawa: Canadian Library Association; London: Chartered Institute of Library and Information Professionals (CILIP), 2010.
6. Library of Congress, *Bibliographic Framework Initiative*, (2021), accessed March 26, 2021, https://www.loc.gov/bibframe/.
7. RDA Toolkit, "Completion of 3R Project," October 15, 2019, https://www.rdatoolkit.org/node/202; Melanie Wacker, "Revised RDA Implementation Timeline: PCC Implementation Will Not Begin before July 2022," Program for Cooperative Cataloging List (PCCLIST), November 10, 2020, https://listserv.loc.gov/cgi-bin/wa?A2=ind2011&L=PCCLIST&P=10930.
8. *The Digital Future of the United States, Part I, the Future of the World Wide Web*, Before the U.S. House of Representatives, Committee on Energy and Commerce, Subcommittee on Telecommunications and the Internet, 110th Cong. (2007) (statement of Sir Tim Berners-Lee, CSAIL Decentralized Information Group, Massachusetts Institute of Technology), accessed March 25, 2021, https://www.govinfo.gov/content/pkg/CHRG-110hhrg39604/html/CHRG-110hhrg39604.htm.
9. Magda El-Sherbini, "RDA Implementation and the Emergence of BIBFRAME," *JLIS.it* 9, no. 1 (2018), https://www.jlis.it/article/view/66-82.
10. Diane L. Boehr and Barbara Bushman, "Preparing for the Future: National Library of Medicine's Project to add MESHRDF URIs to its Bibliographic and Authority Records," *Cataloging & Classification Quarterly* 56, nos. 2-3 (2018): 262–72; Carol Jean Godby and Ray Denenberg, *Common Ground: Exploring Compatibilities between the Linked Data Models of the Library of Congress and OCLC* (Dublin, Ohio: Library of Congress and OCLC Research, 2015), https://www.oclc.org/content/dam/research/publications/2015/oclcresearch-loc-linked-data-2015-a4.pdf; Jungran Park and Andrew Brenza, "Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art," *Information Technology & Libraries* 34, no. 3 (2015): 22–42, http://ejournals.bc.edu/ojs/index.php/ital/article/view/5889; Jackie Shieh and Terry Reese, "The Importance of Identifiers in the New Web Environment and Using the Uniform Resource Identifier (URI) in Subfield Zero ($0): A Small Step that is Actually a Big Step," *Journal of Library Metadata* 15, nos. 3–4 (2015): 208–26; Jackie Shieh, "PCC's Work on URIs in MARC," *Cataloging & Classification Quarterly* 58, nos. 3–4 (2020): 418–27.
11. Jacquie Samples and Ian Bigelow, "MARC to BIBFRAME: Converting the PCC to Linked Data," *Cataloging & Classification Quarterly* 58, nos. 3–4 (2020): 403–17; Hyoungjoo Park and Margaret Kipp, "Library Linked Data Models:

Library Data in the Semantic Web," *Cataloging & Classification Quarterly* 57, no. 5 (2019): 261–77; Lihong Zhu, "The Future of Authority Control: Issues and Trends in the Linked Data Environment," *Journal of Library Metadata* 19, nos. 3–4 (2019): 215–38.

12. Marcia Lei Zeng and Phillipp Mayr, "Knowledge Organization Systems (KOS) in the Semantic Web: A Multi-Dimensional Review," *International Journal on Digital Libraries* 20, no. 3 (2019): 209–30.
13. National Information Standards Organization, *ISO 25964 –the International Standard for Thesauri and Interoperability with Other Vocabularies*, 1.4, (2011/2013), https://www.niso.org/schemas/iso25964.
14. Library of Congress, *Subject Schemes* (2021), https://id.loc.gov/vocabulary/subjectSchemes.html. See the list maintained by the LC. The list currently consists of 371 items.
15. Book Industry Study Group, "BISAC Subject Codes FAQ," accessed March 25, 2021, https://bisg.org/page/BISACFaQ.
16. The Editorial Committee of the Chinese Library Classification, "Overview of Chinese Classification Subject Thesaurus," accessed March 25, 2021, http://clc.nlc.cn/ztfzfbgk.jsp.
17. Tina Gross and Arlene G. Taylor, "What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results," *College & Research Libraries* 66, no. 3 (2005): 212–30; Tina Gross, Arlene G. Taylor, and Daniel N. Joudrey, "Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching," *Cataloging & Classification Quarterly* 53, no. 1 (2015): 1–39.
18. Karen Smith-Yoshimura et al., *Implications of MARC Tag Usage on Library Metadata Practices*, a report produced by OCLC Research in support of the RLG Partnership (2010), 13, https://www.oclc.org/content/dam/research/publications/library/2010/2010-06.pdf.
19. Park and Brenza, "Evaluation of Semi-Automatic Metadata Generation Tools," 22–42.
20. Annif, "How to use Annif," accessed March 26, 2021, http://annif.org/.
21. OCLC, "Inside WorldCat," accessed July 12, 2021, https://www.oclc.org/en/worldcat/inside-worldcat.html.
22. OCLC, "Inside WorldCat."
23. William E. Moen et al., "Learning from Artifacts: Metadata Utilization Analysis" in *Proceedings of the Joint Conference on Digital Libraries–Opening Information Horizons*, (Chapel Hill, NC, June 11–15, 2006), 270–71.
24. Amy P. Eklund et al., "Comparison of MARC Content Designation Utilization in OCLC WorldCat Records with National, Core, and Minimal Level Record Standards," *Journal of Library Metadata* 9, nos. 1–2 (2009): 36–64.
25. Matthew Mayernik, "The Distributions of MARC Fields in Bibliographic Records: A Power Law Analysis," *Library Resources & Technical Services*, 54, no. 1 (2009): 40–54.
26. Smith-Yoshimura et al., *Implications of MARC*.
27. William E. Moen and Penelope Benardino, "Assessing Metadata Utilization: An Analysis of MARC Content Designation Use," in *Proceedings of the International Conference on Dublin Core and Metadata Applications: Supporting Communities of Discourse and Practice—Metadata Research & Applications* (Seattle, WA, Sept. 28–Oct. 2, 2003), https://dcpapers.dublincore.org/pubs/article/view/745.
28. Arlene G. Taylor and Charles W. Simpson, "Accuracy of LC Copy: A Comparison Between Copy that Began as CIP and Other LC Cataloging," *Library Resources & Technical Services* 30, no. 4 (1986): 375–87.
29. Shelia S. Intner, "Much Ado About Nothing: OCLC and RLIN Cataloging Quality," *Library Journal* 114, no. 2 (1989): 38–40.
30. Larson, "Between Scylla and Charybdis."
31. Herbert H. Hoffman, "Evaluation of Three Record Types for Component Works in Analytic Online Catalogs," *Library Resources & Technical Services* 42, no. 4 (1998): 292–303; Herbert H. Hoffman, "Subject Access to Works in Online Catalogs," *Technicalities*, 21, no. 5 (2001): 9–11.
32. Oksana L. Zavalina, Shadi Shakeri, and Priya Kizhakkethil, "Editing of Library Metadata Records and its Effect on Subject Access: An Empirical Investigation," in *Proceedings of the International Federation of Library Associations World Library and Information Congress Satellite Conference: Subject Access: Unlimited Opportunities* (Columbus, OH, August 11–12, 2016), https://pdfs.semanticscholar.org/4989/d3c7ba96a173117264e8f3e1310d0f4dcf0f.pdf.
33. Vyacheslav I. Zavalin, Oksana L. Zavalina, and Rachel Safa, "Patterns of Subject Metadata Change in MARC 21 Bibliographic Records for Video Recordings," in *Proceedings of the Association for Information Science and Technology* 58, no. 1 (2021): 1–5.
34. Vyacheslav I. Zavalin, "Analysis of Genre and Subject Representation in the Children's and Young Adults' Cataloging Program (CYAC) Bibliographic Metadata for Fiction Books," in *Proceedings of the International Association of School Libraries Conference* (Denton, TX, July 12–16, 2021).
35. OCLC Training and Support, "ELvL: Encoding Level," in *Bibliographic Formats and Standards*, accessed March 25, 2021, https://www.oclc.org/bibformats/en/fixedfield/elvl.html.
36. OCLC Training and Support, "042 Authentication Code (NR)," in *Bibliographic Formats and Standards*, accessed March 25, 2021, https://www.oclc.org/bibformats/en/0xx/042.html.
37. OCLC Training and Support, "042 Authentication Code (NR)."
38. Library of Congress Network Development and MARC Standards Office, "MARC 21 Format for Bibliographic Format: Appendix G: Format Changes for Update No. 31

(December 2020)," accessed March 25, 2021, https://www.loc.gov/marc/bibliographic/bdapndxg.html.

39. Shieh and Reese, "The Importance of Identifiers in the New Web," 208–26.
40. Wilson, *Two Kinds of Power*, 69.
41. Eklund et al., "Comparison of MARC"; Mayernik, "The Distributions of MARC Fields in Bibliographic Records"; Moen and Bernadino, "Assessing Metadata Utilization"; Moen et al., "Learning from Artifacts"; Smith-Yoshimura et al., *Implications of MARC*.
42. Pat Riva, Patrick Le Boeuf, and Maja Žumer, *Library Reference Model: A Conceptual Model for Bibliographic Information* (2017): 9, https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf.
43. Library of Congress, "ID.LOC.GOV—Linked Data Service," accessed March 26, 2021, https://id.loc.gov/.