Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 99 (2016) 50 – 63

Procedia Computer Science

International Conference on Knowledge Management, ICKM 2016, 10-11 October 2016, Vienna, Austria

Developing an Empirically-based Framework of Metadata Change and Exploring Relation between Metadata Change and Metadata Quality in MARC Library Metadata

Oksana L. Zavalina[a]*, Vyacheslav Zavalin[a], Shadi Shakeri[a], Priya Kizhakkethil[a]

[a]*University of North Texas College of Information Department of Library and Information Sciences, 1155 Union Circle 311068, Denton 76203- 5017, USA*

Abstract

Managing metadata quality and metadata change contributes to more effective knowledge management in various organizations. This study sought to test and refine the metadata change framework initially developed for characterizing the change in digit al library metadata, in the study of traditional library metadata. As well, the study investigates how the recent shift to the new resource description standard Resource Description and Access (RDA) has impacted the change in library metadata records. Using a content analysis approach, we assessed various quantitative and qualitative aspects of metadata change over time in library metadata records. The study reveals that in the time since the transition to the new standard, library metadata records have received a high level of attention from metadata editors as expressed in overall metadata change frequency and a variety of 3 change categories and 23 subcategories observed. Results suggests the most frequently occurring categories and subcategories of change within the library metadata records, the metadata fields most widely applied in the records and the ones most often undergoing changes, as well as the institutions that most often make changes to the records. Based on the changes observed in our study, we identify relations between specific metadata change categories and subcategories and metadata quality criteria of accuracy, completeness, and consistency. The directions for further research into library metadata change and connections between metadata change categories and metadata quality criteria are suggested.

Peer-review under responsibility of the Organizing Committee of ICKM 2016

Keywords: metadatachange; metadataquality; metadata evaluation; quality control; MARC; RDA;

* Corresponding author. Tel.: +1-940-565-3776; fax: +1-940-369-7898.
E-mail address: Oksana.Zavalina@unt.edu

Peer-review under responsibility of the Organizing Committee of ICKM 2016
doi:10.1016/j.procs.2016.09.100

1. Introduction and literature review

Metadata management, including the efforts on ensuring metadata quality, is an important component of knowledge management. Metadata records, which represent information objects and facilitate access to them within various repositories tend to change over time in response to environmental changes, which include changes in national and international standards for record creation, expansion of controlled vocabularies, and more. [21] Library community has traditionally developed large metadata databases to provide access to information objects held by libraries, and reliance on library metadata creation (a.k.a. cataloging) standards for more than a century has been one of the important ways to ensure metadata quality. Resource Description and Access (RDA) officially replaced its immediate predecessor, the Anglo-American Cataloging Rules (AACR2), as the next-generation standard for library metadata creation in 2013. The new standard was developed as it was believed that the underlying functionality built into the conceptual models behind it – Functional Requirements for Bibliographic Records (FRBR), Functional Requirements for Authority Data (FRAD), and Functional Requirements for Subject Authority Records (FRSAD) – would enhance the quality of library metadata and make it more Semantic-Web-ready. However, three years after RDA implementation, little is known as to how the implementation has impacted the quality of library metadata records.

Since the library community recognizes that generating metadata records with lower quality can negatively impact access by decreasing the visibility of resources within that collection, metadata change to support quality of metadata records has been emphasized in the library cooperative cataloging processes for decades (e.g., [12,14]). Metadata plays a critical role in the retrieval of information objects in both digital libraries and repositories (primarily used for archival and preservation purposes) and the full-text systems (i.e., systems that match the terms used in users' queries to the underlying concepts of documents). The full-text system retrievals function with the help of the assigned or automatically generated keywords (i.e., the terms that represent the conceptual underlying of the documents). However, in the digital libraries, the retrieval of in formation objects or the representation of the objects (e.g., metadata records in library catalogs) still depends on the quality of the supplied metadata. Numerous user studies provide evidence that high-quality metadata in is integral to increasing users' satisfaction and information use (e.g.,[18]). Therefore, to ensure that the users successfully find, identify, select and obtain information objects that they need, high -quality metadata should be provided to describe adequately, information objects contained in both physical and digital libraries. In the digital library research community, a variety of metadata quality criteria have been suggested to guide metadata evaluation; with the following three most widely accepted criteria [15,16]:

- metadata accuracy, measured as the degree to which the data values in metadata record match characteristics of the described information object[19]
- metadata consistency: semantic consistency which refers to an the extent to which the same values or elements are used for representing similar concepts, and structural consistency which is evaluated as a degree to which the same structure is followed in representing information in certain metadata elements[5]
- metadata completeness defined as an the extent to which resources are described using all applicable metadata elements to their full capacity; some assessment criteria include the number of metadata elements per record, the level of application of mandatory and optional elements, presence or absence of blank metadata elements in the records[9].

Many of the metadata quality criteria identified in the digital library community are equally applicable to digital library metadata (created in Dublin Core, MODS, Visual Resources Association Core and other digital library metadata schemes) and traditional library metadata (catalog records created in MARC bibliographic format). The following indicators of quality in traditional library metadata records have been suggested by several authors: using terms from controlled vocabularies as opposed to free-text terms whenever possible [1,2], providing the substantial level of detail [6, 17], the absence of typographical errors [3,4, 8], etc. These indicators match those suggested for digital library metadata quality criteria of consistency, completeness, and accuracy respectively.

Most of the digital library metadata quality studies involve content analysis on statistically significant samples of metadata records [11]. Collection-level metadata records that describe entire collections of information objects as a whole as opposed to individual objects can be examined manually due to the reasonable numbers of metadata records to work with. For instance, as part of a study of collection-level metadata quality in the IMLS DCC aggregation of the United States Institute for Museum and Library Services (IMLS)-funded digital collections, a group of researchers conducted longitudinal analysis of the modifications that had been made by digital collection developers housed at various cultural heritage institutions to collection-level metadata records created by hosting institutions' staff[25]. With the rapid growth of digital libraries and repositories many of which now provide access to their items through millions of metadata records, the evaluation of this item-level metadata relies mostly on computational approaches. This results in the lack of more in-depth qualitative analysis of item-level metadata.

Metadata quality assessment research provides the grounds for corrective and preventive measures to ensure that users can fully benefit from the advantages provided by metadata services[26]. Metadata research suggests the importance of examining metadata change as an indicator of quality to help improve metadata quality. For example, a collection- level metadata study [25] identified the metadata fields, whose data values are updated the most and concluded that subject metadata (including topical, geographical, and temporal) is often the focus of collection-level metadata quality improvement efforts in digital aggregations. Tarver and colleagues[20] conducted one of the first studies with a specific focus on metadata change in a digital library. This high-level computational analysis of metadata change, complemented by in-depth content analysis study[24], resulted in the development of the empirically-grounded model of digital library metadata change [23] which suggests the three major types of metadata change (addition, deletion, and modification) and a variety of more specific change subcategories under each.

The exponential growth of library collections requires high-quality metadata to make materials collected by libraries fully accessible. However, the investigation into the library metadata quality has not yet kept up with the demand for it. In particular, no research studies to date have conducted an in-depth content analysis of metadata change in library catalogs and attempted to build a model of library metadata change. The exploratory study reported here attempts to address this research gap and is expected to make significant contributions in developing an understanding of metadata change and its relation to metadata quality in library catalogs.

2. Methods

This study employed an in-depth content analysis of Machine Readable Cataloging (MARC) library metadata records created according to the RDA metadata creation standard. The records were obtained from the largest global database of library metadata WorldCat (http://www.worldcat.org) which, according to its developer OCLC[13], at the time of writing this paper contains over 372 million of metadata records and is rapidly growing, with a new record added to the database every 10 seconds. The research question that guided this investigation is: What are the characteristics of RDA MARC library metadata change and how these can be related to metadata quality in the library catalogs? In particular:

- Which fields and subfields in library metadata records are changed the most often?
- How do they change? What is the relative frequency of occurrence of major categories of library metadata change such as Addition, Deletion, and Modification and their respective subcategories? What other categories and subcategories of library metadata change can be identified?
- What institutions are most actively involved in creating and/or changing metadata records?
- How does metadata change affect the overall quality of metadata records?

The RDA-based MARC library metadata records were selected as a target for this study. The records for English - language video recordings in DVD format created in English as a language of cataloging were collected at two points in time: January 2013 versions of metadata records were obtained from OCLC Research team in a single MARC XML file consisting of 932 records matching the selection criteria. Random sampling approach was then used to achieve a generalizable sample of a manageable size for in-depth comparative manual content analysis. The 2015 versions of these 369 records were obtained via search by record IDs from World Cat using OCLC Connexion tool (http://connexion.oclc.org) in June 2015.

The earlier and later versions of each metadata record were exported from OCLC Connexion and processed with the help of MARC Edit (http://marcedint.reesenet.net) software and manually compared side-by-side using text editors. A preliminary list of categories of change and their subsequent subcategories was developed for MARC metadata from the list suggested by a non-MARC digital library metadata change study[23]. The occurrence of each category and subcategory of change represents a change of the quality, which facilitated the evaluation of the records' quality.

Therefore, the list was used as a coding manual as well as a measure for determining the changes in the quality, in the study presented in this paper. More categories were added to this list based on researchers' knowledge of MARC standard and experiences in creation and evaluation of MARC metadata.

After completing the initial training and the coding of 15 record pairs not included in the primary sample, the data was coded independently by the authors of this paper based on the coding manual, with overall intercoder agreement of 94.75%. For each field, we checked whether or not it was included in metadata record, noted any change in the field, determined categories, and subcategories of change. For each record, the total number of fields and the proportion of fields with change was calculated, as well as the total number of change subcategories observed in the record. We also calculated the number of fields with multiple instances of the same metadata change subcategory and with multiple subcategories of change and documented the nature of a change within a change subcategory (e.g., fixing a typographical error, changing the form of the name to controlled-vocabulary one, etc.).

Institution codes representing institutions that created and edited each record (contained in subfields $a and $d of MARC field 040) were extracted from the 2015 dataset. Institution names and locations – including ZIP codes or postal codes whenever available – were determined based on institution codes through searching of MARC Organization Codes database (https://www.loc.gov/marc/organizations/org-search.php). For geocoding purposes, location information was used to determine latitude and longitude for each institution using the Batch Geocode function in the Find LatitudeAndLongitude.com. Based on geocoding information, the Tableau tool (http://www.tableau.com/) was used to generate maps to illustrate geographic distribution of record creating and editing institutions.

The results of the comparative content analysis of RDA-based MARC library metadata records are reported and discussed in the next section. They are followed by presentation and discussion of results of the analysis with regards to creating and modifying institutions, as well as to the major metadata quality criteria: completeness, accuracy, and consistency.

3. Findings and discussion

Figure 1 visually represents the library metadata change framework resulting from our study. It shows three broad categories and 23 subcategories of library metadata change for MARC format metadata. Table 1 includes the listing of categories and subcategories of metadata change revealed by our analysis, along with examples from specific MARC fields for each subcategory of change. Subcategories 20 -23 (Table 1) had not been included in our original coding manual and were discovered in the process of analysis. All four additional subcategories belonged to the Modify category; the codes identifying these subcategories are underlined in Figure 1. Each change subcategory code is spelled out and more details on each of the subcategories are presented in Table 1 that follows Figure 1.

Although we assumed a high proportion of RDA-based MARC records to include changes, the finding that all 369 records in the sample contained at least one instance of metadata change was unexpected. Table 1reports the total, mean, median, maximum, and standard deviation of the number of metadata fields per record with instances of change for each metadata change subcategory. It also presents that the Addition change category occurred the most often in our dataset (1926 occurrences), closely followed by the Modification (1421 occurrences); Deletion occurred substantially less often (389 occurrences). The Modification was the most varying category of change observed in metadata records in our study, with a half of all subcategories of metadata change (11 out of 23) belonging to this category; only six subcategories were observed for Addition and 5 for Deletion. In our dataset, we did not observe any instances of the sixth subcategory of Deletion that was suggested by our preliminary draft framework prior to data analysis: the DD subcategory (delete data value from fixed field subfield). This subcategory is made more visible in both Table 1 and Figure 1 through the use of alternative font color.

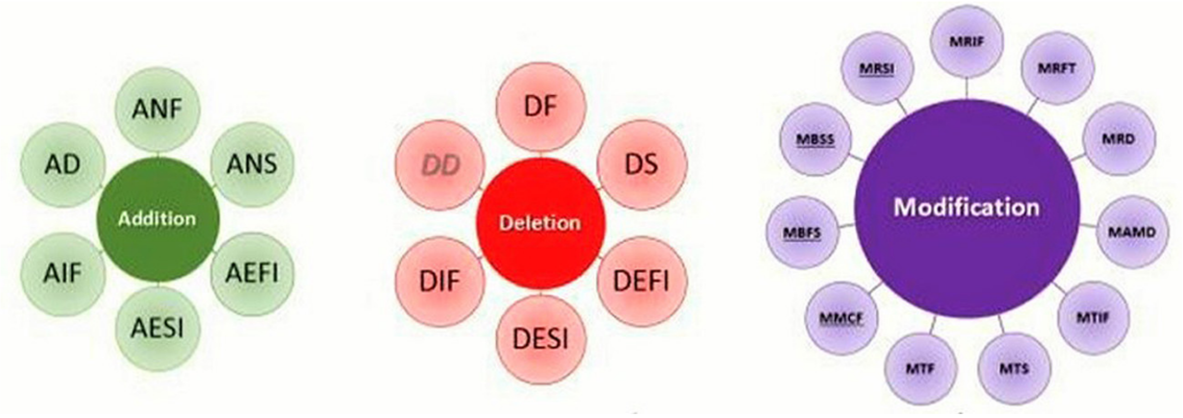

Fig. 1. Framework of library metadata change.

Table 1. Metadata change subcategories: definitions examples, and overall distribution.

Change category code	Change category name and example	Total occurrences	Fields per metadata record with change			
			Max	Mean	Median	Std. Dev
Addition:		1926				
1. ANF	Add new variable field not previously in the record (e.g., a 830 series title added access point)	236	10	0.64	0	1.37
2. ANS	Add new subfield to existing variable field (e.g., a subfield $b to 338 field)	1324	10	3.59	3	1.98
3. AEFI	Add an instance of existing variable field (e.g., a 2nd and 3rd instances of 655 field genre heading)	45	6	0.12	0	0.62
4. AESI	Add an instance of existing subfield (e.g., a 2nd instance of subfield $e to the same 700 personal name added access point field for a relator term to represent contributor's additional role)	307	4	0.83	1	0.54
5. AIF	Add an indicator to existing variable field (e.g., a first indicator 1 for field 041 language – 041 1_– to indicate the item is a translation)	4	1	0.01	0	0.10
6. AD	Add a data value to fixed field subfield (e.g., data value "g" in previously empty Audn subfield to indicate general audience)	1	1	0.003	0	0.05
Deletion:		389				
7. DF	Delete a field completely (e.g., the only instance of 250 edition statement field)	180	7	0.49	0	0.88

8. DS	Delete a subfield completely (e.g., all instances of subfield $4 relator code in 710 corporate name added entry field)	91	2	0.25	0	0.45
9. DEFI	Delete an instance of existing field (e.g., the 4th instance of 500 general note field)	115	4	0.31	0	0.67
10. DESI	Delete an instance of existing subfield (e.g., a 3rd instance of subfield $a in the same 043 geographic area code field)	2	1	0.01	0	0.07
11. DIF	Delete an indicator from existing field (e.g., 2nd indicator "0" in 505 field to represent the fact that contents note formatting is basic not enhanced)	1	1	0.003	0	0.05
12. DD	Delete a data value from fixed field subfield (e.g., a data value in the optional 2nd instance of Dates subfield designed for holding original date of publication when describing reprints)	0	0	0	0	0
Modification:		1421				
13. MRIF	Replace an indicator in existing field with new (e.g., change the 1st indicator in 490 series note field from "0" to "1" to indicate that the title of the series is established as a uniform title in the controlled vocabulary and that field 830 with series title added	15	2	0.4065	0	0.211042
14. MRFT	Replace field tag with new for the same data value (e.g., change field tag from pre-RDA 260 to RDA-based 264 while keeping the same data value)	40	1	0.1084	0	0.311308
15. MRD	Replace existing data value with new (e.g., replace "Motion picture" with "Television program" in 380 field or replace "a" with "i" in Desc subfield of a fixed field)	63	6	0.1707	0	0.566846
16. MAMD	Amend existing data value (e.g., change "two-dimensional moving image" to "three- dimensional moving image" in 336 field; change "160" to "165" in Time subfield of a fixed field)	186	7	0.504065	0	1.073702
17. MTIF	Transpose multiple instances of the same field (e.g., move the 700 personal name added access point field with the name of film director above the other 700 fields with names of the cast)	276	5	0.747967	1	0.871594
18. MTS	Transpose subfields within a field (e.g., move the 041 $h language code of original before 041	264	4	0.715447	1	0.497608
19. MTF	Transpose different fields within a record (e.g., move 505 contents note field from below to above 520 summary note field)	570	6	1.544715	1	1.47397
20. MMCF	Merge content with another field (e.g., merge the data value from 500 general note field with the data value in another note field such as 538 and delete	2	1	0.00542	0	0.073521
21. MBFS	Break an existing field into separate fields (e.g., break down the data value of a single instance of a 505 contents note field containing information on the content of three CDs into three instances of 505 field each containing information on one CD)	1	1	0.00271	0	0.052058
22. MBSS	Break an existing subfield into separate subfields (e.g., split the data value of 300 $c which incorrectly includes both dimensions and additional material information between 300 $c for dimensions and 300$e for additional materials)	1	1	0.00271	0	0.052058
23. MRSI	Replace subfield indicator with new (e.g., change 700 $4 to 700 $e and keep the same data value)	3	1	0.00813	0	0.052058

Our study revealed (Table 1) that addition of a new field or subfield (ANF and ANS) were the most observed subcategories of metadata change with a maximum of 10 occurrences within a single record (Max=7). Although Deletion change category was observed much less overall than Addition and Modification categories, one of its subcategories – deletion of a field (DF) – was observed quite often, up to 7 times per metadata record (Max=7). Similarly, we observed a maximum number of 7 fields per record with instances of MAMD change (modify: amend data value). Interestingly, the results suggest that on average each of the metadata change subcategories occurred less than once within each record, with the exception of ANS (M=3.5) and MTF (modify: transpose fields, M=1. 5).

The following eight subcategories of metadata change (in descending order) have been observed the most often – with at least 100 observations of each change subcategory in RDA-based MARC metadata records: Add new subfield (ANS), Add new field (ANF), Transpose fields within a record (MTF), Add new instance of an existing subfield (AESI), Transpose multiple instances of the same field (MTIF), Transpose subfields within a field (MTS), Amend data value (MAMD), Delete a field (DF), and Delete a field instance (DEFI). Various types of the Addition change category accounted for close to 50% of all metadata change occurrences. As presented in Table 1, ANF has the highest mean number of occurrences per record (M=3.6), followed by MTF (M=1.6), a subcategory of Modification which was found to be the second most widely observed category of metadata change. Also, the results revealed that Deletion was the least frequently occurring category of change in the RDA -based metadata records (i.e., the mean numbers of occurrences per record for all change subcategories were less than one (M<1)).

Library metadata records are created in MARC format (http://www.loc.gov/marc/bibliographic/). The data in a MARC metadata records are organized into fixed and variable fields; each variable field is identified by a three- character numeric tag (e.g., 040, 245, 710, etc.). MARC field tags are defined in blocks according to the first character of the tag, which, with some exceptions, identifies the general function of the field 's data within a record: 0XX Control information, numbers, and code fields; 1XX Main entry fields; 2XX Title and tit le-related fields, edition, imprint, etc. fields; 3XX Physical description, etc.; 4XX Series statements; 5XX Note fields; 6XX Subject access fields; 7XX Added entries other than subject or series; linking fields which represent relations between information objects and metadata records; 8XX Series added entry fields; holdings, location, alternate graphics, etc. fields. Each block includes several specific MARC fields. For example, 1XX block includes the following fields, mostly representing main creators of an information object: 100 Main entry – Personal name; 110 Main entry – Corporate name; 111 Main entry – Meeting name; 130 Main entry – Uniform title. Some of MARC fields can occur in the record only once (e.g., 111), others (e.g., 710 which represents additional corporate bodies responsible for the information object) are repeatable. Each MARC field has at least one subfield, and many of them have several subfields. For example, field 100 can have subfields $a,$b, $c, $d, $e, $q, and $4.

A total of 85 different fields from hundreds of possible MARC metadata fields were observed in one or more of the 369 records. Of these 85 fields, less than a half, only 39, occurred in at least 15% of records in the sample. Table 2 shows the ten most frequently changed MARC fields among the 39 commonly applied fields. Changes in field 040 Cataloging Source occurred in 100% of records as this field is automatically updated whenever the metadata record is edited. Change subcategories observed in this case were mostly ANS (Add a new subfield) or AESI (Add an instance of existing subfield) through the addition of subfield $d. The fields 336, 337 and 338, representing the content type, media type, and format type of an information object, are the most often cited fields that were introduced by RDA standard. Our analysis revealed that these fields – observed in almost all of the records in our study – were changed in more than 50% of the records, mostly through the addition of subfield $b (i.e., the content, media, and carrier type code). Five other fields among the top 10 are non-RDA-specific subject headings fields: personal name, corporate name, topical term, geographic name, and genre. In MARC field 650, ANS was the most observed Addition change subcategory through the inclusion of subfield $0 which, in the case of Faceted Application of Subject Terminology (FAST) subject headings, denotes the FAST authority record number. In MARC field 655 (60.19% of 324 records changed), ANS was also the most observed Addition subcategory, through the introduction of optional subfield $0 which denotes the authority record control number. The addition of subfield $0 points to the Linked Data efforts which are currently underway. The high proportion of records with change in subject-related fields observed in this study is in line with observations of previous studies in digital library environment (e.g., Zavalina et al., 2008) which found subject metadata to be the most frequently edited. Additionally, in the 043 field, which holds the machine-readable code for the geographic area, ANF (addition of a new field) was the most observed change subcategory.

Table 2. Top 10 most widely applied MARC metadata fields and the level of change in them.

Metadata field	Number of records containing field	% of records with change in field
040 Cataloguing source	369	100
650 Topical subject heading	339	98.53
336 Content type	368	51.9
337 Media type	368	53.269
338 Carrier type	367	54.7684
655 Genre subject heading	324	60.19
651 Geographical subject heading	193	95.85
043 Geographic area code	192	31.25
600 Personal name subject heading	70	67.14
610 Corporate name subject heading	56	58.96

Out of the 85 fields observed in one or more records in our dataset in 2013 and/or 2015 version, most have exhibited at least one metadata change event over time between the two data collection points, and only 20 fields, or less than 25%, did not change in any of the 369 records. Half of these unchanged fields belong to the group 0XX of MARC fields (control numbers etc.), three other belong to the 5XX notes group of fields, and three more belong to access points: 110 corporate name main entry, 711 event name added entry, and 810 series added entry corporate name.

To the contrary, some of the records in the analyzed dataset were found to contain more than one change category or subcategory in a given field (for example, a 650 topical subject heading field often contained the addition of new subfield and transposition of field instances at the same time). Figure 2 below shows the number of records for each of the 25 fields with multiple change categories. For most of these 25 fields, multiple change subcategories were observed in only 1 record or a small proportion of records, but field 040 cataloging source and most of the fields in the 6XX subject access fields (especially fields 650, 651, and 655) frequently contained more than one metadata change subcategory.

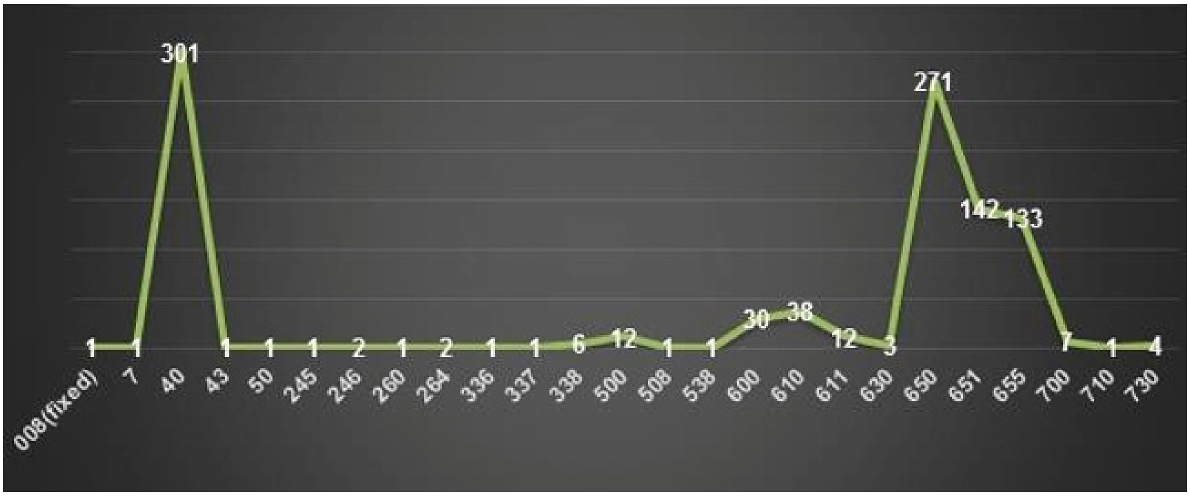

Fig. 2. Distribution of records with multiple metadata change subcategories in a field (n=369).

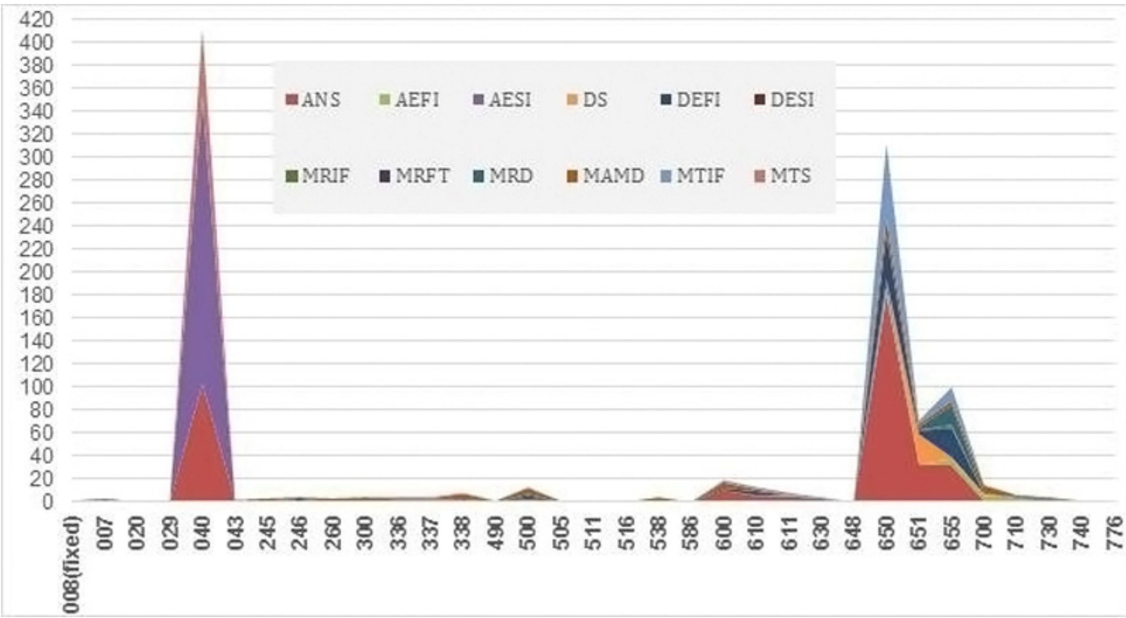

Fig. 3. Distribution of records with multiple instances of the same metadata change subcategory in a field (n=369).

Similarly, a substantial proportion of 85 MARC fields found in the records in our dataset – 32 variable MARC fields and a fixed field – were found to contain multiple instances of the same change subcategory in the same field. For example, more than one new subfield was often added (ANS) to the same field 336 in the record, or data value was amended (MAMD) in more than one instance of a 655 genre field. This was observed for 12 out of 23 metadata change subcategories. For example, as shown in Figure 3, most changes in 040 field comprised of three subcategories: addition of 2^{nd} or further instance of existing subfield $d, followed by addition of new subfield (these included subfield $d modifying institution or subfield $b language of cataloguing), and transposition of subfields within a field (e.g., moving subfield $b above subfield $e etc.).

The left-hand part of Figure 4 shows how the record creation was distributed among 32 institutions that created metadata records in the dataset analysed in this study. The three institutions most actively creating RDA-based metadata records for DVD video resources are US-based university libraries: Stanford, North Carolina State, and Duke. The right-hand part of Figure 4 shows how editing of metadata records was distributed among 34 institutions editing metadata records. It is interesting that the overlap between the group of institutions creating metadata records and the group of institutions editing records is quite moderate: only seven institutions (20.58% and 21.88% of each group) were found to belong to both groups for the records analysed in this study.

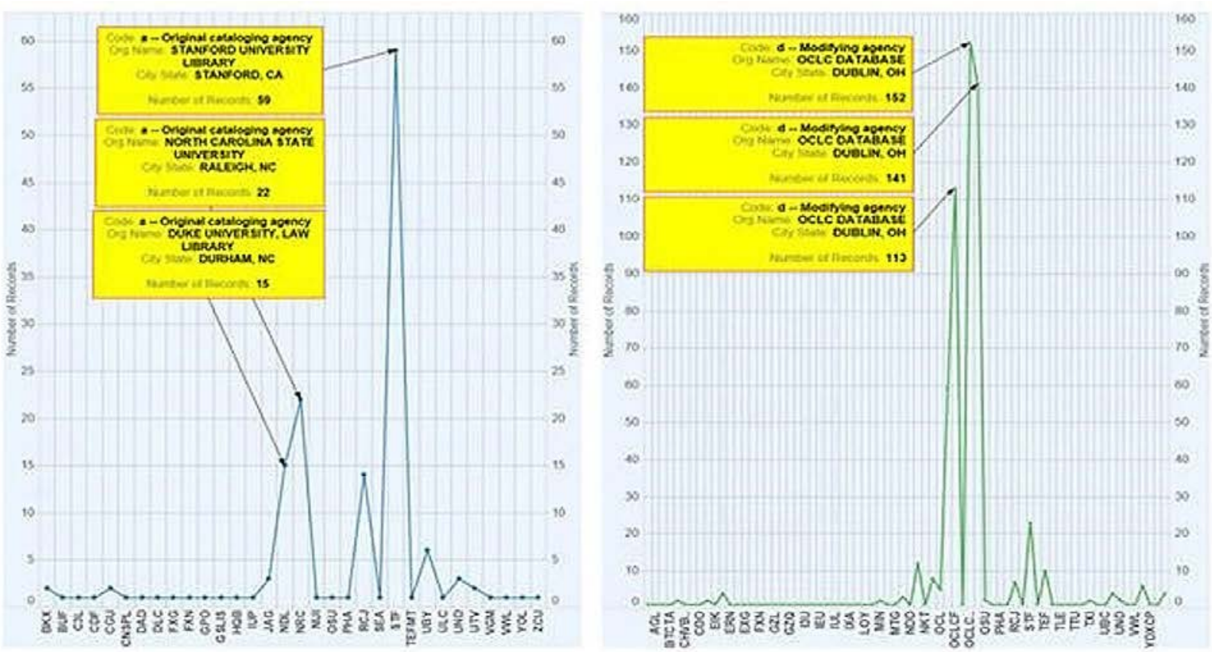

Fig. 4. Institutions creating metadata records and institutions editing metadata records: data values from field 040 subfields $a and $d.

While the codes for modifying institutions in the field 040 subfield $d represent a variety of different institutions responsible for editing the record – libraries, vendors, and others – most of the editing was found to originate from the OCLC team through various (often automated or semi-automated) processes. This is indicated by the high proportion of codes “OCL”, “OCLCA”, “ OCLCF”, “OCLCG”, “ OCLCO”, and “ OCLCQ” added to field 040 subfields $d: we found that only 23.85% of 612 instances of a subfield in our sample of 369 records had the data value other than any of the OCLC-affiliated ones listed above (Figure 4). This finding is in line with aggregated statistical data posted by OCLC for all (over 372 millions) of bibliographic records in World Cat that shows the top 6 most frequently used codes in 040 $d are the ones starting with “ OCL”, with the number of occurrences of the code ranging from 12 million to 211 million as of January 2016 (http://experimental.worldcat.org/marcusage/2016-01-040d.txt).

Institutions that actively contributed to editing the records in our dataset, also included a number of large academic libraries: Stanford University Library (STF), Duke University Law Library (NDL), North Carolina State University (NRC), University of Washington Library (WAU), Columbia University (ZCU), Brigham Young University Library (UBY), Emory University (EMU), Texas State University at San Marcos (TXI), University of California at San Diego (CUS), University of Chicago (CGU) etc. As shown in Figure 5 the institutions, which create and/or edit the RDA- based metadata records are located in various countries around the world.

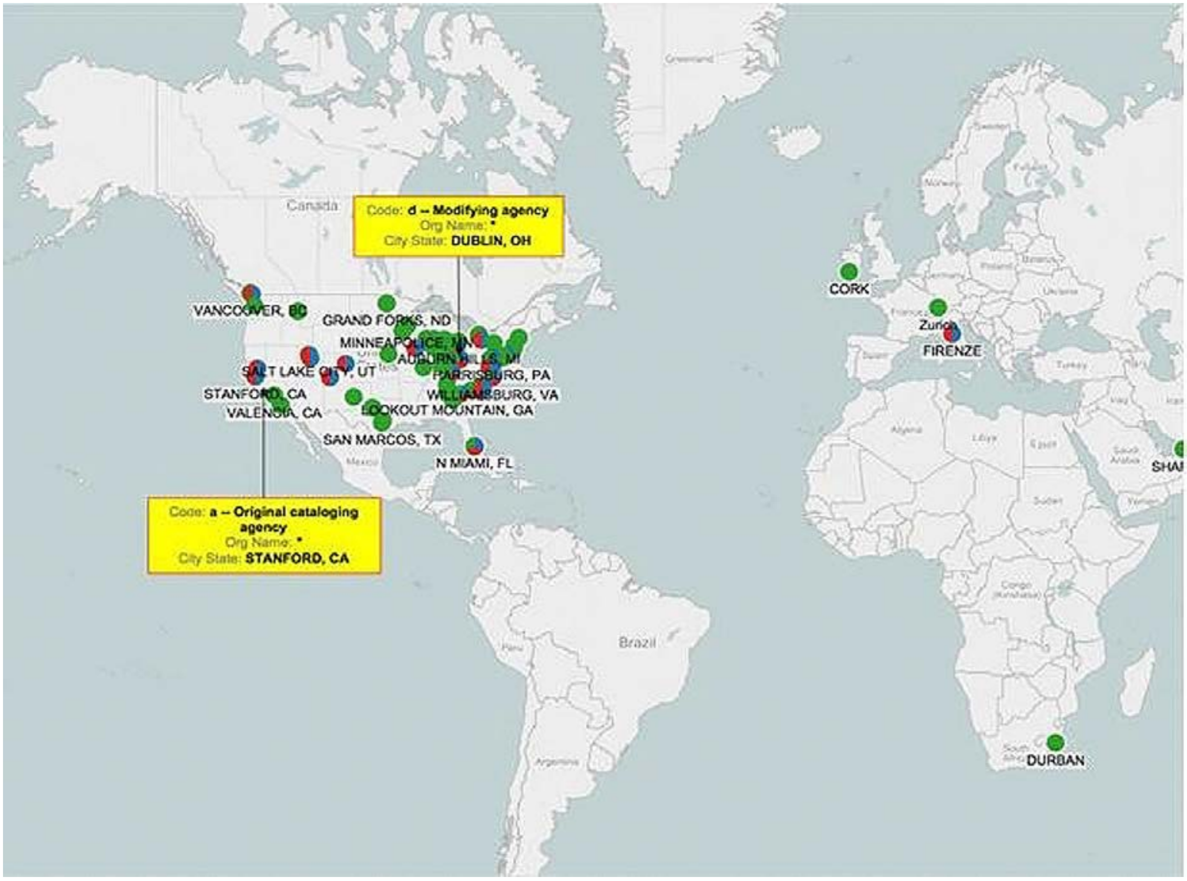

Fig. 5. Geographic distribution of institutions creating and editing metadata records: data values from field 040 subfields $a and $d.

Based on the observations obtained from our analysis of MARC metadata records, it can be argued that metadata quality has been enhanced, especially the completeness, accuracy and consistency of records. The high level of occurrence of ANF and ANS metadata change subcategories points to increased completeness. Many of the amendments of the data value that we observed in various fields either corrected typographical errors that might have negatively affected information retrieval (thus contributing to the accuracy of the record), or provided more detailed information about the object described by metadata record (thus supporting completeness of the record). The completeness and accuracy are also supported by modifying the 6XX subject access fields. For example, the addition of a code "OCLCF" indicates that a record was edited by OCLC FAST team which since September 2013 has been enhancing records by automatically adding Faceted Application of Subject Terminology (FAST) headings. This kind of metadata change improves information access through adding additional access points, thus influencing the completeness of the record. A code "OCLCA" is automatically added to field 040 subfield $d when one or more of included controlled-vocabulary headings representing a name of any kind, a topic, or a genre changes its form in the source controlled vocabulary (e.g., LCSH, etc.) and is replaced through automated processes in the record itself (http://www.oclc.org/content/dam/support/connexion/documentation/client/cataloging/catquickref/connexionclientquickref.pdf); this kind of metadata change supports consistency of metadata records. The codes "OCL", "OCLCQ", " OCLCG" in the 040 field $d indicates some other changes made to the record by OCLC Quality Control Group which for example includes merging two duplicate records, manually correcting mistakes in the record, or adding missing access points such as subject headings, etc., replacing natural-language form of the name with the authorized form of this name from a controlled vocabulary. These kinds of metadata change contribute to accuracy, completeness, and consistency of metadata records thus substantially improving their quality.

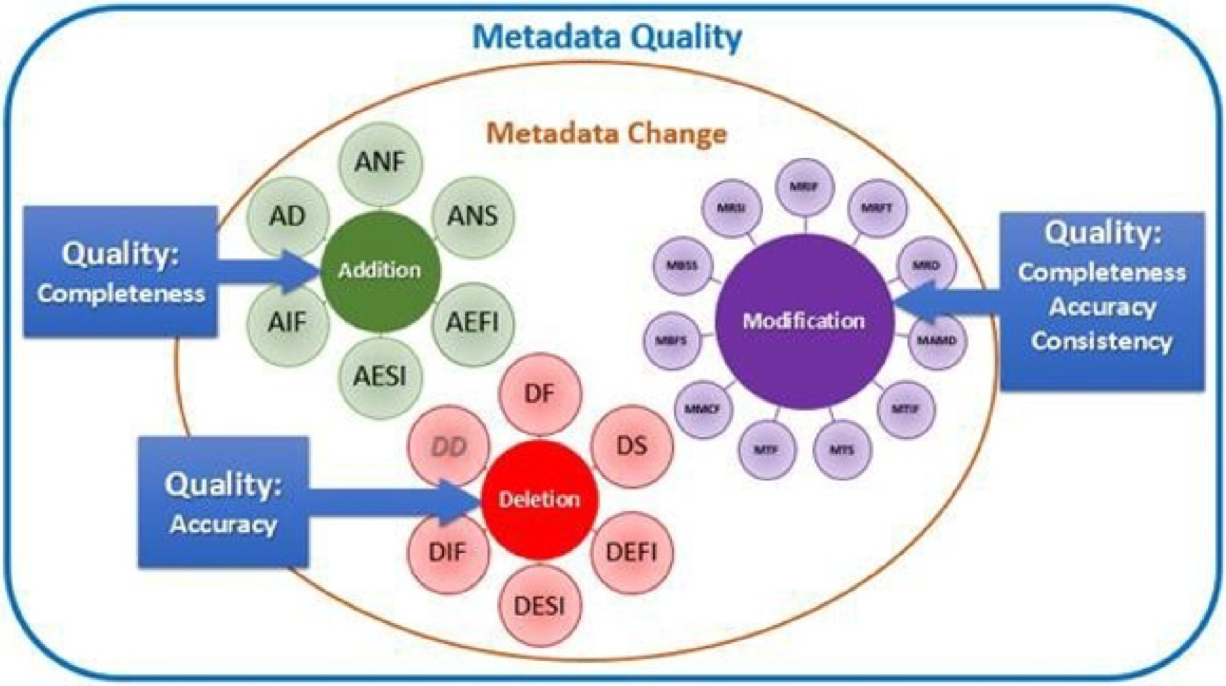

Fig. 6. Metadata change in relation to metadata quality.

Authors of this paper believe that metadata change is integral to metadata quality assurance. Overall, results obtained in this study suggest that the Modification category of metadata change contributes to all three major criteria of metadata quality: completeness, accuracy, and consistency, while Deletion mainly improves the accuracy of metadata records, and Addition mostly contributes to completeness (Figure 6). The majority (over half) of all instances of metadata change observed in our study belongs to Addition category which allows us to make a conclusion that completeness of metadata improved in the high proportion of the records in our sample. Our assumption, based on existing studies, is that metadata change in the Addition category would increase the number and/or quality of access points to information object for the user and therefore will contribute to the higher level of use. For example, a study conducted in a university library setting discovered that addition of the summary (MARC field 520) and table of contents (MARC field 505) fields in the electronic books (e Books) records increased their usage compared to those eBooks without these record enhancements [7]. (Such enhancements can be termed as a metadata change falling under the Addition category). Similar results were obtained in two other studies in which a direct correlation between enhancements in the catalog records through inclusion of table of content and the number of uses of a print book or eBook was noted [10,22]. In the future studies continuing and expanding the research started by our study, a similar approach can be used where changes in the usage of resources before and after metadata change can be evaluated. The study reported in this paper drew the data sample from an environment (OCLC WorldCat) where metadata is aggregated and therefore is not directly connected to information objects ' usage data and does not have a defined end-user community. For these reasons, measuring the effect of metadata change on metadata quality (in terms of user satisfaction and level of use) was not possible for this particular dataset.

4. Conclusions

The study results of which are presented in this paper developed an empirically -based model of metadata change, which categorizes metadata change in MARC library metadata records over time. This model has been tested and is found applicable to the analysis of hundreds of millions of existing MARC metadata records – both RDA-based and pre-RDA. The library metadata community needs a common vocabulary to discuss different kinds of metadata change. It is believed that the metadata change subcategories developed and tested as part of this study would provide such a vocabulary for metadata practitioners, researchers, and educators to compare record versions and discuss metadata change within and across systems.

The data obtained in this research also empirically supports the relation that exists between metadata change and metadata quality. Understanding and successful managing of metadata quality requires a cyclical process that balances the evolving needs of the users, the requirements of national/international standards, and the local environments of the metadata creators. With this in mind, the ability to accurately describe, measure, and communicate change events during the lifecycle of metadata associated with both digital and non-digital objects will be increasingly important and provides the research community with a useful benchmark to assess the quality of MARC or non -MARC metadata records.

This study used RDA-based MARC metadata records for English-language (and English-language-of-cataloging) DVD video resources. Future research will need to evaluate the metadata change for other types of materials in various languages and cataloged by institutions from both English-speaking and non-English-speaking countries. As more time passes after the relatively recent shift to new cataloging standard RDA, longitudinal studies of library metadata change will need to assess the temporal aspects of the effect of this major environmental change in standards guiding metadata records creation on quantitative and qualitative dimensions of metadata change. The model of metadata change proposed by this study will be further tested and refined in these future studies. Further research is needed into how the metadata change categories and subcategories contribute to various aspects of metadata quality: both the three major metadata quality criteria of completeness, accuracy, consistency, and other metadata quality indicators suggested by the literature. To be able to measure the effect of metadata change on metadata quality, these future studies need to involve surveys of user satisfaction and to rely on the metadata change and circulation/use data from local systems as opposed to aggregated environments. This study is expected to make a contribution to building understanding of metadata change and its relation to metadata quality in library catalogs.

References

1 Bade, D. (2002). The creation and persistence of misinformation in shared library catalogs: language and subject knowledge in a technological era: Vol. 211. GSLIS Occasional Papers. Champaign, IL: GSLIS, University of Illinois.

2 Bade, D. (2009). The perfect bibliographic record: platonic ideal, rhetorical strategy or nonsense? *Cataloging& Classification Quarterly, 46*(1), 109-133. doi:10.1080/01639370802183081

3 Beall, J. (2005). Metadata and data quality problems in the digital library. *Journal of Digital Information, 6* (3). Retrieved from https://journals.tdl.org/jodi/index.php/jodi/article/view/65

4 Beall, J. & Kafadar, K. (2004). The effectiveness of copy cataloging at eliminating typographical errors in shared bibliographic records. *Library Resources & Technical Services, 48*(2), 92-101.

5 Bruce, T.R., & Hillmann, D.I.(2004). The continuum of metadata quality: defining, expressing, exploiting. In Hillman, D. and Westbrook, L. (Eds.), Metadata in Practice. Chicago: American Library Association, pp. 238-256.

6 Hanson, H. & Schalow, J. (1999). Two aspects of quality in technical services: automating for quick availability, and identifying problems, effecting solutions. *Library Collections, Acquisitions, & Technical Services, 23(4),* 433-441.

7 Harker, K. & Sassen, C. (2014). Usage of E-books With Enriched Bibliographic Records. Electronic Resources and Libraries Conference Presentation. Retrieved from http://digital.library.unt.edu/ark:/67531/metadc277292/.

8 Mann, T. (1991). Cataloging quality, LC priorities, and models of the library's future. Washington D.C.: Library of Congress.

9 Moen, W.E., Stewart, E.L, & McClure, C.R. (1998). The Role of Content Analysis in Evaluating Metadata for the U.S. Government Information Locator Service(GILS): Results from an Exploratory Study. Retrieved from http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm

10 Morris, R. C. T. (2001). Online tables of contents for books: Effect on usage. Bulletin of the Medical Library Association 89(1): 29-36.

11 Ochoa, X., & Duval, E.(2009). Automatic evaluation of metadata quality in digital repositories. International Journal of Digital Libraries, 10, 67-91.

12 OCLC.(2015). Quality assurance. In Bibliographic Formats and Standards), chapter 5. Retrieved from https://oclc.org/bibformats/en/quality.html
13 OCLC. (2016a). A global Library Resource. Retrieved from https://www.oclc.org/worldcat/catalog.en.html
14 OCLC. (2016b). Quality Control. Retrieved from http://oclc.org/worldcat/catalog/quality-control.en.html
15 Park, J. (2009). Metadata quality in digital repositories: a survey of the current state of the art. *Cataloging& Classification Quarterly, 47* (3), 213-228.
16 Park, J. & Tosaka, Y. (2010). Metadata quality control in digital repositories and collections: criteria, semantics, and mechanisms. *Cataloging & Classification Quarterly, 48* (8), 96-715.
17 Shedenhelm, L. D., & Burk, B. A. (2001). Book vendor records in the OCLC database. *Library Resources & Technical Services, 45(1),* 10- 19.
18 Stvilia, B., Gasser, L., Twidale, M., Shreeves, S., & Cole, T. (2004). Metadata quality for federated collections. In S. Chengulur Smith, L. Raschid, J. Long, & C. Seko (Eds.), Proceedings of the International Conference on Information Quality ICIQ 2004 (pp. 111–125). Cambridge, MA: MIT.
19 Stvilia, B. et al. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology, 58* (12), 1720-1733.
20 Tarver, H., Zavalina, O.L., Phillips, M., Alemneh, D., & Shakeri, S. (2014). How descriptive metadata changes in the UNT Libraries' Collections: a case study, *Proceedings of the International Conference and Workshop on Dublin Core and Metadata Applications*, Austin, Texas, October 8-11, 2014.
21 Thornburg, G., & Oskins, M. (2007). Misinformation and bias in metadata processing: Matching in large databases. *Information Technology and Libraries, 26* (2), 15-26.
22 Tosaka, Y., & Weng, C. (2011). Reexamining content-enriched access: Its effect on usage and discovery. College & Research Libraries 72(5), 412-427. Retrieved from http://www.ala.org/nmrt/
23 Zavalina, O.L., Kizhakkethil, P., Alemneh, D., Phillips, M., & Tarver, H.S. (2015). Building a framework of metadata change to support knowledge management. *Journal of Information and Knowledge Management, 14* (1), 1-16.
24 Zavalina, O.L., & Kizhakkethil, P. (2015). Exploration of metadata change in a digital repository. Proceedings of the iConference 2015. Newport Beach, California, March 24-27, 2015.
25 Zavalina, O.L., Palmer, C.L., Jackson, A.S., & Han, M.-J. (2008). Evaluating descriptive richness in collection-level metadata. *Journal of Library Metadata, 8* (4), 263-292.
26 Yasser, C.M. (2011). An analysis of problems in metadata records. *Journal of Library Metadata. 11*(2), 51-62. doi:10.1080/19386389.2011.570654.