Quality Over Time: A Longitudinal Quantitative Analysis of Metadata Change in RDA-based MARC Bibliographic Records Representing Video Resources

Oksana L. Zavalina  
University of North Texas  
1155 Union Circle #311068,  
Denton, TX 76203  
Oksana.Zavalina@unt.edu

Vyacheslav Zavalin  
University of North Texas  
1155 Union Circle #311068,  
Denton, TX 76203  
Vyacheslav.Zavalin@unt.edu

Shawne D. Miksa  
University of North Texas  
1155 Union Circle #311068,  
Denton, TX 76203  
Shawne.Miksa@unt.edu

ABSTRACT
Everyday hundreds of millions of metadata records provide users access to information. To remain functional, metadata schemas changes in response to shifts in user needs and standards of information representation. This poster presents results of the longitudinal quantitative analysis of change occurring in a select group of library metadata records over time. This study traces the evolution of a sample of OCLC WorldCat metadata records created in MARC format according to new RDA standard of information representation. The same records were obtained at four data collection points at approximately one year intervals. Findings reveal a significant increase in the number of instances of a variety of metadata elements: both those introduced by the RDA standard and traditional pre-RDA elements. This study, along with related research, contributes to the understanding of metadata change and its relation to improvements in metadata quality and information access.

Keywords
Metadata evaluation, metadata change, library metadata, MARC, RDA; cataloging

INTRODUCTION AND BACKGROUND
The content of metadata records in large databases developed over time is influenced by various changes in the surrounding context. These environmental changes, as summarized by Thornburg and Oskins (2007), include changes in national and international standards for record creation, drifting or expansion of controlled vocabularies, changes in the goals or focus of institutions that contribute to the database (e.g., change from local repository to shared resource), outsourcing of record creation, change in relative proportions of different types of materials or topical coverage, and important changes in objects that are described by metadata records (e.g., electronic resources changing URLs, websites changing structure and contents, etc.). To keep pace with these environmental changes, metadata managers need to continually revisit and update metadata records, or in other words, make necessary changes to metadata records at regular intervals.

Researchers of information quality (Stvilia et al., 2004; Stvilia & Gasser, 2008) have pointed out the need to measure the metadata change and its outcomes for the users. However, almost no research identifying and measuring metadata change has been published in information science literature. Only a handful of studies have analyzed change in digital library metadata. This is in part due to the lack of information systems that include a metadata versioning functionality. Only two studies attempted to identify and measure change in metadata records in digital library environment, with a focus on item-level and collection-level metadata respectively (Tarver et al., 2014; Zavalina et al., 2008). Recently, an in-depth study in which a team of researchers qualitatively categorized metadata change in digital library metadata and compared it with traditional library metadata change categories was conducted (Zavalina et al., 2015; Zavalina & Kizhakkethil, 2015; Zavalina, Shakeri, & Kizhakkethil, 2015).

Existing studies on metadata change focus mainly on digital libraries which only came into existence in the last 15-20 years. However, for decades prior to introduction of digital libraries, metadata change was (and remains now) an important aspect of cooperative efforts in the library cataloging community. Catalogers have always edited bibliographic records obtained in bibliographic databases such as Online Computer Library Center’s (OCLC) WorldCat, Research Library Group’s (RLG) Research Libraries Information Network (RLIN), etc., to better meet local user needs and contribute to improving the quality of existing records to be shared among institutions (OCLC, 2015, 2016). Changes to library metadata records with the goal of improving their quality have been encouraged by the bibliographic utilities. For instance, OCLC has offered substantial financial credits to its members based on the quantity and level of improvements made to the records. Staff members of bibliographic utilities cataloging quality control units also routinely edit records both manually and automatically (e.g., by merging duplicate records for the same manifestation of a work etc.).
The topic of library metadata change is closely related to the topic of quality of bibliographic records in library catalogs. Quality in library cataloging has been defined as maintaining authority control (Bade, 2002 et al.), absence of typographical errors (Beall, 2005; Beall & Kafadar, 2004; Mann, 1991), the avoidance of duplicate records (Norgarde et al., 1993), and the level of information provided in the bibliographic data, often referred to as record enhancement (Hanson & Schalow, 1999; Shedenhelm & Burk, 2001). These criteria for library metadata quality greatly overlap with criteria of digital library metadata quality such as completeness, accuracy, and consistency (e.g., Bruce & Hillmann, 2004) which has been commonly accepted as three major criteria of quality by both digital and traditional library community (e.g., Park & Tosaka, 2009). Despite almost 50 years of metadata change activities in the library cataloging community and the assumed significant effect of metadata change on the quality of metadata, which in turn facilitates access to information for the users, the only published study to date that specifically analyzed change in library metadata records is the OCLC Research team’s 2013 user study into the effect of metadata record augmentations through automatic addition of terms from the Faceted Application of Subject Terms (FAST) controlled vocabulary (Mixer & Childress, 2013). However, the study focused on only a single set of fields in library metadata records – the subject metadata fields. The question still remains as to how a library metadata record as a whole changes over time and what this means for information access.

While user studies of the effect of metadata change in information access are important, we believe that the first step is to gain a clear picture of how much, and in what ways, metadata records change over time through empirical investigations. The objective of our study, some results of which are presented in this paper, is to bridge the gap in research into systematically identifying and measuring metadata change in library metadata records.

METHODS

The following research question guided our empirical investigation thus far: How does the application of various metadata elements in library metadata records change over time? To answer this, we relied on quantitative comparative content analysis of a sample of RDA-based MARC metadata records from the WorldCat database (http://www.worldcat.org). Based on metadata change factors suggested by Thornburg and Oskins (2007), our assumption for this study was that a major environmental shift such as change in international standards for metadata record creation – namely the recent shift from Anglo-American Cataloging Rules (AACR2) to Resource Description and Access (RDA) standard in 2013 – would naturally spur change in metadata records. Therefore, as a target group of metadata records we selected records based on new RDA standard. Additional selection criteria were introduced to generate a homogeneous sample of a manageable size: records describing one kind of information objects (English-language visual materials in DVD or Blu-Ray format), and records with English as a language of cataloging. Two versions of the same WorldCat records at two points in time – in January 2013 (3 months prior to official shift to RDA) and in April 2014 (approximately a year after the shift) – were obtained through OCLC Research team. In 2013, 932 records in the WorldCat database met the criteria specified above; over 20,000 records met the criteria in 2014. We selected for comparative analysis a random sample of 369 records that appeared in both the 2013 and 2014 versions of the dataset. Next, the 2015 and 2016 versions of the same 369 records were obtained directly from WorldCat database using the export tool in OCLC Connexion application in May 2015 and May 2016. Thus, we retrieved four datasets representing the evolution of the same metadata records over time, with approximately equal one-year-long intervals.

In our analysis, we evaluated and compared the four (4) datasets for the frequency of occurrence of each MARC 21 variable field that occurred at least once in our sample. Fixed fields that appear in every MARC record by default and that are non-repeatable – LDR, 001, 005, 008 – as well as some fields that are automatically generated and displayed every time the record is open (e.g., 049) were excluded from analysis as the number of occurrences can never change for these. In answering our research question, we were interested to see how the new elements of library metadata records introduced by RDA cataloging guidelines (e.g., MARC fields 264, 334, etc.), as well as traditional metadata record elements (e. g., 041, 246, 300, 490, 518, 650, 830, etc.), were applied and how the level of their application changed over time. MARC Edit, Microsoft Excel, and information visualization Tableau tool were used in our analysis.

FINDINGS AND DISCUSSION

Between 2013 and 2016, each of the metadata records in our sample underwent editing at least once, with an average of 4.97 editing events per record. Most of the 369 metadata records in our original 2013 dataset had been edited at least once prior to our first data collection point, as indicated by a total of 343 occurrences of MARC 21 field 040 subfield $d (Figure 1). Any time a change is made and the record is saved in the OCLC database, an instance of subfield $d with the code for institution that updated the record is automatically added to the record (i.e., this subfield is not present in the records that were not edited.) In the 2014 versions of the same 369 metadata records, the number of occurrences of this subfield increased almost three-fold, to 961, meaning that a total of 618 editing events happened in these 369 records over the period of 15 months between the two first data collection points (Figure 1). It is worth noting that this number does not represent the total number of changes as each editing event could include change to one or more fields of the metadata record. The trend continued between 2014 and 2015 data collection points as the total number of occurrences of 040 $d increased by 56.5% to 1,504 in 2015 dataset. Between 2015 and 2016 data collection points, the total number of instances of 040 $d raised by 22.6% to 1,844 which indicates gradual slowing down of the initially very fast pace of metadata change in RDA-based records over time.
Most of the hundreds of MARC 21 metadata elements are optional or required if applicable, while some (e.g., 245; 260 or 264 imprint information; 300 physical description for all kinds of information objects; 007 physical description codes for non-print materials) are mandatory. A total of 91 MARC variable fields were found in one or more versions of metadata records in our dataset; a number of these metadata elements exhibited change in the number of occurrences between 2013 and 2016. Figure 2 below shows percentages of records that included RDA-specific MARC 21 fields in 2013, 2014, 2015, and 2016 versions of the same 369 records in the sample. Figure 3 visually represents the same information for non-RDA-specific MARC 21 fields. The grouping of MARC fields into RDA-specific and non-RDA-specific is based on official documentation published by the United States Library of Congress Network Development and MARC Standards Office (e.g., https://www.loc.gov/marc/RDainMARC.html).

As shown is Figure 2, the levels of application increased for most of the RDA-specific fields, with the most substantial growth for fields 344, 347 and 340. Surprisingly, for some of the RDA-specific fields (e.g., 257, 338, and 380), the level of application somewhat decreased. The use of field 264 increased slightly (from 23.85% to 27.22% % of records) while the use of its traditional alternative – field 260 – decreased only slightly, from 77% to 73.6% of records (Figure 3) which might indicate the gradual switch in cataloging practice from using fields 260 to using its RDA counterpart 264.

As demonstrated by Figure 3, it was not only the RDA-specific MARC 21 metadata fields that increased in the level of application in the same set of records between data collection points. Several non-RDA-specific fields are more often applied in 2016 versions of the records then they were in 2013. The most drastic increase in the number of occurrences (from 35.5% to 54.2% of records) was observed for the 043 field which holds a machine-readable geographic area code that represents the place the information object is about and in field 050 (a gradual increase from 24.1% to 42.3% of records) which holds the LCC classification number. This increase might be due to semi-automatic record enhancement processes in which the data values for some of the fields are generated based on the data values in other fields. The increase in the proportion of records including field 019 (OCLC control number cross-reference) indicates that the duplicate records for some information objects represented by records were merged by the OCLC quality control team. Interestingly, field 648 which holds chronological subject headings generated in the process of creating FAST headings, exhibited a drop in the level of application between 2013 and 2014 data collection points but returned almost to its 2013 level by 2015 data collection point. The application of another 6XX subject field – 651 geographic heading – also substantially decreased between 2013 and 2015 data collection points (from 50.9% to 42.7% of records) but then somewhat increased by 2016.

We expected the level of change to be the highest for the 3XX group of MARC fields which are largely composed of new fields introduced by RDA cataloging code (336, 337, 338, 340, 344, 380 etc.). However, as shown by our field-level analysis, the increase happened only for some fields in this group. There was a decrease in the number of occurrences of some of the new metadata elements. This was unexpected and will require further investigation into subfield-level quantitative analysis and qualitative manual analysis of metadata records. Overall, though, the addition of a metadata element occurred much more often than a deletion of an element. This finding contradicts those of digital library metadata change studies which have reported a higher level of deletion. Such a difference between the patterns of change of metadata records in digital library and traditional library environments might be due to the nature of metadata standards used in these environments and the difference in the way information systems process MARC 21 data from, for example, how Dublin Core or other digital library metadata is processed.

![Figure 1. Metadata editing events in 369 records during 2013-2016](image1.png)

![Figure 2. Application of MARC21 RDA-specific metadata fields in 2013-2016 (% of records, n=369)](image2.png)
Figure 3. Application of MARC 21 non-RDA-specific metadata fields in 2013-2016 (% of records, n=369)
CONCLUSION
The study seeks to provide exploration and measurement of metadata change in the WorldCat library metadata that relies heavily on the RDA cataloging standard and MARC 21 metadata scheme. It is the first study of metadata change for this type of metadata that systematically analyzed each metadata element as opposed to only focusing on a specific group of elements.

Our findings indicate that RDA-based MARC records were extensively edited, with almost five editing events per record on average, over a relatively short period of time of three years. According to our data, the peak of metadata change occurred soon after the official transition to the RDA information representation standard: between the 2013 and 2014 data collection points. The pace of change in metadata records gradually slowed down over the next two years, however a substantial proportion of records underwent editing events between 2014 and 2016 data collection points. This may be due to catalogers gaining a firmer grasp of the new cataloging rules and standards as well as improvements in the bibliographic system overall. It also speaks to possible implications for improving cataloging practices, and cataloging education in general.

This analysis of library metadata change is part of a series of studies that will help to inform metadata management decisions such as setting priorities in metadata record editing. Although the results of this study are generalizable only to RDA-based English-language-of-cataloging MARC metadata records in WorldCat for English-language video materials in DVD and/or BluRay format, the sample used in this study is not representative of all kinds of materials. Also, this study presents only a high-level quantitative empirical investigation. A deeper study focusing on more granular units of analysis (MARC subfields as opposed to MARC fields) and a manual comparative content analysis of a sample of metadata records at several points in time, is being conducted by the authors of this paper. It looks beyond field occurrences and into specific types of metadata change, including various modifications of data values, indicators, and tags, and will complement the results of the study reported here.

REFERENCES