Exploration of Metadata Change in a Digital Repository

Oksana L. Zavalina, University of North Texas
Priya Kizhakkethil, University of North Texas

Abstract
This paper presents preliminary results of the ongoing research project that explores the changes occurring in metadata records over time. We use a large regional distributed digital library that versions its metadata records as a target of our case study. The preliminary findings in particular reveal what the most prevalent types of metadata change are, which metadata elements receive the most attention from those editing metadata records, and how the types of metadata change vary across metadata elements.
Keywords: metadata; change analysis; measurement; digital library evaluation
Citation: Zavalina, O.L., Kizhakkethil, P. (2015). Exploration of Metadata Change in a Digital Repository. In *iConference 2015 Proceedings*.

Contact: oksana.zavalina@unt.edu, priya.kizhakkethil@unt.edu

1 Introduction

The concept of change in texts, strings, files, scripts, etc., as well as mechanisms for identifying it – such as edit distance (e.g., Bille, 2005)[1] – have been explored in computer science field. This encompasses research related to file comparison tools used for isolating differences between files, programs applications, and ontologies, including different versions of the same entities: DIFF, COMM, PRETTY DIFF, and PROMPTDIFF (Cheney, 2010; Horwitz, 1990; Noy et al., 2004)[2].

In information science, several studies among the body of research in metadata quality (Stvilia et al., 2004; Stvilia & Gasser, 2008)[3] suggested the link between metadata change and quality of metadata and emphasized 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. We attribute this lack to unavailability – until recently – of open-source or inexpensive proprietary information systems that allow versioning that makes such analyses possible.

We have been able to identify only two studies that attempted to do this at a very broad level. One of them appeared six years ago. As part of the study of collection-level metadata quality, the IMLS DCC aggregation researchers conducted a small-scale analysis of the revisions that had been made by digital collection developers to metadata records created by the hosting institution staff (Zavalina, Palmer, Jackson, & Han, 2008) [4]. They categorized metadata revisions into two broad categories – change and addition – and compared the frequency of these two types of revisions for the elements in collection-level IMLS DCC metadata records.

A recent study (Tarver et. al., 2014)[5] analyzed metadata change in the large database where versioning is enabled, over a period of several years. The quantitative characteristics that researchers measured included overall distribution of the frequency of metadata change, the number of changes in metadata record length and access status, etc. The authors of that study found that almost 40% of metadata records in their database underwent one or more editing events over time, with substantial number of metadata revisions resulting in increased completeness of metadata records. The study revealed that while most metadata editing episodes made with the purpose to improve the quality of metadata records result in increase of metadata record length, surprisingly, some changes decrease

[1] Bille, P. (2005). A survey on tree edit distance and related problems. *Theoretical Computer Science*, 337(1-3), 217-239.

[2] Cheney, A. (2010). Pretty Diff - Documentation. Retrieved from http://prettydiff.com/documentation.php ; Horwitz, S. (1990). Identifying the semantic and textual differences between two versions of a program. *ACM SIGPLAN Notices, 25*(6), 234-245. DOI: http://dx.doi.org/10.1145/93548.93574 ; Noy, N., Kunnatur, S., Klein, M., & Musen, M. (2004). Tracking changes during ontology evolution. *Lecture Notes in Computer Science, 3298*, 259-273.

[3] Stvilia, B., Gasser, L., Twidale, M., Shreeves, S., & Cole, T. (2004). Metadata quality for federated collections. *Proceedings of ICIQ04, 111-12*; Stvilia, B., & Gasser, L. (2008). Value based metadata quality assessment. *Library & Information Science Research*, 30 (1), 67-74. Retrieved from http://dx.doi.org/10.1016/j.lisr.2007.06.006.

[4] 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.

[5] Tarver, H., Zavalina, O., 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.

record length. Authors concluded that further investigation was needed into reasons for this somewhat unexpected finding as well as into more granular dimensions of metadata change at the level of individual records, metadata elements, and data values.

The ongoing study the most interesting results of which are presented in this early work / preliminary results paper is beginning to bridge the gap in research into identifying and measuring metadata change in information science research.

2 Methods

The following research question guided the investigation: *What are the characteristics of metadata change?* In particular: How and when do the metadata records change? How does the amount of metadata change (e.g., as expressed in the number of elements with change in the record) correlate with such quantitative characteristics of the metadata record as its age, the number of editing events, and the fluctuations in the length of the record? Which fields in metadata records are changed the most often across multiple collections? What categories of change can be identified? What is the relative frequency of occurrence of metadata change categories?

To answer these research questions, the authors used a manual in-depth content analysis of item-level metadata records (specifically the metadata records that have undergone changes) in the large-scale digital repository University of North Texas (UNT) Digital Collections. At the time of analysis, this digital library contained almost 700,000 of metadata records in various metadata schemes, including Dublin Core and locally-developed UNTL metadata scheme based on Dublin Core. Almost a third of these records, over 200,000, had been edited at least once over a period of 4.5 years between the time of implementing a system that allows versioning of metadata records and the time of data collection in spring of 2014.To arrive to a manageable yet representative sample for in-depth manual exploratory metadata change analysis, the researchers applied the following criteria: the records selected for analysis had to be:

- Dublin Core records (this metadata scheme is widely used in digital repositories around the world; analyzing the change in Dublin Core records would allow for future comparative analysis of metadata change in multiple digital repositories)
- edited at least once
- initially created at different points over the extended period of time, starting with October 1, 2009 – soon after the versioning system was implemented – and ending substantially earlier than the time of data collection (the records initially created on or before December 31, 2012, or 15 months before data collection time, were selected)
- created by human metadata creators as opposed to automated processes
- last edited in January-April of 2014
- visible to the end users (i.e., having “unhidden” status both at the time of record creation and at the time of data collection)
- representing different collections in the aggregation, and describing different kinds/genres of information objects (images, conference and journal papers in various fields of knowledge, dissertation theses, and other textual materials).

This sampling approach resulted in 157 Dublin Core metadata records. Initial and latest versions of each record were subjected to comparative content analysis. Two researchers analyzed the records using the coding based on three broad categories of change: addition, deletion, and modification. In the process of analysis, additional, more granular metadata change categories emerged. The overall intercoder reliability constituted 91.46%. Analysis of this sample helped us find answers to most of the research questions.

The number of versions of metadata records in the sample ranged from 2 to 17. A total of 35 records (over 22% of records in the sample) had only 2 versions: initial and second (i.e., latest). The next largest subset of records (33 records or 21% of records in the sample) had 4 versions, followed by records with 5 versions (26 records or almost 17%). A subsample of eleven metadata records with 4 versions was selected for analysis of relative distribution of changes among the metadata editing events which result in new versions of the records. The purpose of this analysis was to answer the research question: at which stage do most of the metadata changes occur?

3 Findings

3.1 Metadata change frequency and variability

Our first observation in analyzing the dataset (Figure 1) was that most of the twenty one metadata elements did not exhibit any changes in a high proportion of metadata records in our dataset. The Language metadata element was the most stable, with not a single change observed. Additional seven metadata elements – Title, Date, Collection, Institution, Rights, Resource Type, and Format – did not contain any changes in 90% or more of the records. The Coverage, Contributor, Identifier, Creator, and Publisher, were the only five metadata elements which remained unchanged in less than a half of analyzed metadata records.

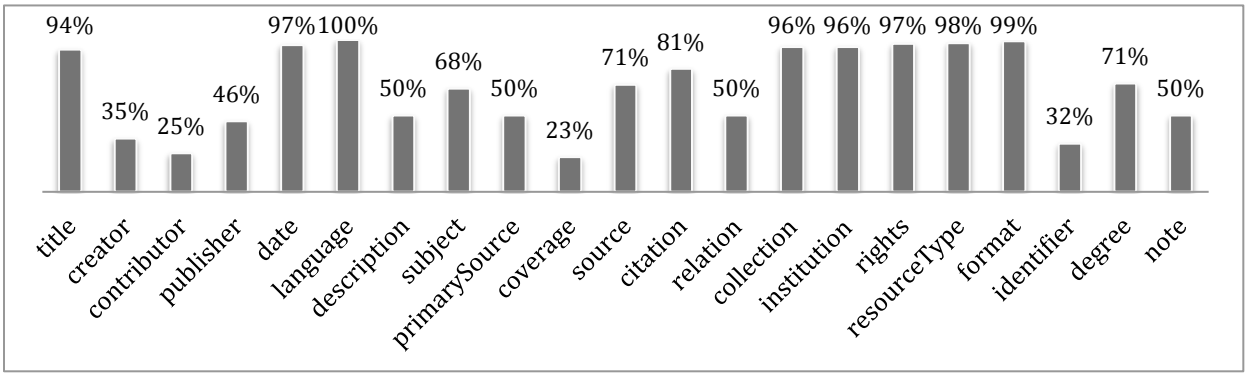

Figure 1. No change, % of records (initial and latest versions only; n=157)

To the contrary, a number of records in the sample exhibited multiple changes. For eleven metadata elements we observed more than one change type or subtype in the same record (Figure 2). For example, Creator and Note metadata elements contained multiple change types in more than a third of records; Contributor and Identifier elements contained multiple change types and subtypes in approximately a quarter of metadata records each. Sections 3.3 and 3.4 of this paper report our observations of various metadata change types and subtypes.

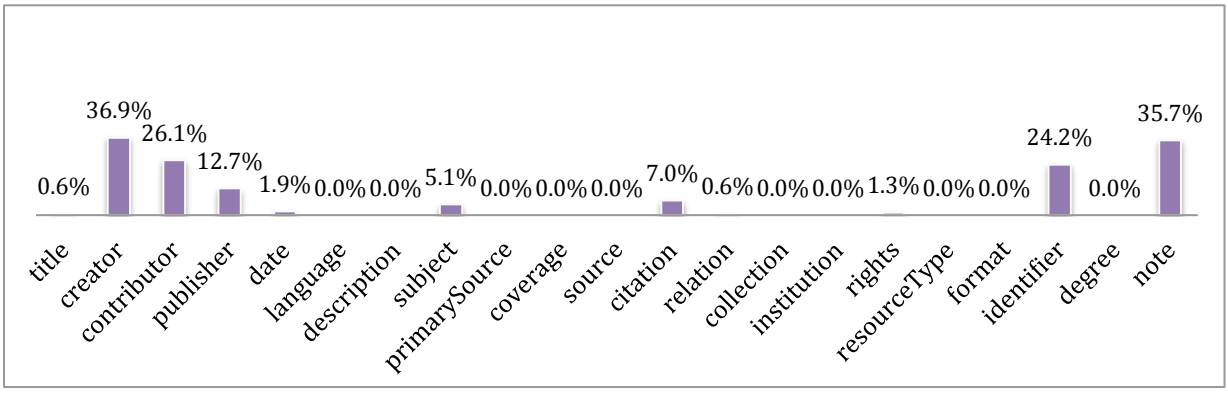

Figure 2. Multiple change types, % of records (initial and latest versions only; n=157)

We also observed that records in the sample exhibited the change in the number of instances of one or more metadata elements: in most cases through addition of second and further instances of a metadata element, but some deletions were also observed (more details on addition and deletion are presented in the sections 3.3 and 3.4 of this paper). As shown in Figure 3, the metadata elements that exhibited the change in the number of instances between the initial and latest record version in the highest proportion of metadata records were Identifier (almost 55% of records), Note (over 40%), Subject (close to 32%), and Citation (11.5%). On the other side of the spectrum, three metadata elements – Language, Source, and Primary Source – did not exhibit change in the number of element instances in any records in our sample. For the remaining 15 metadata elements, the percentage of records with the

change in the number of element instances varied between 0.6% and 7%.

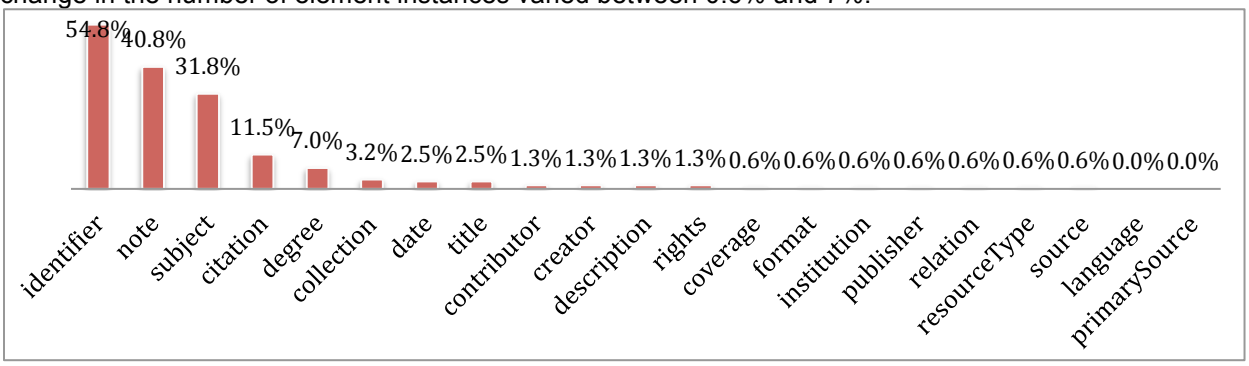

Figure 3. Records with changes in the number of instances of descriptive metadata elements (initial and latest versions only; n=157)

In the 157 records in the sample, we observed a total of 1263 instances of metadata change (Table1). Due to the nature of our sampling which focused on metadata records with at least one change, the minimum number of metadata fields with changes in our dataset was 1. Although in some cases, we observed as many as 18 metadata fields with changes, no more than one-third of twenty-one metadata fields in the record underwent changes on average.

Table 1. Metadata change distribution: variability measures (initial and latest versions of record; n=157)

	Total number of changes observed	Range		Mean number of fields with change per record	Median number of fields with change per record	Standard Deviation (number of fields with change per record)
		minimum number of fields with change per record	maximum number of fields with change per record			
any change in metadata record	1263	1	18	6.68	7	3.64

3.2 Metadata change: correlations

The changes in the analyzed 157 metadata records in most cases resulted in higher record length measured as the number of characters. The average increase in record length from initial to the latest edited version constituted 10.67% of the initial record length. However, a small proportion (3.2%) of the records in the sample decreased in size as a result of editing – one of them by as much as 33% of its initial record length in characters. This happened as a result of deletion of 7 previously empty fields in this record in the process of editing.

We measured the correlation between the age of the metadata record – expressed in the number of days, hours, minutes, and seconds between the records creation and the date of data collection (April 2014 for our study) – and the amount of change in the metadata record expressed as:

- the number of versions of this metadata record in the database
- the number of edited metadata fields in this record
- the difference in record length between the initial and the latest versions.

Our analysis revealed a moderate negative correlation for the first two indicators of metadata change: Pearson's r=-0.23944 was observed between the age of the record and the number of versions of this record; Pearson's r=-0.248 was observed between the age of the record and the number of metadata fields with changes in the record. A moderate positive correlation (Pearson's r=0.49472) between A strong positive correlation (Pearson's r=0.7741) was observed between the number of versions of a metadata record and the number of edited fields in the record. The correlation between the record length change and the number of edited fields in the record was also positive but weak (Pearson's r= 0.1167).

3.3 Major types of metadata change

Three major types of metadata change were identified: addition, deletion, and modification. Our findings (Table 2) indicate that deletion occurred the most often, with 534 instances, closely followed by modifications (475 instances); additions were observed substantially less often (254 instances) than two

other types of metadata change. The number of metadata fields with one or more deletions ranged from 0 to 9 per metadata record, with the highest mean and median: 3.38 and 4 respectively. The number of metadata fields with one or more modifications ranged the most, from 0 to 14 per record, with the mean of 3.01 and the median of 3. The number of metadata fields with additions exhibited the lowest variability, with the range from 0 to 8 per record, the mean of 1.62 and the median of 1 per record.

Table 2. Distribution of major metadata change types per metadata record (initial and latest versions only; n=157)

	Total number of changes observed	Range		Mean number of fields with change per record	Median number of fields with change per record	Standard Deviation (number of fields with change per record)
		minimum number of fields with change per record	maximum number of fields with change per record			
addition(s)	254	0	8	1.62	1	0.98
deletion(s)	534	0	9	3.38	4	1.76
modification(s)	475	0	14	3.01	3	2.17

Of the three major metadata change types, modifications occurred in the largest overall number of metadata fields – 16 out of a total of 21 (Figure 4), followed by deletions (11 fields), and additions (10 fields). Three metadata elements – Creator, Publisher, and Description – were the most modified elements, with modifications in these elements found in more than 40% of records. Four metadata elements – Coverage, Contributor, Primary Source, and Relation – underwent the most deletions (more than 40% of records each). Additions most often occurred in Identifier, Note, and Subject metadata elements.

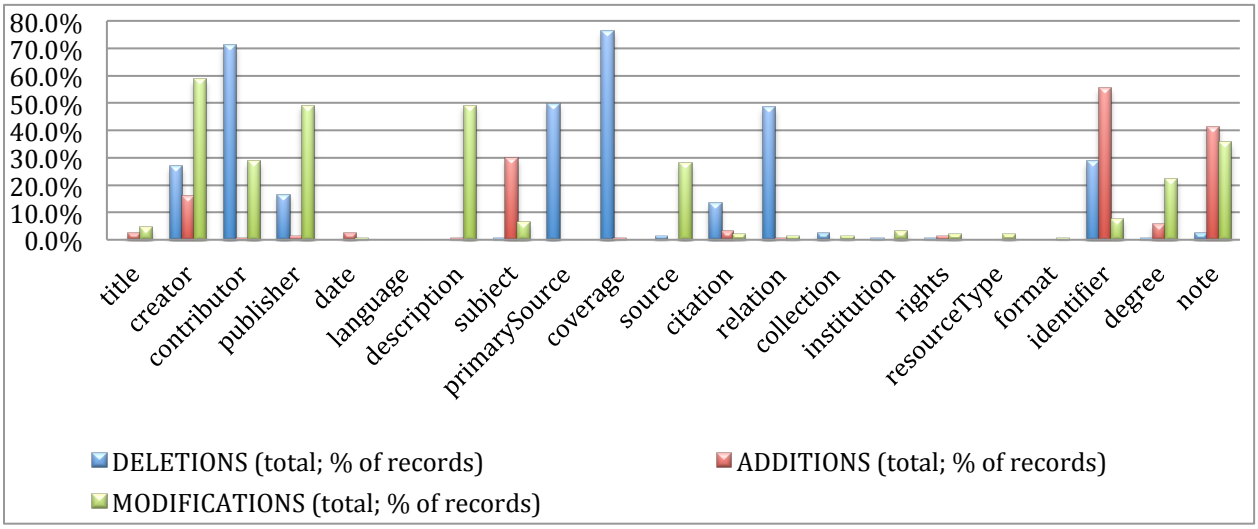

Figure 4. Relative frequency of distribution of major change types, % of records (initial and latest versions only; n=157)

3.4 Subtypes of major metadata change types

Table 3 shows the frequency of occurrence of each of the subtypes of addition, deletion, and modification for each of the elements in the metadata records. The order of metadata elements in this table is the same in which the elements appear in the metadata records in the target digital library.

Table 3. Metadata change subtypes observed, % of records (initial and latest versions only; n=157)

Metadata element	ADDITION			DELETION				MODIFICATION						
	field or subfield		qualifier of a field/ subfield	field or subfield			qualifier of a field/ subfield	populating empty field or subfield	replacement		amendment		transposition	
	new	2nd+ instance		empty	populated (the only instance)	populated (the 2nd+ instance)			data value	qualifier of a field/ subfield	data value	qualifier of a field/ subfield	data values	fields/ subfields
title	-	2.5	-	-	-	-	-	-	0.6	-	3.2	0.6	-	-
creator	15.3	1.3	-	26.8	-	-	-	-	-	-	31.2	-	0.6	14.0
contributor	0.6	-	-	69.4	1.9	-	-	-	-	-	22.3	-	-	-
publisher	0.6	-	0.6	15.3	1.3	-	-	-	4.5	-	20.4	-	-	24.2
date	-	2.5	-	-	-	-	-	-	0.6	-	-	-	-	-
language	-	-	-	-	-	-	-	-	-	-	-	-	-	-
description	0.6	-	-	-	-	-	-	-	-	-	49.0	-	-	-
subject	1.3	28.7	-	-	0.6	-	-	0.6	-	-	2.5	3.2	-	-
primary source	-	-	-	49.7	-	-	-	-	-	-	-	-	-	-
coverage	0.6	-	-	76.4	-	-	-	-	-	-	-	-	-	-
source	-	-	-	1.3	-	-	-	0.6	-	-	27.4	-	-	-
citation	1.3	1.9	-	5.7	0.6	7.0	-	0.6	0.6	0.6	0.6	-	-	-
relation	-	-	0.6	48.4	-	-	-	0.6	-	-	0.6	-	-	-
collection	-	-	-	-	-	2.5	-	0.6	0.6	-	-	-	-	-
institution	-	-	-	-	0.6	-	-	0.6	0.6	-	1.9	-	-	-
rights	-	-	1.3	0.6	-	-	-	1.3	0.6	-	-	-	-	-
resource type	-	-	-	-	-	-	-	0.6	-	-	1.3	-	-	-
format	-	-	-	-	-	-	-	0.6	-	-	-	-	-	-
identifier	25.5	21.7	8.3	12.1	-	-	16.6	7.6	-	-	-	-	-	-
degree	-	5.7	-	0.6	-	-	-	0.6	-	-	21.7	-	-	-
note	3.2	31.2	28.7	2.5	-	-	-	28.7	5.7	5.7	1.3	-	-	-

Based on analysis of initial and latest versions of metadata records, we identified the following subtypes of addition in metadata element or field of the record: addition of element qualifier and addition of a field or subfield, further subdivided into addition of a new field or subfield not included in the initial version of the metadata record and addition of a subsequent instance of a field or subfield that was present in the initial version. As shown in Table 3, addition of the new field and subsequent instances of existing fields was observed in almost identical proportion of metadata elements, and addition of a qualifier was observed less often. Overall, only two metadata elements out of twenty one – Identifier and Note – exhibited all three subtypes of addition. These same two metadata elements were the most affected by addition, with the highest proportion of records undergoing additions. Subject and Creator element were the third and fourth most affected by addition.

Our analysis revealed only one metadata element with deletion of a qualifier. Filed or subfield – either empty or populated (the latter further subdivided into deletion of the only instance of populated field and deletion of one or more of the multiple instances of the field not resulting in the absence of the field in the edited record) – was deleted quite often, in the total of 18 metadata elements. Empty field was deleted the most often, with Coverage, Contributor, and Primary Source metadata elements the most affected.

We identified four major subtypes of modification: populating previously empty field or subfield, amendment, replacement, and transposition. The first kind of modification occurred in twelve metadata elements out of twenty one, most frequently in the Note element. Complete replacement of a data value or field's qualifier with the new one was observed in eight elements, most frequently in Contributor and Note. Replacements of data values were observed much more often than replacements of qualifiers. Amendments – modifications that are less drastic than complete replacement – were observed substantially more often than two other subtypes of modifications discussed above. Amendments to data value were observed in a total of twelve metadata elements. The Description field demonstrated this kind of amendment the most frequently, followed by Creator, Source, Contributor, Degree, and Publisher. Finally, a transposition – a situation when the data values within the same field or the instances of a field or subfield are rearranged and the order in which they appear changes between the initial and the latest version of the metadata record – was observed in only two metadata elements: Publisher and Creator. The vast majority of this kind of metadata change occurred with data values.

Amendment of a data value, deletion of an empty field or subfield, and populating of an empty field or subfield were the three most widely occurring subcategories of change overall, with 13, 12 and 12 metadata elements affected by these kinds of metadata change respectively. Transposition of data values, amendment of a field/subfield qualifier, deletion of a populated second or further instance of the field or subfield, and deletion of a field/subfield qualifier were observed in the lowest number of metadata elements – 1 or 2 each. Qualifiers were rarely deleted or modified, but quite often added.

Overall, the Citation metadata element exhibited the widest variety of metadata change, with 9 subtypes of metadata change observed in this element. A total of 8 subtypes of metadata change were observed in the Note element, 7 in the Publisher element, and 6 in Subject and Identifier elements.

More details on the types and subtypes of metadata change, with numerous examples, can be found in Zavalina et al. (2015) [6] study which used the same sample of metadata records for testing the general framework of metadata change.

3.5 Metadata change in relation to metadata editing events

The in-depth comparative analysis of multiple versions of a small subsample of metadata records that underwent three editing events – resulting in a total of four versions of a metadata record – shows that most (over 70%) of the changes occur during the first editing event (i.e., between the initial and second version of the metadata record), and that the overall amount of change tends to reduce with each next editing event. As shown in Table 4, deletion is by far the most frequently occurring type of change in the first editing event. At the same time, this type of change was not found at all in the second and third editing events. While additions occurred in the second editing event almost half as often as in the first editing event, they were quite rare in the third editing event. To the contrary, the amount of modifications which was the second highest in the first editing event remained lower but steady between the second and third editing events.

[6] Zavalina, O.L., Kizhakkethil, P., Phillips, M., Alemneh, D., & Tarver, H. (forthcoming in 2015). Building a framework of metadata change to support knowledge management. *International Journal of Knowledge Management*.

Table 4. Distribution of metadata change across editing events
(records with 3 editing events; n=11)

	Total number of changes observed	Range		Mean number of fields with change per record	Median number of fields with change per record	Standard Deviation (number of fields with change per record)
		minimum number of fields with change per record	maximum number of fields with change per record			
1st editing event:	**68**	**3**	**8**	**6.09**	**6**	**1.30**
addition(s)	14	0	3	1.27	2	1.10
deletion(s)	32	0	5	2.91	3	1.51
modification(s)	22	0	3	2.00	2	1.10
2nd editing event:	**15**	**1**	**3**	**2.27**	**2**	**0.65**
addition(s)	6	0	2	0.55	0	0.69
deletion(s)	0	0	0	0.00	0	0.00
modification(s)	9	0	2	0.82	1	0.60
3rd editing event:	**11**	**1**	**3**	**2.00**	**2**	**0.89**
addition(s)	2	0	1	0.18	0	0.40
deletion(s)	0	0	0	0.00	0	0.00
modification(s)	9	0	2	0.82	1	0.75

4 Discussion and Conclusion

Five metadata elements – Coverage, Contributor, Creator, Identifier, and Publisher – were changed in 50% or more of the 157 metadata records analyzed in this study. Publisher and Identifier were also among the five elements with widest variety of categories of metadata change, along with Citation, Note and Subject metadata elements. Since Identifier field is commonly used for URLs and "iink rot" – changes in URLs that often result in limited accessibility of digital information object – is a known problem of metadata quality, the high frequency of change in Identifier fields of metadata records is expected. The changes to data value in Coverage could also be expected to some extent as this subject metadata element is more open to so called "cataloger judgment" of a metadata creator. However, the purely descriptive metadata elements such as Contributor, Creator and Publisher are substantially more straightforward – much like the Language element that did not exhibit any changes in our study – and therefore we did not expect these elements to be among the most frequently changed.

Among the major types of metadata change, modification occurred the most frequently overall. Most of the records in our main sample (n=157) and subsample of records with four versions (n=11) had modifications in one or more metadata elements. Creator, Publisher, and Description were the most modified metadata elements. Deletion occurred less often overall than modification and more often overall than addition; this type of change, however, was observed the most often in the first editing event for a subsample of records with four versions. Coverage, Contributor, Primary Source, and Relation metadata elements underwent the most deletions. Another major metadata change type, addition, occurred the least often. This particular does not confirm our intuitive expectation that addition would be found substantially more often than deletion. Additions most often occurred in Identifier, Note, and Subject metadata elements.

Our initial assumption was that amendment – a subcategory of modification – would be the most widely occurring kind of metadata change. This assumption held true as amendment of a data value, along with two other subcategories of metadata change – deletion of an empty field or subfield, and populating of an empty field or subfield – occurred the most often. On the other side of the spectrum, four subcategories of metadata change were observed the least often: transposition of data values, deletion of a populated second or further instance of the field or subfield, amendment of a field/subfield qualifier, and deletion of a field/subfield qualifier. This observation is only partially consistent with our expectation for much less changes in the qualifiers of metadata fields or subfields compared to the amount of changes to data values contained in these fields or subfields.

Only two of our assumptions regarding correlations between the indicators of metadata change were confirmed: a moderate positive correlation between the age of metadata record and the difference in record length between the initial and latest versions and a strong positive correlation between the number of versions of metadata records and the number of edited fields in the record. The surprising findings

include negative correlations between the age of the metadata record and indicators such as the number of versions of this metadata record in the database and the number of edited metadata fields in the record, as well as the very low positive correlation between the change in record length over time and the number of edited fields in the record.

The finding about the first editing event bringing about most of the changes to the metadata records is consistent with our expectations based on anecdotal evidence. However, the fact that one of the types of metadata change – modification – continues at a steady rate in subsequent metadata editing events, is surprising.

The study reported in this paper seeks to provide in-depth exploration and measurement of metadata change on a small purposive sample taken from a single large digital repository. Results of our analysis of metadata change will help to inform metadata management decisions such as setting priorities in metadata quality control and metadata record editing in the target digital repository. Although the sample is representative of digital collections and types of information objects in the UNT digital repository, the study's results are not – and are not intended to be – generalizable beyond this digital repository. Now that major digital content management tools such Fedora, Islandora, and Hydra provide metadata versioning capabilities, many digital libraries and repositories will build the data corpus for study of metadata change. This exploratory study helps identify some areas for future exploration that will be addressed by further, more in-depth, mixed-methods studies, which will need to examine characteristics of metadata change and explore its relation to metadata quality in multiple digital repositories.

5 Table of Figures and Tables