DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2015

Exploratory Analysis of Metadata Edit Events in the UNT Libraries' Digital Collections

Hannah Tarver
University of North Texas Libraries, USA
hannah.tarver@unt.edu

Mark Phillips
University of North Texas Libraries, USA
mark.phillips@unt.edu

Abstract

This paper presents the results of an exploratory analysis of edit events performed on records in the University of North Texas Libraries' Digital Collections during calendar year 2014. By comparing the amount of time that editors worked on records for certain item types and collections, we were able to isolate different categories of activities ("creating" vs. "editing") and to generalize rough benchmarks for expected editing durations depending on project criteria.

Keywords: metadata creation; metadata editors; edit events; benchmarks; editing activities

1. Introduction

One ongoing challenge for any metadata creation operation involves estimating the amount of time needed to create (or normalize) metadata for a particular project as well as the costs for doing the work. A reasonable estimate of time helps to build realistic timelines for internal or grant-funded projects, gauge the number of staff needed to meet deadlines, and assess the amount of funding required. To address this need, we decided to perform an exploratory analysis of data within the University of North Texas (UNT) Libraries' Digital Collections.

The Digital Collections comprise three large digital library interfaces: the UNT Digital Library (http://digital.library.unt.edu), The Portal to Texas History (http://texashistory.unt.edu), and The Gateway to Oklahoma History (http://gateway.okhistory.org). The UNT Digital Library primarily contains items owned, licensed, or created by UNT community members. The Portal is collaborative and contains materials owned by more than 250 partner institutions from across the state of Texas, while the Gateway hosts materials owned by the Oklahoma Historical Society. Materials from these collections are in a single, unified infrastructure and all items in our system use the same locally-qualified Dublin Core metadata with twenty-one possible fields (UNT, 2015). Records may be created in-house or by partner institutions, resulting in a large number of editors.

All of the digital library infrastructures for the Digital Collections, including public and administrative interfaces, were built in-house from open source software. Administratively, all item records are accessed via a single metadata editing environment (see Appendix A) locally referred to as the "Edit System." The Edit System loads the current version of a metadata record (which can range from a blank template to a complete record) into a user interface that allows users (i.e., metadata editors) the ability to complete or modify the record and then publish it. At this point, the Edit System saves the most current version and re-indexes the record. Each time an editor interacts with a metadata record, the Edit Event system (see Appendix B) logs the duration and basic metadata information. The analysis presented in this paper is based on events logged by the Edit Event system.

2. Methods

The research questions that guided this exploratory study are: Can metadata event data be used to establish and verify benchmarks within a metadata environment by looking at general

https://doi.org/10.23106/dcmi.952137010

information such as editor or record identity and length of edits? Can metadata edit event data be used to understand the activities of specific users within collections in a metadata system?

Our metadata system creates a log entry when a user opens a record to begin editing, starting a timer for the specific edit session of that record. When the user publishes the record, the Edit Event system queries the log entry and records the duration of the edit in seconds with the editor's username, record identifier, status (hidden or unhidden), record quality -- a completeness metric based on values for eight required fields (title, language, description, subject, collection, partner institution, resource type, format) -- and changes in status or quality (see Table 1). Unless otherwise noted, all duration counts in this analysis are represented in seconds.

TABLE 1: Sample metadata Edit Event system entry.

ID	Event Date	Duration	Username	Record ID	Record Status	Record Status Change	Record Quality	Record Quality Change
73515	2014-01-04T22:57:00	24	mphillips	ark:/67531/metadc265646	1	0	1	0

With this information we can easily see the number of metadata edits on a given day, within the month, and for the entire period we've been collecting data. We can also view the total number of edits, the number of unique records edited, and finally the number of hours that our users have spent editing records within a given period.

We decided to limit this analysis to the calendar year lasting from January 1, 2014 to December 31, 2014, to have a concrete period of time with a reasonable number of data points. The logs contained a total of 94,222 metadata edit events for that year, across 68,758 unique records. These events represent a full range of edit types for materials in our collections. In some cases records were created from blank or near-blank templates by staff members or partner institutions; in other cases, edits were made to correct errors, fix formatting, or add new information to completed records.

In addition to the metadata edit events, we extracted information from the UNT Libraries' Digital Collections related to the individual records: the contributing partner institution, collection code, resource type, and format data for each edited record. We also manually coded the 193 unique metadata editors in the system to classify each as a UNT-Employee or Non-UNT-Employee, and to assign a "rank" of librarian, staff, student, or unknown.

The information was merged and loaded into a Solr index, used as the base datastore for this analysis. We made use of built-in functionality of the Solr index (e.g., StatsComponent, Simple Faceting, and Pivot Faceting) and wrote Python scripts to interact with the data from Solr as needed.

3. Findings

To address the research questions, we first performed basic analysis on the dataset for some of the primary factors including: who is editing the records, what they are editing, and length of edits.

3.1. Who

A total of 193 unique metadata editors logged 94,222 edit events during 2014. As Figure 1 shows, the ten most prolific editors (5% of population) made 57% of overall metadata edits; the graph quickly tapers down to the "long tail" of users who have a lower number of edit events. Since we are reporting on the activities within our own system, it is not surprising that the authors are both listed in the top 5%, as well as others employed in the department.

https://doi.org/10.23106/dcmi.952137010

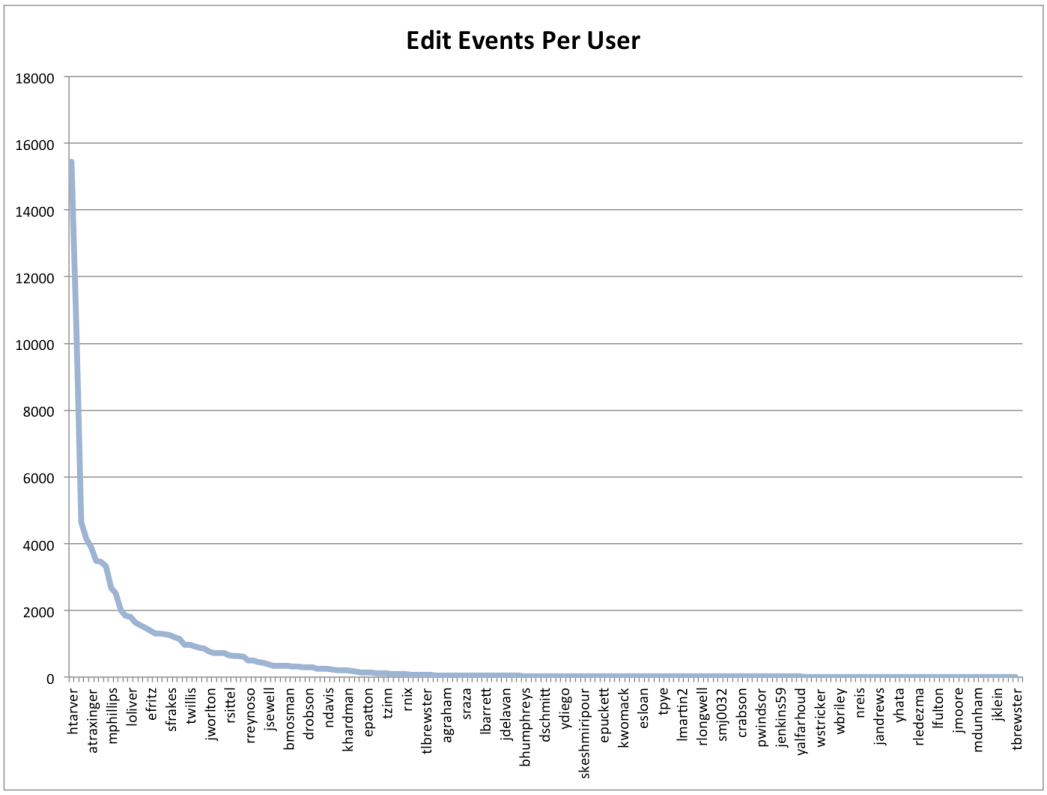

FIG. 1. Distribution of edit events, per editor.

Of the 193 editors in the dataset, 135 (70%) were classified as Non-UNT-Employee and 58 (30%) were classified as UNT-Employee. For the edit events, 75,968 (81%) were completed by a user classified as a UNT employee and 18,254 (19%) by a non-employee user. We also broke this down based on assigned rank of librarian, staff, student, or unknown (see Table 2).

TABLE 2: Statistics for the editors in the system based on their rank.

Rank	Edit Events	Percentage of Total Edits (n=94,222)	Unique Users	Percentage of Total Users (n=193)
Librarian	22,466	24%	16	8%
Staff	12,837	14%	13	7%
Student	41,800	44%	92	48%
Unknown	17,119	18%	72	37%

A clear majority (44%) of all of the edits were completed by students, while librarians and staff members combined accounted for 38% of the edits. The number of students includes both UNT employees -- students employed to do metadata work -- and 65 non-employee students who edited records as part of an assignment in a UNT metadata course.

3.2. What

The dataset contained 94,222 edit events occurring across 68,758 unique records, for an average of 1.37 edits per record. The maximum number of edits for a single record was 45, though most of the records -- 53,213 records (77%) -- were edited just once. Roughly 14% (9,937 records) were edited two times; 5% (3,519 records) were edited three times; records with four or more edits per record only account for 4% of the total dataset.

https://doi.org/10.23106/dcmi.952137010

To see the distribution of edits, we categorized records by the partner institution listed in each record and analyzed statistics for the ten most represented partners in the dataset (see Table 3).

TABLE 3: Most edited items by partner institution.

Partner Code	Partner Name	Edit Count	Unique Records Edited	Unique Collections
UNTGD	UNT Libraries Government Documents Department	21,932	14,096	27
OKHS	Oklahoma Historical Society	10,377	8,801	34
UNTA	UNT Libraries Special Collections	9,481	6,027	25
UNT	UNT Libraries	7,102	5,274	27
PCJB	Private Collection of Jim Bell	5,504	5,322	1
HMRC	Houston Metropolitan Research Center at Houston Public Library	5,396	2,125	5
HPUL	Howard Payne University Library	4,531	4,518	4
UNTCVA	UNT College of Visual Arts and Design	4,296	3,464	5
HSUL	Hardin-Simmons University Library	2,765	2,593	6
HIGPL	Higgins Public Library	1,935	1,130	3

Many partners who are heavily represented have edits spread across multiple collections. However, there are also differing trends regarding the ratio of edits to records. Figure 2 quickly shows which partners often make multiple edits per record as opposed to those partners that tend to have only one record edit event per record. In some cases, such as the editing done by Houston Public Library, the number of edits is roughly double the number of records, versus editing relationships that are nearly one-to-one (e.g., Hardin-Simmons University Library).

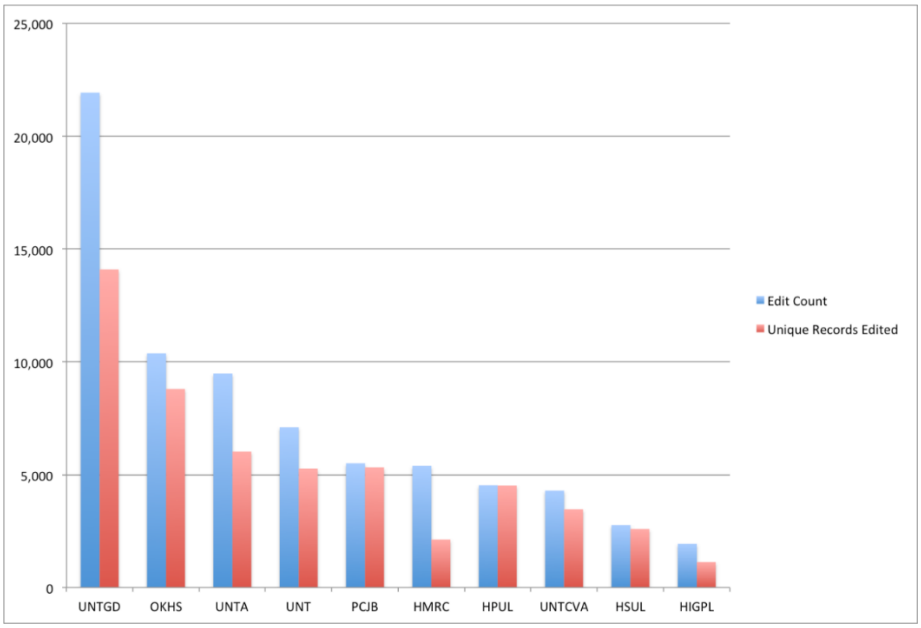

FIG. 2. Comparison between edit count and unique records for the top ten partners.

Since edits may be distributed across multiple collections (see Table 3), we analyzed edit events by collections and determined the ten collections that had the most edited items (see Table 4).

https://doi.org/10.23106/dcmi.952137010

TABLE 4: Most edited items by collection.

Collection Code	Collection Name	Edit Events	Unique Records Edited
TLRA	Texas Laws and Resolutions Archive	8,629	5,187
ABCM	Abilene Library Consortium	8,481	8,060
TDNP	Texas Digital Newspaper Program	7,618	6,305
TXPT	Texas Patents	7,394	4,636
OKPCP	Oklahoma Publishing Company Photography Collection	5,799	4,729
JBPC	Jim Bell Texas Architecture Photograph Collection	5,504	5,322
TCO	Texas Cultures Online	5,490	2,208
JJHP	John J. Herrera Papers	5,194	1,996
UNTETD	UNT Theses and Dissertations	4,981	3,704
UNTPC	University Photography Collection	4,509	3,232

The distribution of edit events and unique records also varies by collection. Since items can have assignments to multiple collections, some of the data in Table 4 overlaps. For example, the John J. Herrera Papers were part of the Texas Cultures Online project, which explains why the editing trends look similar (see Fig. 3). However, there were other edits to the Texas Cultures Online project which were not part of the Herrera papers, so the numbers are not an exact match.

Figure 3 shows the relation of edit events and unique records by collection. There are some slightly different trends, but this information is helpful in our system because collections often encompass discrete projects, while edits to partner items may be spread across multiple projects.

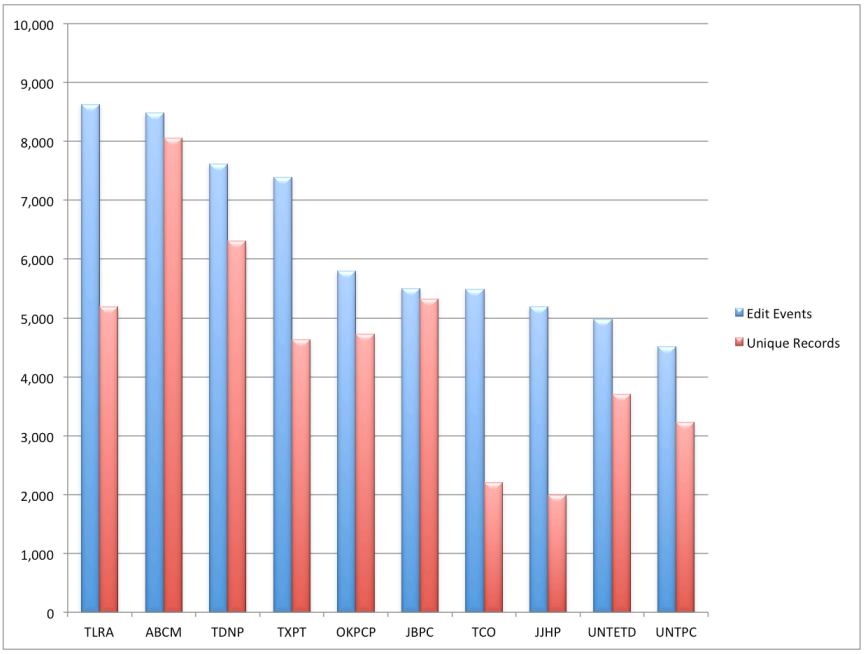

FIG. 3. Comparison between edit count and unique records for the top ten collections.

3.3. How Long

Without a time aspect, it would be difficult to formulate benchmarks or generalize conclusions from the raw data. The duration of edits in this dataset ranged from only 2 seconds to over 119 hours. To better visualize the distribution, the duration of each edit event was grouped into "buckets" of hours and minutes. A majority of edit events -- 93,378 (99%) -- lasted for 60

https://doi.org/10.23106/dcmi.952137010

minutes or less. Of these events that happened within an hour, 75,981 (81%) of the events lasted less than 6 minutes and 17,397 (19%) lasted 7-60 minutes (see Fig. 4).

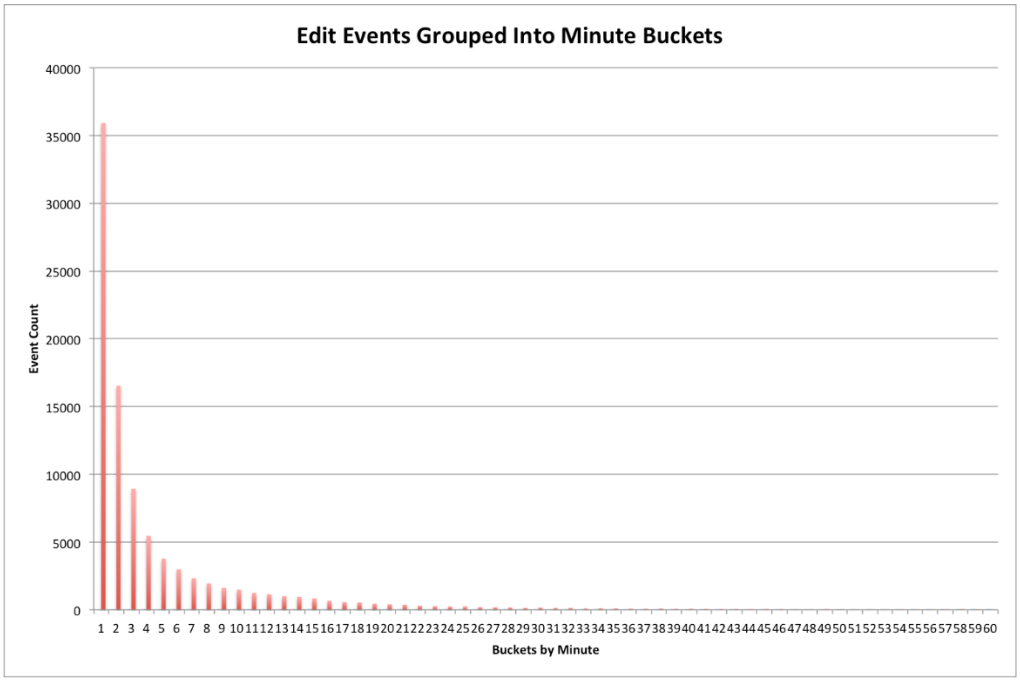

FIG. 4. Distribution of edits up to 60 minutes in duration (n-93,378)

Since a relatively large number of events (35,935) lasted less than one minute, we graphed this subset to see where those edit events fell within the distribution by number of seconds (see Fig. 5).

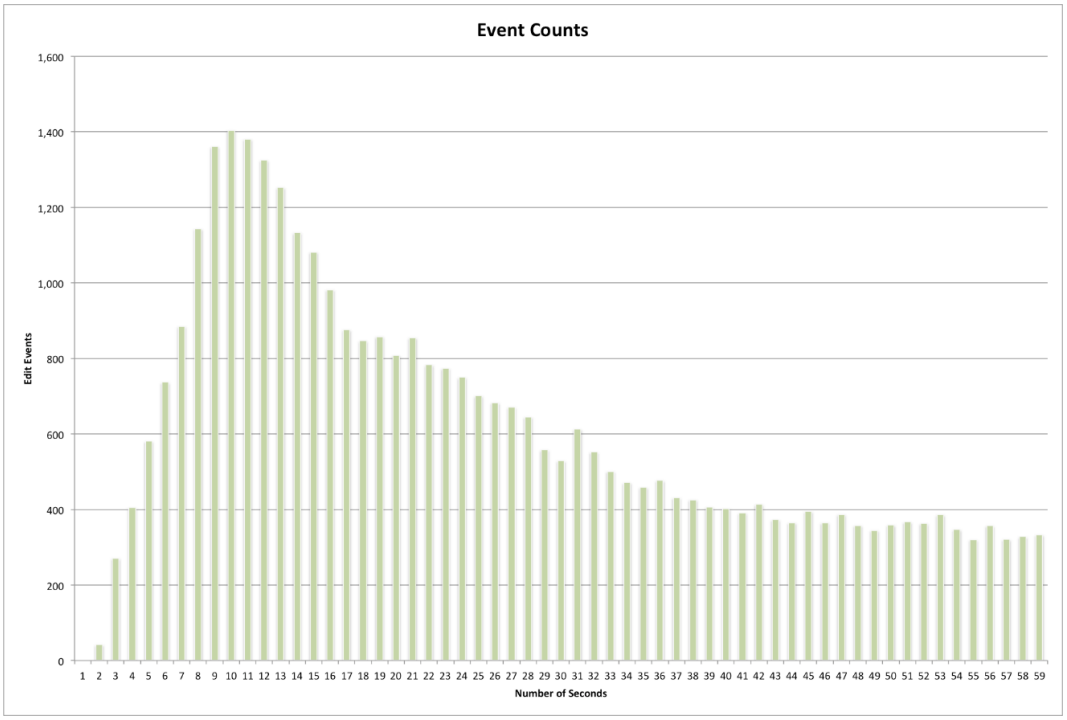

FIG. 5. Distribution of edits up to 60 seconds in duration (n-35,935)

https://doi.org/10.23106/dcmi.952137010

We wanted to eliminate excessively long edits that do not represent "normal" editing and may be errors. Based on the distributions, we chose a range encompassing a majority (97.6%) of edits, establishing a ceiling of 2100 seconds (35 minutes). This threshold leaves 91,916 remaining events; the other 2,306 events were ignored for all further calculations. The average duration of edits lasting 2100 seconds or less was 233 seconds with a standard deviation of 345.48.

4. Discussion

Based on these who/what/how long questions, we wanted to draw reasonable benchmarks for editing activities in our system and to better gauge editing activities. Further comparisons show where times based on kinds of records and edits can provide useful information.

4.1. Time by Item Type

Some records may take longer than others because more information is available to enter, or because it takes more time to skim information on text items than to look at an image and describe it. Although there will always be outliers, the average amount of time by resource type should demonstrate general trends. Table 5 displays edit times by type, including minimum and maximum duration, number of records, total edit time, average (mean), and standard deviation (stddev).

TABLE 5: Average duration of edits (in seconds) by resource type.

Resource Type	Min	Max	Records	Total Time	Mean	Stddev
image_photo	2	2,100	30,954	7,840,071	253.28	356.43
text_newspaper	2	2,084	11,546	1,600,474	138.62	207.30
text_leg	3	2,097	8,604	1,050,103	122.05	172.75
text_patent	2	2,099	6,955	3,747,631	538.84	466.25
physical-object	2	2,098	5,479	1,102,678	201.26	326.21
text_etd	5	2,098	4,713	1,603,938	340.32	474.40
text	3	2,099	4,196	1,086,765	259.00	349.67
text_letter	4	2,095	4,106	1,118,568	272.42	326.09
image_map	3	2,034	3,480	673,707	193.59	354.19
text_report	3	1,814	3,339	465,168	139.31	145.96

As expected, text items tend to take longer, though edit time for photographs is also high. This may be due to the number of photograph records created from scratch, especially when other sources were consulted. The largest spike is in the average time for patent records; this is likely because patent records are being created from near-blank templates and require a large amount of information. We also use patent records for library students or volunteers to experiment with creating metadata, so a number of these editors are new and may tend to take longer than experienced editors.

Based on this information, we can say that editors should expect to spend roughly 10 minutes per patent record, once they are familiar with the system. Some item types are more ambiguous. For example, photographs have a lower average time, but they are a mix of records written from scratch and those edited less extensively. It is still helpful for editors and supervisors to know that most of the time, editing photograph records for longer than 5 minutes is excessive. In this case, more information about the collection would provide a better sense of expected average times.

4.2. Time by Collection

In general, we have internal knowledge about which collection records were primarily created from scratch versus those that required cleanup or less extensive additions. While editors may

https://doi.org/10.23106/dcmi.952137010

conduct different kinds of activities within a collection, the average amount of time (see Table 6) should still give a sense of time spent on records, especially if combined with other information.

TABLE 6: Average duration of edits (in seconds) by collection

Collection Code	Collection Name	Min	Max	Edit Events	Duration Sum	Mean	Stddev
TLRA	Texas Laws and Resolutions Archive	2	2,083	8,418	1,358,606	161.39	240.36
ABCM	Abilene Library Consortium	3	2,100	5,335	2,576,696	482.98	460.03
TDNP	Texas Digital Newspaper Program	3	2,095	4,940	1,358,375	274.97	346.46
TXPT	Texas Patents	5	2,084	3,946	563,769	142.87	243.83
OKPCP	Oklahoma Publishing Company Photography Collection	4	2,098	5,692	869,276	152.72	280.99
JBPC	Jim Bell Texas Architecture Photograph Collection	3	2,095	5,221	1,406,347	269.36	343.87
TCO	Texas Cultures Online	2	1,989	7,614	1,036,850	136.18	185.41
JJHP	John J. Herrera Papers	3	2,097	8,600	1,050,034	122.10	172.78
UNTETD	UNT Theses and Dissertations	2	2,099	6,869	3,740,287	544.52	466.05
UNTPC	University Photography Collection	3	1,814	2,724	428,628	157.35	142.94

Table 6 presents average edit times by collection for the ten most edited collections. The same spike for patents appears here since the Texas Patent collection has a one-to-one relationship with the patents (resource type). There is also a higher-than-expected average for the Jim Bell Texas Architecture Photograph Collection, even when compared to similar collections (e.g., the University Photography Collection). However, the primary editor for this collection often opened many records so that they would be loaded and waiting; this action skewed the data since the system calculates duration based on when the record was opened, rather than on activity.

4.3. User Activities

Our second research question focused on identifying kinds of editing activities by user. We looked at statistics for the ten most active editors (see Table 7).

TABLE 7: Statistics of edits by user for the top ten editors.

Username	Min	Max	Edit Events	Duration Sum	Mean	Stddev
htarver	2	2,083	15,346	1,550,926	101.06	132.59
aseitsinger	3	2,100	9,750	3,920,789	402.13	437.38
twarner	5	2,068	4,627	184,784	39.94	107.54
mjohnston	3	1,909	4,143	562,789	135.84	119.14
atraxinger	3	2,099	3,833	1,192,911	311.22	323.02
sfisher	5	2,084	3,434	468,951	136.56	241.99
cwilliams	4	2,095	3,254	851,369	261.64	340.47
thuang	4	2,099	3,010	770,836	256.09	397.57
mphillips	3	888	2,669	57,043	21.37	41.32
sdillard	3	2,052	2,516	1,599,329	635.66	388.30

In some cases, editing activities are more apparent when editors are working at different levels on a set of items. Figure 6 shows the average edit by editor for legislative text (type) items.

https://doi.org/10.23106/dcmi.952137010

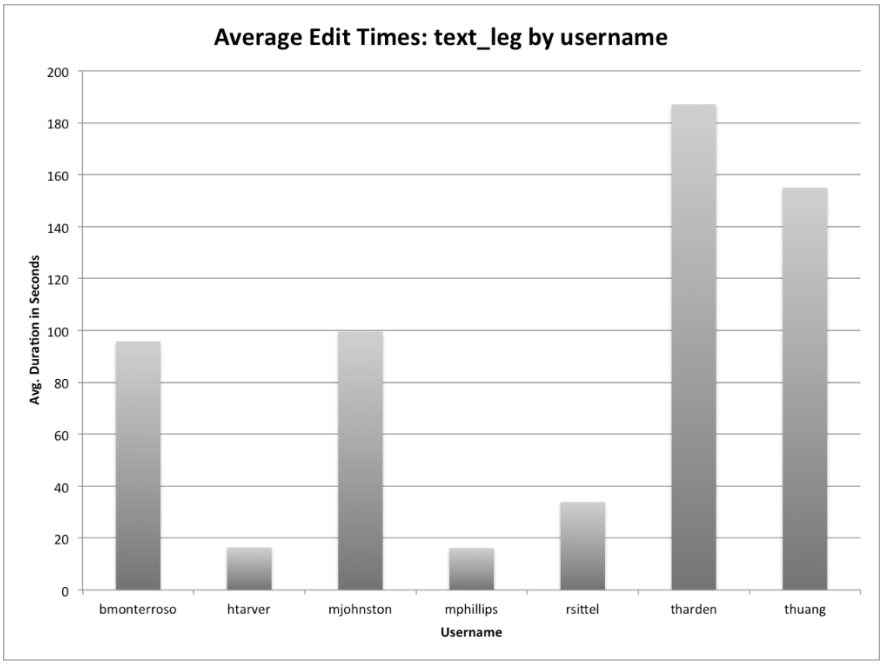

FIG. 6. Average duration of edits by user for legislative text items.

Most records with a legislative text resource type had little starting information. Several of the editors in the set -- htarver, mphillips, and rsittel (all classified as librarians) -- have significantly lower average times than other editors, suggesting that they were performing less extensive edits, compared to editors who spent longer amounts of time on the items. It also shows a trend where students (mjohnston, tharden, and thuang) are primarily "creating" records by adding significant amounts of information while librarians and staff (including bmonterroso) generally supervise by making minor changes and corrections.

While this is not entirely conclusive, we can distinguish "new record creation" versus "minor edits" when compared to average times of similar types, other items in the collection, or against various users. In the future, this provides an opportunity to isolate activities and find a reasonable average time per record creation based on comparable collections.

5. Conclusions

This paper describes an exploratory analysis of the 94,222 metadata edit events logged by 193 editors in the UNT Libraries' Digital Collections from January 1 to December 31, 2014. Based on data collected from the Edit Events system and information known about the records and editors, we discovered that multiple variables affect editing times, but we can generalize about the kinds of activities and how close various edits come to a "normal" or average duration.

5.1. Benchmarks

We particularly wanted to know if we could use gathered editing data to define general characteristics for certain kinds of editing projects within our system. Overall, we discovered that edits of any kind are unlikely to take more than 35 minutes and the average time for those edits is only three minutes and fifty-four seconds.

When monitoring a project, it may be useful to see if the average time is near four minutes or if it differs significantly. However, based on the analyses in the previous section, we can also take into account the resource type and kind of collection. For a text-based collection, we would expect the average time for "creation" to be closer to ten minutes, rather than the system average.

https://doi.org/10.23106/dcmi.952137010

Additionally, we could use average duration to determine the kinds of edits -- in particular, whether users are acting as "creators" and primarily making large additions or significant changes versus "editors" enacting relatively minor edits and corrections. Likewise it should be possible to identify users of browser automation tools, such as Selenium[1], to streamline the editing process.

Distinguishing between "creators" and "editors" could be applied to tracking projects when users with different roles are working on a collection; e.g., two editors "creating" records and keeping them hidden while a third (supervisor) reviews and publishes the records. We would expect the first two editors to have similar duration averages while the third user might have a substantially lower average. Project-level benchmarks could be based on average times by role.

In terms of our research questions, we *can* determine the general kinds of editing activities and create project-based benchmarks based on similar project variables from information in this study.

5.2. Next Steps

Building on this initial study, several comparisons could augment precision in our benchmarks. Pairing the number of metadata edits per collection and partner institution with the average user durations would make it possible to identify administrative editors in the system, or those who are metadata "creators." This may lead to more accurate item-type or collection-level benchmarks when general averages do not fit a project well.

Additionally, it is possible to calculate the total amount of time spent on a given record by adding the edit durations, either by one user or for all users. This information could be valuable for establishing the average amount of time needed to fully complete a metadata record.

One area of interest, which we were not able to explore in this particular study, is to assign hourly costs to users based on ranks (librarian, staff, student, or unknown) and to calculate approximate costs paid by UNT for employee editors versus time "donated" by non-employees. Additionally, if more information can be gathered about the editors -- such as metadata experience -- it may be possible to determine if other variables affect average durations and the cost-per-record.

5.3. Further Study

The analysis presented in this paper is a first step. Although statistics for other institutions may be affected by differences in system interfaces or kinds of collections, staff at other repositories could collect similar data to see if trends match our findings and build benchmarks for editing their collections. It may also be helpful for other groups to use similar criteria identified in this study as a starting point, particularly resource type and collection information since those seem to provide a reasonable cross-section for benchmarking average metadata creation times for many materials. Additional work could also pinpoint which criteria or combination of criteria are most useful for outlining benchmarks based on this kind of data.

Exploring data and the aggregate statistics in this dataset may allow researchers to help metadata editors and administrators produce higher quality metadata records for less overall cost.

References

Phillips, Mark Edward. (February 2015). UNT Libraries 2014 Editors Event Dataset [dataset]. http://digital.library.unt.edu/ark:/67531/metadc502990

University of North Texas Libraries (2015). Input Guidelines for Descriptive Metadata (revised version). Retrieved from http://www.library.unt.edu/digital-projects-unit/input-guidelines-descriptive-metadata

[1] http://www.seleniumhq.org/projects/ide/

https://doi.org/10.23106/dcmi.952137010

Appendix A: UNT Editing System

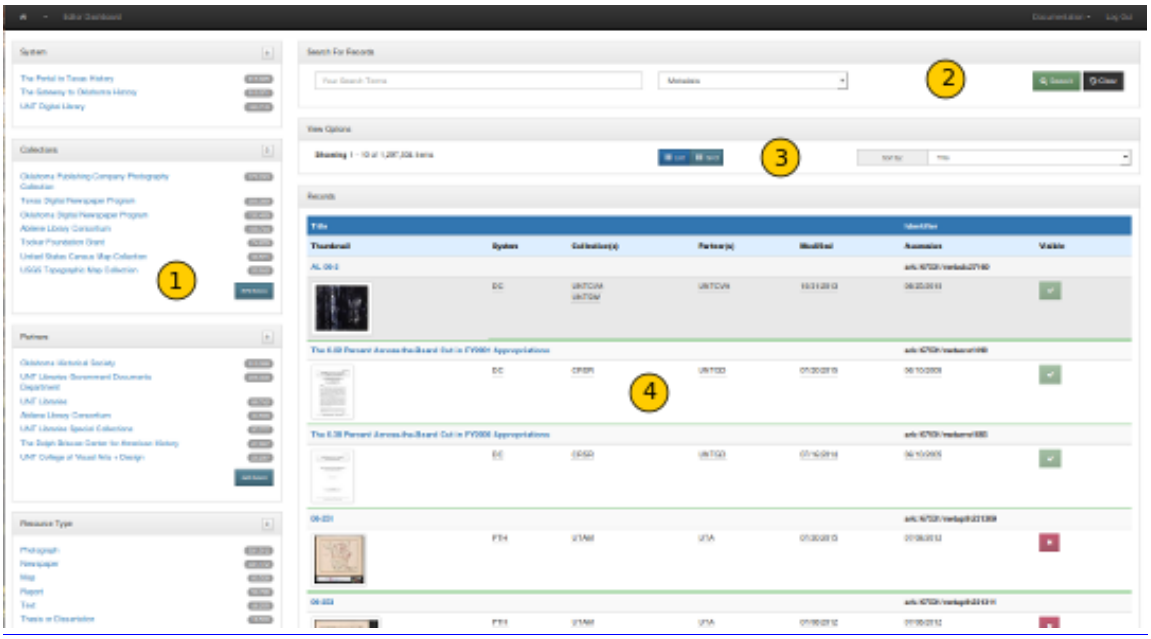

FIG. 7. Screenshot of the user Dashboard in the Editing System.

When a user logs into the Edit System, he sees a list of all records for which he has access. Clicking a title or thumbnail will open the item record in a new tab or window (see Fig. 8).

Dashboard Features

1. Facet options to narrow records by system interface, collection, partner, resource type, and public visibility (when applicable).
2. Search bar to find terms in records, with a drop-down menu to limit searches to a specific field.
3. Options to display item records in a thumbnail grid or list (shown here) and a drop-down menu to sort by the dates records were added or modified, item creation dates, or unique ARK identifiers.
4. List of item records displaying the title, thumbnail, system, partner, collection, date added and modified, ARK identifier, and public visibility status for each.

https://doi.org/10.23106/dcmi.952137010

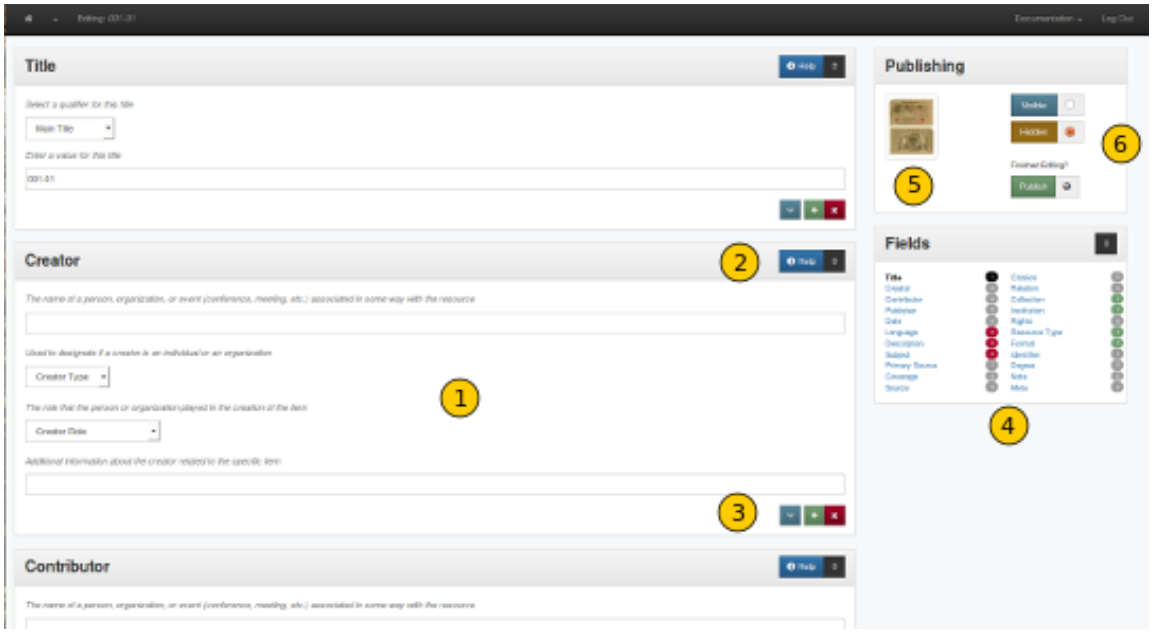

FIG. 8. Screenshot of an item record in the Editing System.

This is the view of a metadata record containing an incomplete template for an item.

Record Fields

1. Text box(es) and/or drop-down menu(s) appropriate for the field are displayed in a bounded box with a title bar.
2. The title bar for each field includes a "Help" link to the guidelines for the field (which open in a pop-up modal), as well as an icon to collapse the field.
3. At the bottom of the field, buttons allow a user to insert symbols and add or remove field entries.

Navigation

4. All of the fields are listed on the right side of the screen and are clickable so that an editor can go directly to a specific field. A bubble next to each field title lists the number of entries in the field; the bubbles are color-coded to show if required fields have values (red = no value, green = value present) and to highlight invalid dates or insufficient subjects (yellow).
5. Clicking the thumbnail opens a new tab displaying all images (pages, views, etc.) for the item.
6. Radio buttons let an editor change the status (visible to or hidden from the public) and the "Publish" button saves all changes to the record.

https://doi.org/10.23106/dcmi.952137010

Appendix B: UNT Edit Event System

FIG. 9. Screenshot of the Edit Event System dashboard.

The Edit Event system dashboard displays current statistics at the time the page is accessed. Each of the buttons is clickable, to show additional statistics for specific dates, users, or records.

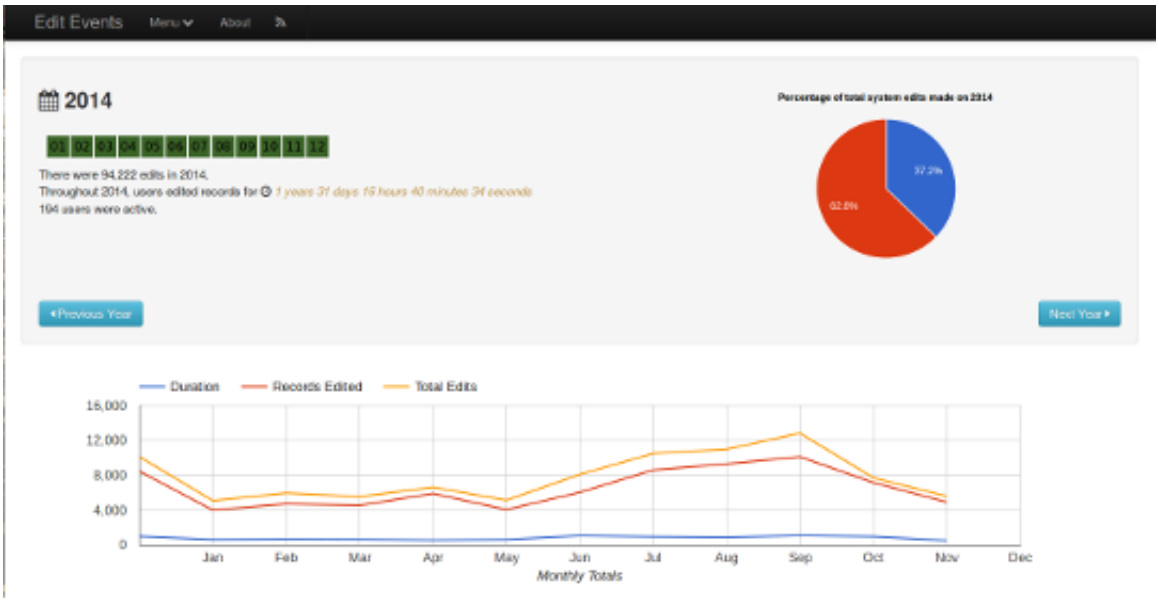

FIG. 10. Screenshot of the statistics page for the 2014 calendar year.

This is an example of a more specific page, showing overall information for 2014. Months are listed across the top to limit by a particular month, followed by a summary of various statistics associated with the chosen date (e.g., total edits, total time, number of editors, etc.). The pie chart shows how the number of edits during the year (blue, 37.1%) compares to the number of other edits (red, 62.9%) logged by the system. The graph at the bottom has lines showing average duration, number of records edited, and total edits throughout the year. Similar statistics are displayed at every level, depending on relevant information for that date, user, or record.

https://doi.org/10.23106/dcmi.952137010