IFLA WLIC 2016 COLUMBUS

Submitted on: 01.06.2016

A Case Study of Metadata Creation in the University of North Texas Libraries' Digital Collections

Hannah Tarver
University of North Texas Libraries, Digital Projects Unit, Denton, USA.
E-mail address: hannah.tarver@unt.edu

Oksana Zavalina
University of North Texas College of Information, Denton, USA.
E-mail address: oksana.zavalina@unt.edu

Mark Phillips
University of North Texas Libraries, Digital Libraries Division, Denton, USA.
E-mail address: mark.phillips@unt.edu

Abstract:
This paper presents a case study of the work carried out by the University of North Texas (UNT) Libraries and the intersection of metadata modeling, metadata input rules and documentation, metadata quality assessments, and technology to empower metadata editors to create high-quality metadata. Further, the examples illustrate how each of the pieces has informed the development of the others. The goal of the paper is to provide a cohesive overview of how documentation, information organization, metadata, and technological infrastructure function together in large digital collections to address multiple needs.

Keywords: metadata creation, metadata documentation, technical specifications, digital collections.

INTRODUCTION

The University of North Texas (UNT) Libraries started actively collecting and generating digital content around 2002 and had publicly-available materials online as early as 2004, making the Digital Collections more than ten years old. Currently, the Digital Collections comprise more than 1.5 million objects that reside in a single technical infrastructure, accessible through three separate user interfaces: The Portal to Texas History (a collaborative site hosting the digital materials from more than 300 partner institutions and private collectors across the state of Texas), the UNT Digital Library (a repository of government documents, and materials owned by UNT or created by the UNT community, including scholarly output), and the Gateway to Oklahoma History (collected materials from the Oklahoma Historical Society, primarily newspapers and newspaper photos).

In 2009, at the conclusion of the Interface Optimization for Genealogists (IOGENE) grant[1] funded by the Institute of Museum and Library Services (IMLS), the Portal launched a new iteration of the public interface, which was later used for all of the Digital Collections. A team of programmers developed the software in-house using open source components. The system has two primary pieces: the Coda repository where original master files are archived, replicated, and maintained; and the Aubrey access system which organizes and displays the derivative web copies and current descriptive metadata. The UNT Libraries makes use of Archival Resource Key (ARK) identifiers for unique identification and as part of the persistent identifier strategy for resources added to the Digital Collections.

The UNT Libraries has locally created software to manage the whole range of services required in a modern digital library. In 2014 the UNT Libraries created an Open Source Software Policy[2] that encourages and discusses various ways that our programmers can release and make contributions to software projects. Additionally, the Software Development Team in the Digital Libraries Division has started to release and develop more of their software on a public GitHub repository.[3] The UNT Libraries also developed a policy clearly designating all metadata and catalog records created at the UNT Libraries as Public Domain and available under a CC0 license.[4]

At UNT, we adhere to the idea of "perpetual beta" and roll out new services and features as needed, or as components are completed and put into production. Now (spring 2016) we are testing a beta version of the Portal user interface; it is the first complete redesign since 2009. As the Digital Collections at the UNT Libraries continue to mature and grow, several of the underlying models, documentation, and technologies play a role in providing a stable framework for content description and a wide range of technological development. The following sections detail UNT metadata decisions, tools, and practices.

METADATA AT UNT

Metadata in the Digital Collections is based on locally-qualified Dublin Core, with the addition of several descriptive metadata fields to meet local needs (see Table 1, with local UNT fields in bold). Although we allow some flexibility in the usage of these fields, all item records contain the same twenty-one possible fields, regardless of collection, resource type, or source. One additional administrative field -- Meta -- documents information automatically captured by the system, or special flags (such as the system/s in which the item is visible), but it is not generally editable. These twenty-two fields comprise the local "UNTL" (UNT Libraries) metadata format around which we build systems and functionality.

Since the Digital Collections' content is built through collaborative relationships within the university and across the state, metadata is created and edited by a variety of users from UNT and from partner institutions. The majority of new record creation is completed by student employees hired on the UNT campus, however, editors represent a full spectrum of trained library professionals, locally-trained students and staff members, and volunteers with no previous experience.

To address issues of quality, consistency, and the need for guidance at several levels (for both experienced and new editors), metadata managers created extensive internal and publicly-facing documentation.[5] We start with a general concept of completeness, by establishing the eight field values we require for every item in order to meet the criteria for a "minimally-viable" record in the Digital Collections.[6] For some of the fields, such as Title and

Description, specific qualified information is required (e.g., a Main Title and Content Description) even if additional information is available or relevant for the item (e.g., a Serial Title or Physical Description). In the Portal and the Gateway, every record must have at least one subject term from the UNT Libraries Browse Subjects (UNTL-BS), which is a hierarchical vocabulary, created and maintained locally.

Field	*Qualified*	*Repeatable*	*Vocabularies*	*Description*
Title	Yes	Yes	Qualifier only	Published title, or a short tagline identifying the item
Creator	Yes	Yes	Type, Role	Entity with primary responsibility for creating the item
Contributor	Yes	Yes	Type, Role	Entity with secondary responsibility for creating the item, or a provenancial connection
Publisher	No	Yes	N/A	Name and location of an entity that formally published the item
Language	No	Yes	Controlled list	Language of any visible or audible text in the content of the item (may be "No Language")
Date	Yes	Yes	Qualifier only	Known/approximate date or date range when the item was created, harvested (websites), submitted or accepted (patent applications), or when an embargo on a digital item will lapse
Description	Yes	Yes	Qualifier only	Attributes of the physical object (e.g., material or size) and of the item's content
Subject	Yes	Yes	Qualifier, UNTL-BS	Terms that describe the content of the item, from controlled vocabularies and/or keywords
Primary Source	No	No	Controlled list	Whether or the item is considered a primary source (e.g., a photograph or original letter) or not (e.g., a book written based on other sources)
Coverage	Yes	Yes	Qualifier, Time period	Information about the geographic, geospatial, and temporal extent of the item's content
Source	Yes	No	Qualifier only	A larger item or event from which the item is taken or derived
Citation	Yes	Yes	Qualifier, Peer reviewed	Component parts of the formal citation for the item, such as volume, issue, or page start and end
Relation	Yes	Yes	Qualifier only	Explanation and link to another item that has a particular kind of connection to the item described
Collection	No	Yes	Controlled list	Broad category for like items based on an archival collection, funding source, topic, or other criteria
Institution	No	No	Controlled list	Partner department or institution that owns or manages the physical materials
Rights	Yes	Yes	Qualifier, Access, License	Information about access privileges, licensing, and other relevant statements for the item
Resource Type	No	No	Controlled list	Classification of the original item's physical or content type
Format	No	No	Controlled list	Classification of the original item's physical type
Identifier	Yes	Yes	Qualifier only	Any alpha-numeric code used to identify the item (assigned by a publisher, partner institution, or another entity)
Degree	Yes	Yes	Qualifier only	Information primarily identifying thesis/dissertation degree aspects; also used for UNT departments in which an item was created
Note	Yes	Yes	Qualifier only	Additional contextual, administrative, or item-related information not documented elsewhere in the record

Table 1. List of editable fields in UNTL metadata (required fields are highlighted; locally-added field names are bold).

To introduce brand-new editors to metadata creation, an online "Quick-Start Metadata Guide" outlines the most common usages for each field.[7] Additional webpages provide in-depth descriptions of the expectations and appropriate usage for each field, broken down into bullet points, with example values covering as many scenarios as possible from the basic to the unusual.[8] These pages also allow relatively easy upkeep, to document new examples and precedents as needed.

For certain, highly-specialized materials we also create collection-specific metadata guidelines that delineate the field requirements and usage for describing those items.[9] Although this approach requires some additional maintenance, it saves time and encourages additional consistency for large, on-going collections.

SYSTEM/TECHNICAL DEVELOPMENT

The local success of the UNT Libraries' Digital Collections is due, in large part, to a combination of the conceptual and organizational models of metadata as well as a strong technical foundation on which to build metadata services for both internal and external users. Many of the technical components also exploit the fact that all digital objects use the same possible fields and are therefore built on the same framework.

The UNTL metadata format has a standardized XML serialization that is outlined by an XML Schema.[10] Building on this UNTL format is a Python library for reading, writing and performing common operations on these metadata files, called pyuntl.[11] Additionally, pyuntl includes several other components, e.g.: the programmatic implementation of our "completeness" metric, and conversions from UNTL to other formats such as standard Dublin Core records, and highwire press metadata tags used by Google Scholar.

In a broad sense, the metadata framework also encourages shareability. One of the core values and goals that the UNT Libraries has considered in the development of its digital library infrastructure and the various access, delivery, and preservation systems is that partners who work with The Portal to Texas History, the UNT Digital Library, or the Gateway to Oklahoma History should always have access to the full metadata for their collections. All metadata in the system is available for harvesting via an OAI-PMH endpoint globally for each access system or on a partner or collection basis as needed. The OAI-PMH repository makes available the raw UNTL metadata, the OAI_DC (normalized Dublin Core) metadata, and RDF/XML (Resource Description Framework/eXtensible Markup Language) metadata for each item record. A variety of other metadata formats are also available through a standard API to objects in the Aubrey content delivery system.

Within metadata records, technical specifications can control field values. Whenever possible, we manage information with a limited number of options as a controlled vocabulary for descriptive metadata (see Table 1), preservation metadata, and other system activities. These vocabularies are managed in a UNT Libraries system[12] and documented by the UNTL Controlled Vocabularies App.[13] Each vocabulary is available in a variety of formats including JSON (JavaScript Object Notation), Python data structures, and RDF/XML. Terms in the vocabularies have unique identifiers that meet the expectations of linked, open data; the pyuntl library can convert string representations in the UNTL metadata qualifiers into the unique identifiers for each term in a standardized way.

In addition to various controlled vocabularies, the Digital Collections make use of the UNT Name App[14, 15] to perform name authority control[16] for a number of collections,[17] including the UNT Scholarly Works repository[18] and the UNT Theses and Dissertation collection.[19] At this time, staff members add new name authority files manually, so the primary focus is on names that are not controlled elsewhere and which provide improved local access (such as UNT faculty member names). Name authority files in the UNT Name App have unique identifiers and a field to include alternate identifiers as a way to link out to other established authorities when available, e.g.: the Virtual International Authority File (VIAF), Open Researcher and Contributor IDs (ORCIDs), ResearchGate IDs, International Standard Name Identifiers (ISNI), UNT faculty profiles, or others.

These various services allow for the creation of metadata tools and provide mechanisms for handling metadata on a technical level.

MAKING CONNECTIONS

There are multiple ways that the technical infrastructure, metadata scheme and user interfaces function together. These intersections improve user experiences for both metadata editors and public end-users.

Metadata Editing Interface

To make metadata editing easier and more consistent for creators of different experience levels, the metadata editing interface uses the technical specifications to provide contextual information. Links from each field allow an editor to look at the full written guidelines and examples, and the layout of the metadata form also provides visual cues. For example, a list of the fields on the right side of the screen allows for navigation among the fields and also displays color-coding to show required fields that are not completed (see Appendix A).

Over the past few years, the user interface for the metadata editing system has undergone changes to incorporate additional features that promote practices consistent with our standards. For example, at UNT we've adopted the Extended Date/Time Format (EDTF)[20] to better represent complex dates in a machine readable way.[21] The EDTF makes it possible to represent concepts that are a common for cultural heritage items using standardized formatting, such as created "in the summer of 2001" (2001-22) or "March of 1903" (1903-03). In order to provide feedback to metadata creators as they format EDTF strings, date fields validate against the EDTF standard through a JSON web service[22] that makes use of applications called django-edtf[23] and edtf-validate.[24] If metadata editors enter non-EDTF-valid dates, the metadata form displays a reminder and highlights the field. Additionally, notes for certain fields appear when pieces of information are missing (e.g., when editors enter values without qualifiers).

Controlled vocabularies managed within the system are also directly connected to the metadata entry form. All qualifiers and controlled lists display as drop-down menus, so that editors must choose a term from the list, rather than manually entering terms. Similarly, a UNTL-BS tool appears in a pop-up modal when that qualifier is chosen in the Subject field. Although the modal previously allowed users to enter terms manually, it has been adjusted so that only valid, controlled terms can be added to the record from the current list.

The practice of developing local tools to manage the various activities in our digital library provides flexibility to incorporate modules into the metadata editing system as needed. The UNT Name App connects to designated name fields (creator, contributor, and publisher) to

provide type-ahead functionality for editors so that there is immediate integration of names added to the Name App into the metadata editing system. A metadata editor can either choose a name from the authorized list or enter an alternative name, since the UNT Name App is used only for select collections.

All of these tools make metadata creation easier for editors and reduce errors whenever possible.

Making Metadata Flexible

At the most basic level, the way that the fields are coded into the system ensures that we only add new fields when absolutely necessary, and encourages flexible usage of fields as an alternative. In some cases, the technical infrastructure has directly affected how we adapt metadata to serve various purposes. For example, since serial and series titles are indexed to provide faceting options to narrow search results, we often use series titles as a way to collocate groups of items within a larger collection. This provides a reasonable alternative to creating new collections for every grouping of items that need to be designated for partner or user needs.

Another option is to use the full-text metadata indexing that is already a part of the system by working formatted phrases in the content descriptions of every record in a collection or group. Some partners have specific needs or desires about how they will be able to use/search their own collections within the Portal or the Digital Library that we can meet with this technique. One partner institution had a collection of portraits on glass-plate negatives, primarily images of individuals and small groups, so they wanted to be able to search for the number of persons visible in each image. To facilitate this, we adjusted our usual guidelines to start every description with the statement "Photograph of [#] person(s)." Since the phrasing is identical in every record, someone who knows that this is built into the records could easily search for a specific number of people. For a UNT grant project, we used a similar approach working with specialists to develop a content description that lists every physical characteristic of biological mussel specimens in a formulaic way so that researchers can easily find sub-sets of specific mussel types.

In some cases, the need to more clearly express information (especially across various kinds of collections) has led to changes in the technical implementation. Several years ago, we added a qualifier to the Source field, which had previously been a simple text field. Now an editor can specify the kind of "source" from a list that ranges from larger objects for which the item is a part (e.g., a book, journal, newspaper, etc.) to events at which the item originated or was presented (e.g., a conference, academic course, lecture series, etc.).

Metadata as Data

One important aspect of the Digital Collections metadata editing system is the versioning of metadata records, introduced in the infrastructure during fall 2009. All versions of each record have been saved since its implementation, so that there is always a way to look back at each version (see Appendix B), or to roll back to a previous version if needed. Another component, deployed in October 2013, documents each edit event associated with a metadata editor account (noting length of time, the record involved, and the completeness and visibility of the item at the end of the edit) and compiles overall editing statistics (see Appendix C).

Saving version and event data provides an opportunity to analyze various aspects of metadata creation and editor activities. We have begun initial studies to determine some of the

information that we can learn by looking at metadata versions,[25] comparing changes in records,[26] and evaluating editor times.[27] In the future, we hope to use this data to learn more about determining metadata quality metrics and cost-benefit analysis for metadata creation. If other institutions have similar data, this will also allow for comparative studies to learn more about metadata creation across different systems.

DISCUSSION

A strong conceptual and technical infrastructure for metadata creation, backed by policies that encourage open development and sharing of data, is an important underpinning of the UNT Libraries' Digital Collections. They form a firm foundation to develop existing services and to build new services for both the internal and external users of digital library metadata and content.

In fact, many of the features implemented in the user interface during the 2009 redesign were directly connected to metadata values by leveraging the available data. Most of the facets available to narrow search results are pulled from fields with controlled vocabularies, or which have standardized formatting (e.g., dates and place names). The subject browse categories are based on UNTL-BS terms, so only records with those subjects are findable through that browse option. Adding these features also created a feedback loop for editors, since it highlighted many of the areas where metadata was less consistent than expected, or where mistakes had been made in isolated records.

Based on the experiences at UNT, there are some general recommendations that may be helpful for other institutions. One consideration is how many fields to include in a metadata schema and how open it is to new fields or different usage across a collection. Maintaining consistency within the metadata schema and usage across our holdings has had several benefits, not only functionality in the Digital Collections user interfaces, but also cross-collection connections, ease of sharing through initiatives such as the Digital Public Library of America (DPLA), and a more uniform dataset for analyses.

Regardless of fields, consistency may be the most useful general practice since it allows for the possibility of metadata-based services, but also enables easier migration or future changes, if necessary. Institutions may want to look for fields that can be managed partly or entirely by controlled vocabularies; this improves consistency and also makes metadata creation easier for editors. Whenever possible, controlled vocabularies and any other authority control should align with five-star linked open data[28] and adhere to the highest level possible. In some cases, there may be relatively easy steps to move established controlled vocabularies forward in the linked-data continuum if they do not meet the expectations of five-star data.

Although authority control and schema development can be complex and seem overwhelming, there are many established vocabularies and standards available publicly. Institutions can easily reuse any of these standards that apply, or use them as a starting point for a locally-developed vocabulary or standard. In the Digital Collections, several of our vocabularies are based on established lists from other organizations with local modifications. For example, the "role" vocabulary[29] for creators and contributors uses applicable values from the MARC relator codes[30] and includes additional terms (such as "Artisan" or "Retailer") that are not available in the original list; the vocabulary for relation types[31] is based on the Dublin Core list of Relation refinements,[32] again with local additions (e.g., "Has Translation" and "Is Transcription of").

Finally, written documentation has been invaluable for managing the metadata in our Digital Collections, in both the technical and value input aspects. Ensuring that all metadata editors and technical staff have access to complete metadata documentation, even if it is not publicly available, promotes consistency and opens a dialogue to allow for tools and discussions about how metadata effectiveness can be used to best advantage. In our system, all aspects of metadata creation and management have developed interdependently, sometimes in unexpectedly positive ways. We hope that sharing case studies, such as our experiences at UNT, may provide insight for other institutions developing and managing metadata or considering system changes.

REFERENCES

1. Hartman, C.N. & K.R. Murray. (February 2010). Optimizing the user experience in a rapid development framework: final project report, February 2010. http://digital.library.unt.edu/ark:/67531/metadc32974/
2. http://www.library.unt.edu/policies/other/open-source-software-policy
3. https://github.com/unt-libraries/
4. http://www.library.unt.edu/policies/other/locally-created-cataloging-and-metadata-records-rights-policy
5. Tarver, H. (2010). Better guidelines, better functionality: how metadata supports the cycle of system improvement at the University of North Texas. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, pp. 165-176. http://digital.library.unt.edu/ark:/67531/metadc29323/
6. http://www.library.unt.edu/digital-projects-unit/minimally-viable-records
7. http://www.library.unt.edu/digital-projects-unit/quick-start-metadata-guide
8. http://www.library.unt.edu/digital-projects-unit/input-guidelines-descriptive-metadata
9. http://www.library.unt.edu/digital-projects-unit/project-specific-guidelines-and-documents
10. http://digital2.library.unt.edu/untl.xsd
11. https://github.com/unt-libraries/pyuntl
12. http://digital2.library.unt.edu/vocabularies/
13. https://github.com/unt-libraries/django-controlled-vocabularies
14. http://digital2.library.unt.edu/name/
15. https://github.com/unt-libraries/django-name
16. Tarver, H., L. Waugh, M.E. Phillips, & W. Hicks. (2013). Implementing name authority control into institutional repositories: a staged approach. http://digital.library.unt.edu/ark:/67531/metadc172365/
17. Waugh, L., H. Tarver, & M.E. Phillips. (2014). Introducing Name Authority into an ETD Collection. *Library Management*, *35*(4/5), pp. 271-283. http://digital.library.unt.edu/ark:/67531/metadc279706/
18. http://digital.library.unt.edu/ScholarlyWorks/
19. http://digital.library.unt.edu/explore/collections/UNTETD/
20. http://www.loc.gov/standards/datetime/
21. Tarver, H. & M.E. Phillips. (July 2013). Lessons learned in implementing the Extended Date/Time Format in a large digital library. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, pp. 60-70. http://digital.library.unt.edu/ark:/67531/metadc174739/
22. http://digital2.library.unt.edu/edtf/
23. https://github.com/unt-libraries/django-edtf
24. https://github.com/unt-libraries/edtf-validate
25. Zavalina, O.L., & P. Kizhakkethil. (2015). Exploration of metadata change in a digital repository. Proceedings of iConference 2015, Newport Beach, California. http://digital.library.unt.edu/ark:/67531/metadc503265/
26. Zavalina, O.L., P. Kizhakkethil, D. Alemneh, M. Phillips, & H. Tarver. (2015). Building a framework of metadata change to support knowledge management. *Journal of Information and Knowledge Management*, *14* (1), 1-16. http://digital.library.unt.edu/ark:/67531/metadc505014/
27. Tarver, H., O.L. Zavalina, M. Phillips, D. Alemneh, & S. Shakeri. (2014). How descriptive metadata changes in the UNT Libraries' Collections: a case study. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, pp. 43-52. http://digital.library.unt.edu/ark:/67531/metadc406345/
28. Berners-Lee, Tim. (July 7, 2006) "Linked Data," W3C. http://www.w3org/DesignIssues/LinkedData.html

29. http://digital2.library.unt.edu/vocabularies/agent-qualifiers/
30. http://www.loc.gov/marc/relators/relators.html
31. http://digital2.library.unt.edu/vocabularies/relation-qualifiers/
32. http://dublincore.org/documents/dcmi-terms/

APPENDIX A: UNT EDITING SYSTEM

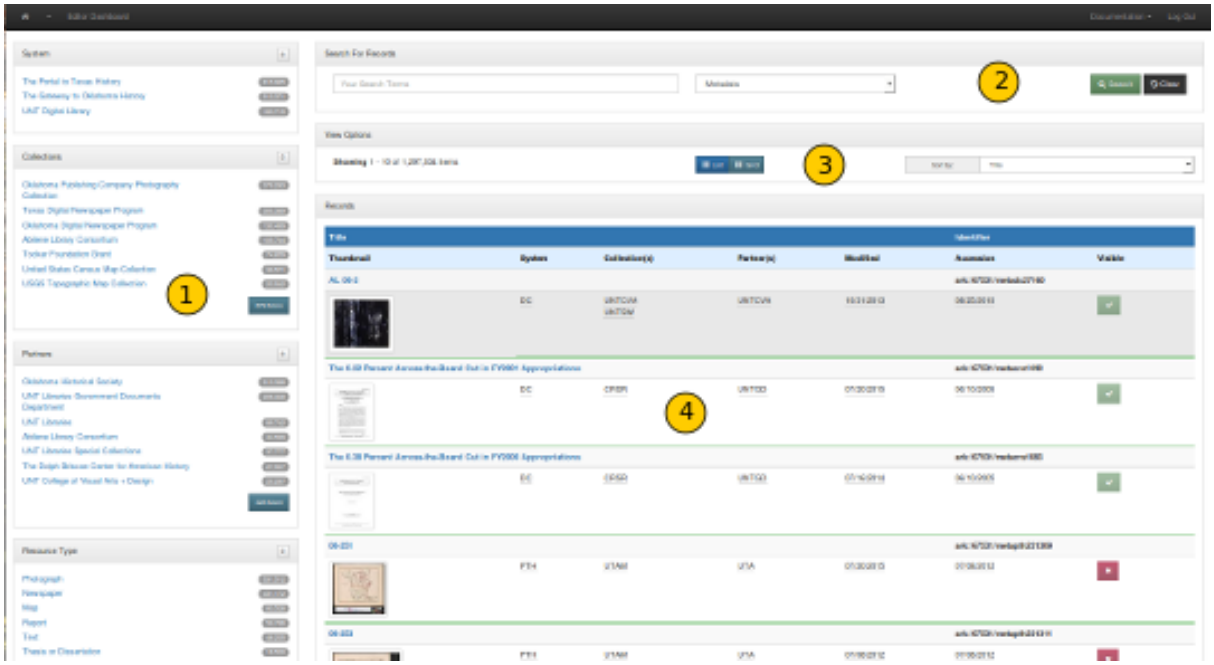

Figure 1. Screenshot of the user Dashboard in the Editing System

This is the Dashboard view that a user sees upon logging into the Editing System for the Digital Collections. Each editor sees the list of records to which he has access. Individual records can be opened by clicking the title or thumbnail (see Figure 2).

Dashboard Features:

1. Facets allow editors to narrow records by applicable criteria such as system interface, collection, partner, resource type, and public visibility.
2. A search bar to find values anywhere in the metadata, or in a specific field.
3. Display options for viewing items as a list (shown here) or a grid and a drop-down menu to sort by titles, dates that records were added or modified, item creation dates, or unique ARK identifiers.
4. Item records listed with the title, thumbnail, system, partner, collection, date added and modified, ARK identifier, and visibility status.

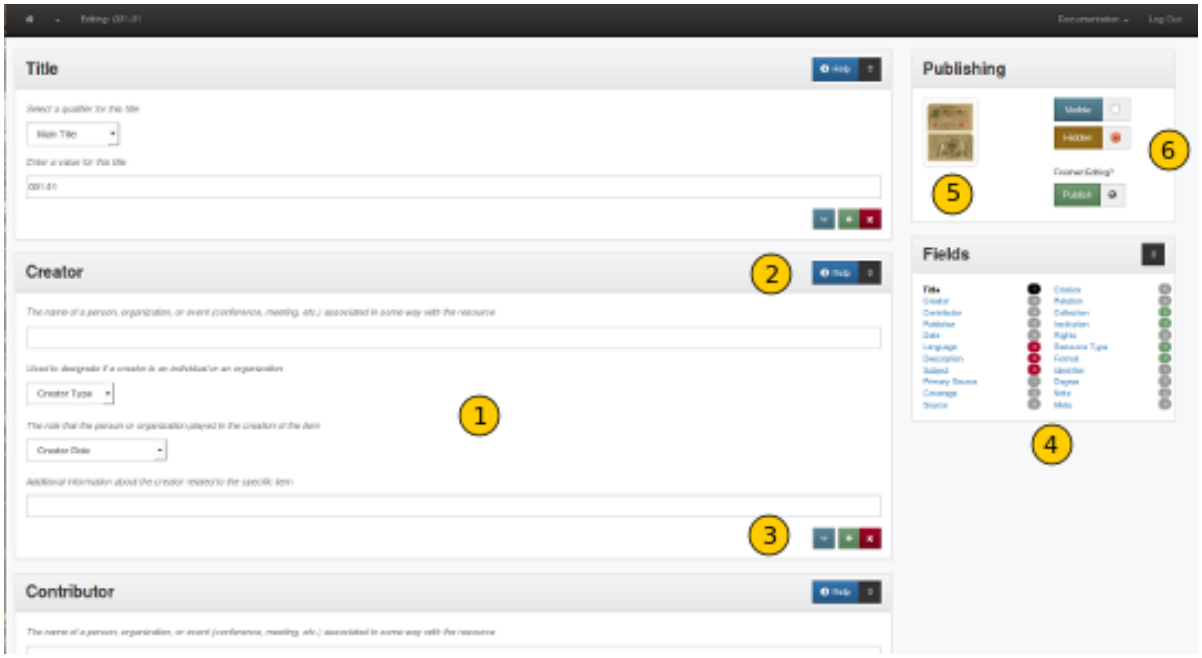

Figure 2. Screenshot of an item record in the Editing System.

The image in Figure 2 shows the metadata form for an item record; only some values are entered in this record.

Record Fields

1. Each field is bounded in a box and has appropriate text box(es) and/or drop-down menu(s) for the metadata values.
2. A “Help” link in the title bar for each field links to the guidelines (which open in a pop-up modal); an icon to collapse the field is also in the upper-right corner of each field.
3. Buttons in the lower-right corner for each field allow users to insert symbols, add, and remove entries.

Navigation

4. A list of fields on the right side of the screen displays the number of entries for each field and provides a clickable list to let an editor jump to a specific field. The entry numbers are color-coded for required fields (red = no value, green = value present) and to highlight invalid dates or insufficient subjects (yellow).
5. The thumbnail is clickable and opens a zoom-able version of all pages/images associated with the item in another tab or window.
6. Buttons in the upper-right section control public visibility of the item (visible or hidden) and allow the editor to “Publish” (save) the version of the record.

APPENDIX B: ITEM RECORD HISTORY

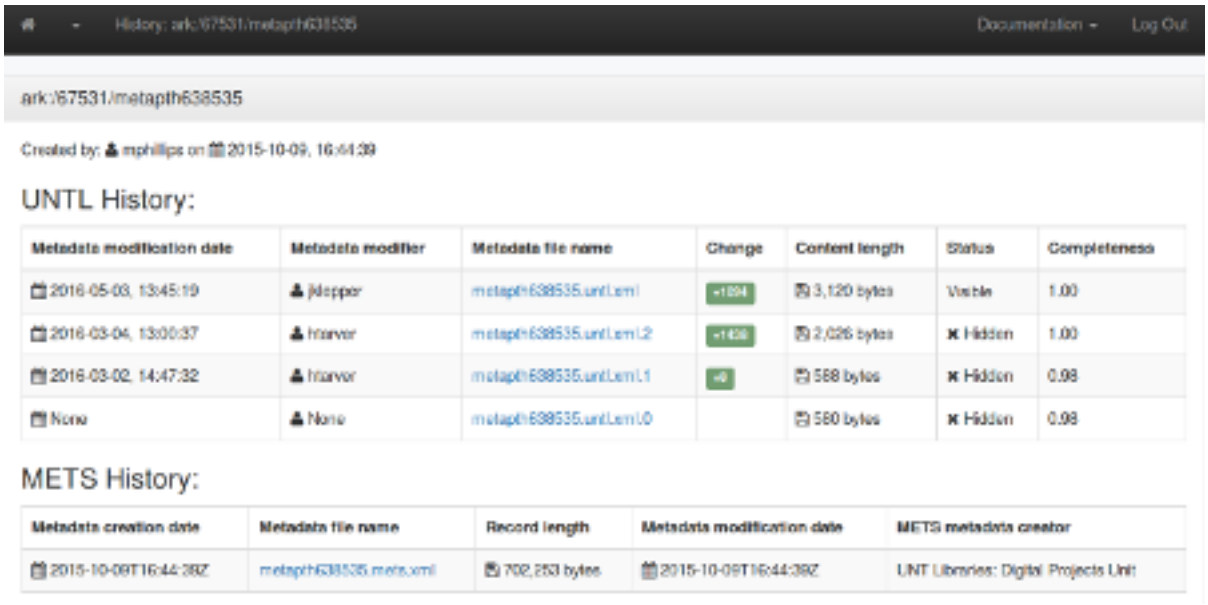

Figure 3. History of an item record including modification dates, editor names, record version, amount of change (by number of characters), size of record, visibility status, and completeness score.

APPENDIX C: UNT EDIT EVENT SYSTEM

Figure 4. Screenshot of the Edit Event System dashboard.

Figure 5. Screenshot of the statistics page for the 2015 calendar year.