Utilizing the EPIC Framework to Improve Metadata Infrastructure

Hannah Tarver[1,*,†], Mark Edward Phillips[1,†]

[1]University of North Texas Libraries, 1155 Union Circle #305190 Denton, TX 76203-5017, United States of America

Abstract

The UNT Libraries manage digital collections comprising more than 4 million items, creating challenges for assessing and improving descriptive metadata quality. Previously, we introduced an iterative model called EPIC (evaluate, prioritize, identify, correct) to guide ongoing assessment work. This paper focuses on changes to infrastructure that have been introduced to support each step in this framework as a case study for ways that other organizations might envision or plan similar work.

Keywords

iterative editing, digital infrastructure, metadata quality, metadata assessment

1. Introduction

Over time, digital collections have become more common and increasingly visible on the Internet, representing massive amounts of digitized and born-digital content. The Digital Public Library of America (https://dp.la) currently aggregates records from more than 53 million items from 45 different hubs (representing hundreds of institutions); Europeana (https://www.europeana.eu/en) also aggregates more than 60 million items from institutions across the European Union; other institutions worldwide maintain or contribute records for even more cultural heritage materials and scholarly works.

As digital collections grow, it becomes even more important for metadata to be complete and accurate to help users find specific items within these collections, and to ensure that user functionality (such as search filters or browse options) will work appropriately. While print materials such as books and technical reports may be findable through OCR, or a combination of full-text searching and descriptive metadata, many cultural heritage materials are image-based (such as maps, photographs, artworks, etc.) or have little printed text (handwritten letters, music scores, historic legal documents, etc.); users may have to rely solely on the completeness and accuracy of metadata to find and access these materials [1]. Similarly, addressing metadata quality can support increasing accessibility online. A 2018 study focused on users with print disabilities noted that items were often not findable due to misspellings and other metadata errors and also suggested increased use of faceting or similar functionality [2], which often draw from values in item records.

DCMI-2025 International Conference on Dublin Core and Metadata Applications

*Corresponding author.

†These authors contributed equally.

hannah.tarver@unt.edu (H. Tarver); mark.phillips@unt.edu (M. E. Phillips)

0000-0003-2344-9268 (H. Tarver); 0000-0002-9679-6730 (M. E. Phillips)

https://doi.org/10.23106/dcmi.952584682

Additionally, the process of reviewing metadata records or assessing whole collections and making corrections or enhancements increases in difficulty proportionate to size. During a 2019 survey, metadata practitioners were asked about methods of evaluation and a majority who answered (73%) said that manually checking records was the primary or only way that they evaluated metadata quality in their collections [3]. Although certain errors can only be found by looking at an individual record, manual evaluation techniques are not particularly scalable, especially as digital holdings grow to hundreds of thousands or millions of items [4]. This suggests an ongoing need for tools and strategies to approach metadata assessment and quality improvement.

2. Background

The Digital Collections hosted by UNT Libraries comprise more than 4 million items that are archived in a single, home-grown digital library system and are accessible via three different public interfaces. The Portal to Texas History (https://texashistory.unt.edu) contains cultural heritage materials contributed by partner institutions across the state; the UNT Digital Library (https://digital.library.unt.edu) primarily contains materials owned by the university (including special collections) or created at the university, such as scholarly works by faculty and staff; the Gateway to Oklahoma History (https://gateway.okhistory.org) contains cultural heritage materials from the Oklahoma Historical Society and other partner institutions in that state.

While the item types represent a broad range of image- and text-based materials — including photographs, newspapers, artworks, letters, maps, technical drawings, books, etc. — as well as audio/visual materials, the Digital Collections have a single, standardized metadata schema applied to every metadata record. The UNTL schema is locally-qualified and Dublin Core-based, with 21 possible DC and local metadata fields. There are 8 fields required in every record: main title, language, content description, two subjects, resource type, format, collection, and partner institution [5]. All other fields may be used according to local formatting and usage guidelines when they apply to an item.

In 2024, the Portal celebrated its 20-year anniversary, commemorating two decades of work in digital libraries, although the Digital Collections have changed over that time. During this period, we have done extensive research and work around managing digital collections and, specifically, assessing and improving the quality of metadata in the growing number of item records [6, 7]. In particular, as the amount of metadata has increased, new challenges have also arisen in relation to the practicality of reviewing or evaluating metadata records — both on an individual level and at more scalable system or collection points — as well as making ongoing corrections and enhancements to records to improve findability and functionality within public interfaces [8].

3. The EPIC Framework

Based on the work at UNT related to digital collections and metadata creation, we introduced the idea of the EPIC framework in 2020 [9], to more clearly refine the approach that we have taken to metadata quality and how it may be useful for other institutions. EPIC stands for

https://doi.org/10.23106/dcmi.952584682

Evaluate, Prioritize, Identify, and Correct; it is intended to be an iterative process, as depicted in Figure 1, to constantly improve various aspects of metadata records, while acknowledging that they may never all be considered "complete" or "perfect." This framework is also intended to be system agnostic, so that it can apply to any local context. For example, not every institution would have access to the same kinds of tools so "Evaluate" may not look the same for different collections, but any digital library can do some level of metadata assessment.

Additionally, this framework can support long-term metadata quality work, including iterative changes to individual records, as well as organizational planning. Comprehensive metadata quality assessment and enhancement requires time and resources that include personnel and the acquisition or creation of tools or programs to make this work easier as collections expand.

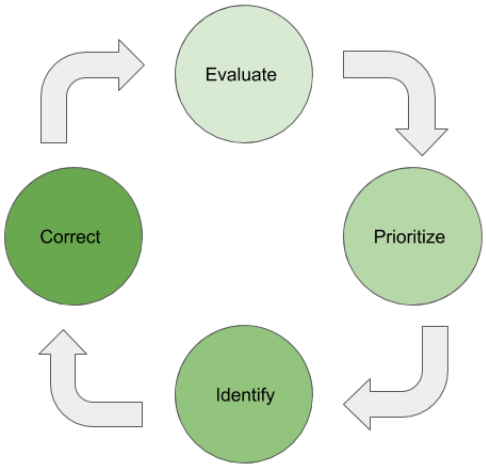

Figure 1: Illustration of the EPIC Model.

Although some work has been done to describe the initial research and development of the EPIC framework [10], this paper attempts to provide more details about the local approach to this process. It also focuses more specifically on steps taken to build out new components or enhance existing infrastructure to support metadata quality work in each of the phases.

3.1. Evaluate

The first step requires some level of assessment for a collection of records to determine what issues already exist or need to be reviewed. This could encompass a number of issues for any aspect of quality, including value formatting (e.g., personal names should be inverted, but some are not), field usage (e.g, published books should generally have at least one named author, but some records have no creator listed), or applicability (e.g., the subject "pangolins" was added to an entire collection of photos, but is not relevant for all of them).

There are many ways to evaluate aspects of metadata, but one area of significance not always highlighted is the quality and depth of metadata input guidelines (or similar documentation,

https://doi.org/10.23106/dcmi.952584682

which may include a Metadata Application Profile). The same previously-mentioned 2019 survey found that around 65% of responses to the question about evaluation methods mentioned using metadata guidelines for determining metadata quality [3]. For editors checking their own work, or even metadata professionals evaluating records system-wide, the person reviewing metadata values or field usage must have clear instructions and documentation to determine when values are correct (or incorrect) and which records or values need further review.

Metadata documentation has been an ongoing project at UNT, starting with a complete overhaul of formatting more than a decade ago [11]. As the next step, in 2022, all metadata documentation moved out of the library website infrastructure and into a Sphinx instance [12]. This required all of the guidelines to be reformatted from Markdown into restructured text (ReST), now stored in a GitHub repository [13]. The pages are linked from the UNT Libraries website and pulled into the UNT domain, so the documentation appears integrated but does not take up space directly in the website content.

The decision to move the documentation involved several factors, particularly the difficulty of maintaining large, complex pages of content in the Libraries website, which is primarily intended for more basic "flat" pages of information about departments and services. No mark-up language is perfect, but ReST allows more flexibility for certain kinds of formatting, especially tables. Our existing metadata guidelines use a number of tables to group relevant information, put guidelines and examples side-by-side, and generally make information more organized and readable for editors who are skimming to find specific references. As noted in the 2010 paper regarding guideline formatting, we continue to prefer HTML-based documentation over statically-created PDF versions because it allows for easier linking, updating, and incorporating into other tools and interfaces.

All of the guidelines were thoroughly reviewed and updated during the re-formatting process. Storing the guidelines in GitHub also creates automatic version control and allows for more collaboration in creating and updating documentation, including ways to flag problems or make suggestions using "issues" within the GitHub infrastructure. This supports ongoing efforts to update usage guidelines with additional examples or document new precedents whenever issues arise.

Moving out of the website structure also provided opportunities to expand into additional documentation related more tangentially to metadata creation, including authority issues and metadata assessment. Not only can editors consult the guidelines when evaluating a metadata record, information specifically about how to do assessment with available tools and what editors should look for are now also available in conjunction with the usage documentation. These factors all help metadata creators and managers evaluate metadata quality more easily in the Digital Collections as a regular part of metadata curation and management.

3.2. Prioritize

Perhaps the most subjective part of this process is prioritizing issues to decide which should be addressed first. Depending on local needs and resources, this could be based on how easy an issue is to fix, the impact on user functionality, availability of subject specialists, or other criteria.

At UNT, the Digital Collections have a number of tools that have been integrated into

https://doi.org/10.23106/dcmi.952584682

the edit system and can be accessed by any editor (or supervisor) to review metadata values and field usage in various ways. Although some systems use spreadsheets to enter metadata (or to export and re-import corrected metadata), the UNT system is not compatible with spreadsheets; since most fields are qualified, the data tends not to be rectangular. Additionally, there are "helper" components integrated directly into the item record level to provide support or warnings. Features can be added at almost any time, depending on the complexity and developer availability; larger design changes are also implemented on a regular, rotating basis. In addition to the general functionality, this section will focus on some of the ways that we are continuing to enhance these features:

Count. The first tool is called "count" and can display the number of entries-per-record for a field (or field and qualifier), which is useful for finding missing values or outliers. For example, an editor could look at the number of creator entries for a photograph collection and see if there are records with no creator — which could be expected, or may need review, if the photographer is known — or records that have three or four creators, which generally would not make sense for a photograph. Similarly, the counts could be limited to just values labeled "photographer" (since we include a role for every creator or contributor), and if it does not match the number of total creator entries, some values may be mis-labeled.

Facet. Second, the facet tool lists all of the unique values for a specific field (or field-qualifier combination), which is useful for finding misspellings or variations in terms that ought to be the same. For example, an editor could look at all of the subject values in a collection and find places where proper names were entered slightly differently or incorrectly. Limiting to a specific subject type (such as named persons or LCSH) would also show values that do not match formatting guidelines or deviate from a vocabulary's authorized terms.

Cluster. Finally, cluster can be used to take all of the values for a field (or field-qualifier combination) and either run them through an algorithm to find terms that have the same normalized value (similar to functionality in OpenRefine), or to sort all of the values according to characteristics such as length or alphanumeric pattern. This could uncover issues such as inverted and non-inverted names that would have the same normalized key but may not sort close together in a large list (e.g., John Smith and Smith, John), values that are extremely long or short (e.g., 1-character subjects), and values that don't match expected patterns (e.g., YYYY-MM-DD for dates or local identifiers that have standard formatting).

Aside from the general functions of these tools, there are also features that help editors prioritize issues, as well as a number of components that have been added to improve these tools in the last year. For example, in cluster, an editor can search results by the most members (i.e., clusters with the most variations in strings that normalize to a single value) or by the most records (i.e., total records affected, regardless of the number of variations). This can make it easy to find clusters that will have the most impact — either by normalizing a large number of variations or changing one variation that will affect a large number of total records — especially when there are a large number of clusters to review.

A newer feature assists with flagging values that may be incorrect, including terms that do not match certain controlled lists and any extremely similar values that sort adjacently. If an editor is reviewing a particular collection, there may be a number of values that are misspelled or normalized differently that are flagged; if an incorrect value is only in one or two records, it is generally worthwhile to fix it immediately. Other times, a value may be in a larger number of

https://doi.org/10.23106/dcmi.952584682

records and the issue gets added to one of our running lists of problems for editors to fix during spare time. Issues that may require more time or expertise also go on lower-priority lists until someone has time to resolve them.

3.3. Identify

The identification step is about determining or isolating the specific set of records affected by a particular issue or concern. Historically, this was more complicated within our local workflows because our tools were only available outside of the edit interface, by downloading records via OAI-PMH and then doing various analysis using Python [14]. Since then, we have integrated metadata tools directly into the editing system (count, facet, and cluster). For descriptive metadata in item records, this step now largely relies on those tools linking directly from a value or criterion being evaluated to the records that need to be addressed or reviewed.

However, we have been expanding metadata activities into records that serve various administrative, authority, and other purposes to support or enhance the value of item records. In particular, during 2022, we introduced a database to create title records for serials and series as well as authority records for locations [15]. All of these records are maintained within administrative interfaces in our existing Django infrastructure, but also have public-facing records.

Title records. These records have fields similar to traditional MARC-based serial records and document information related to a specific title, including: the kind of publication, the frequency, the start/end dates for the title, a publisher, and preceding/succeeding titles. Each record is also assigned a unique identifier that is also incorporated into the permanent URL for the title record. In the public view, a title record will automatically display the current digital holdings by keying off an identifier included in each item record — either an existing OCLC or LCCN, or the unique UNT title identifier. Assigning new identifiers at the title level allows us to represent items without previous catalog records and also provides the flexibility to create a holdings record for a series of items that may all have different OCLC numbers by including the UNT identifier in those records. Public browse options include a list of "curated" titles (i.e., that have title records) and item records connect directly to title records when applicable.

Location records. These records are intended primarily to document the current authorized version of place names used within the Digital Collections to ensure that they are controlled across item records. Each location record has only a handful of possible fields including the component parts (administrative divisions) for each place name string, some general descriptive flags for special types of locations (such as historic places), identifiers for GeoNames and Wikidata, latitude and longitude, a note, and a check box to mark locations "verified" by UNT staff. Although there is not currently a browse-able public interface for location records, they do display publicly with a direct URL including the unique record identifier. Integrating some of these features into public interfaces is slightly more complicated, since we want to continue to allow for the flexibility of editors to add new locations to item records when needed, without disrupting the current ways that the Digital Collections use that data.

When these databases were first introduced, all places already in the Digital Collections had automatically-generated records, although many had only a place string (e.g., no identifiers, and not verified); all newspaper titles already in the system also got automatically-generated

https://doi.org/10.23106/dcmi.952584682

records. Those automatic records need review in some cases to ensure that information carried over correctly and some required manual enhancement (e.g., verifying locations or designating preceding/succeeding titles). In each of these databases we only create records when the title or location exists in the Digital Collections (or will, in the immediate future). For example, the goal would never be to add an authority record in our system for every place name that could exist, just for the places that are already represented in item records (or for items that are awaiting upload). However, adding significant amounts of administrative metadata on top of the existing item description work creates new challenges.

One issue is knowing what records need to be created for new titles or locations. Item metadata is added by a large number of editors (usually 80-90 unique editors per month) from various library departments and even other institutions, but administrative records for non-newspaper serials and titles or locations added after the initial generation must be created manually by staff in the Digital Libraries Division. To assist with this, a list of locations or titles "not_in_database" is created in the interfaces for each of the public sites (though not linked to user functionality). In each case, the system compares values in public records against the relevant database and then displays any values that do not have a matching administrative record. This can also highlight errors more quickly by showing, for example, when a non-authorized place name is entered for a verified location, or when a title is mis-matched with an identifier and doesn't get collocated into the holdings list.

While this primarily assists in identifying "missing" administrative records, this work still directly impacts users by providing more consistency in metadata records and additional functionality (such as title records with contextual data). In the future, this administrative metadata will be leveraged to provide more support to metadata editors and enhancing user features in the public interfaces.

3.4. Correct

Despite the various tools now available within the editing system for the Digital Collections, the system was originally designed without batch editing options, requiring editors to open individual records to correct or change information one-at-a-time. For minor errors or typos, this is not a significant issue, but some issues require normalizing formatting or values across thousands of records due to incompatible source values, changes over time related to formatting decisions or even values themselves (e.g., LCNAF that have been updated with death dates, countries or administrative units that have shifted over time, etc.), and other factors.

Over the years there have been different tools and approaches used by metadata editors trying to automate some of this work. These often involve the use of browser automation tools that can be programmed with behaviors and repeated in succession across open tabs in browsers. This provided one way of editing records without having to manually open, navigate, change, and then submit the changes to the system. The challenge with this approach is that it was entirely based on the html interfaces of the edit system. If the interface changed to improve the human user experience, previously created actions or recipes would no longer work. Another challenge is that simple edits with predictable field ordering was the general use case for this workflow; it would not be feasible as the changes became more challenging, or when it was not easy to predict the order of instances of an element to change, for example in some records a

https://doi.org/10.23106/dcmi.952584682

```
[
    {
        "op": "test",
        "path": "/language/0",
        "value": {
            "content": "eng"
        }
    },
    {
        "op": "replace",
        "path": "/language/0",
        "value": {
            "content": "spa"
        }
    }
]
```

Figure 2: JSON Patch to test and change language value.

term being changed could be in the second subject position and in others the third. While many of these automation tools allow for high levels of programming and customization, the work needed to set up and test the automations made this approach challenging to use for all but a few metadata editors, especially as the number of "easy" edits has also been largely exhausted, leaving changes that require more human attention.

To better address the challenge of batch editing, we needed an approach that simplified the process and allowed for a different level of scaling. Beyond changes identified through the revision process, there are situations where large-scale modifications of records would be required to enable new features or data models within the metadata records to be implemented (e.g., adding appropriate standardized rights statements to every record). These kinds of tasks are almost impossible without a way of programatically updating records within the system.

Currently, the Digital Libraries' Software Development Unit is working on a two-phase approach to this work. First the core component was to develop a programmatic interface that would assist in the automated editing of individual records, locally called the Edit API. This is a low-level API that allows submission of changes to modify a record; these changes would be validated, queued for applying, and applied to the record, before re-indexing the content. This has been built in a similar way conceptually as the Item Metadata API: Write at the Internet Archive [16]. In this model, the web standard of JSON Patch [17] is used to submit changes to the system that should be applied to the underlying metadata file. These patches are submitted at the record level where they are queued in a standard queuing system for future processing. An example JSON Patch for testing and then changing the language value of a record is presented in Figure 2. This allows for large numbers of edits to be put into the queue and then processed asynchronously within the system. Once verified and processed, the records are re-indexed and the changes are visible in the production system.

The next piece being built is a command-line tool for helping a human user interact with the

https://doi.org/10.23106/dcmi.952584682

Edit API — called the Metadata Edit Helper — and provides an interface to add, remove, and replace elements within records with a few simple commands. This can be applied to a single record or a list of records, allowing for a way to apply edits to dozens or even hundreds of records at once. The Metadata Edit Helper provides an abstraction to the end user where they can indicate what needs to change, and the tool will help to figure out where in the metadata record element the values are and change them accordingly. At this time the tool is used by higher-level metadata managers who have experience working with the command-line and who have worked within the editing systems for many years.

These two component tools are the important building blocks to provide a true "batch" editing interface for the metadata in the UNT Libraries' Digital Collections. The next step will be to design a web-based interface that incorporates lessons learned in the work so far that will allow less technically-oriented users the ability to edit across their records instead of just editing individual records. There are many unanswered questions about the final implementation at present, including, of course, concerns about the scope of editing that users will have, the ability to roll back changes if a mistake is made, and who will or will not have access to these tools. We hope that over the next year we are able to answer some of these questions with an initial release of a tool for our users. This will provide a new level of editing and correcting records that will allow us to meet the requirements of scale as our collections continue to grow.

4. Conclusions

Although the EPIC framework is a direct manifestation of how we have conceived of metadata and worked with descriptive records over the past two decades, we have continued to expand how we think about the component parts and address possible gaps or opportunities for improvement in our own system and workflows. The framework was originally developed as a way of approaching the metadata assessment and correction work to support our local digital collections, but we believe it could be useful for other organizations that may be grappling with similar issues. This systematic and iterative approach is helpful at an initial stage — e.g., as an organization may be considering a new or comprehensive assessment project in preparation for system migration or local initiatives — but it can also serve as a way to identify areas that would most benefit from additional resources or innovation, even for collections that have robust assessment workflows. This paper also shows how non-technical elements, like documentation, guidelines, and tutorials are an important component of a large digital library infrastructure.

Using the EPIC framework provides a way to discuss how new technological innovations can assist one or many components of the framework and how they might fit into existing local processes or resource constraints. This paper positions the EPIC framework as a way of thinking about improving the infrastructure to support the creation and curation of metadata in our organizations, but is only one specific aspect of nurturing a healthy metadata ecosystem. The inclusion of EPIC during planning provides a lens for expressing how improvements either directly or indirectly support different components in the life cycle of metadata improvement. Our hope with this paper is to demonstrate how EPIC can be used in different ways and support conversations and other work related to metadata quality.

https://doi.org/10.23106/dcmi.952584682

References

[1] J. Attig, A. Copeland, M. Pelikan, Context and meaning: The challenges of metadata for a digital image library within the university, College & Research Libraries 65 (2004) 251–261. doi:10.5860/crl.65.3.251.

[2] W. M. Beyene, T. Godwin, Accessible search and the role of metadata, Library Hi Tech 36 (2018) 2–17. doi:10.1108/LHT-08-2017-0170.

[3] S. Gentry, M. Hale, A. Pyant, H. Tarver, R. White, R. Wittmann, Survey of benchmarks in metadata quality: Initial findings [white paper], 2020. URL: https://digital.library.unt.edu/ark:/67531/metadc1637685/.

[4] A. Tani, L. Candela, D. Castelli, Dealing with metadata quality: The legacy of digital library efforts, Information Processing & Management 49 (2013) 1194–1205. doi:10.1016/j.ipm.2013.05.003.

[5] University of North Texas Libraries, Minimally-viable records, 2022. URL: https://library.unt.edu/metadata/minimally-viable-records.html.

[6] M. E. Phillips, H. Tarver, Experiments in Operationalizing Metadata Quality Interfaces: A Case Study at the University of North Texas Libraries, International Conference on Dublin Core and Metadata Applications 2018 (2018). doi:10.23106/dcmi.952139004.

[7] H. Tarver, O. Zavalina, M. Phillips, A case study of metadata creation in the university of north texas libraries' digital collections, 2016. URL: https://library.ifla.org/id/eprint/1325/.

[8] H. Tarver, Current data on the digital collections: A statistical report, 2024. URL: https://digital.library.unt.edu/ark:/67531/metadc2360958/.

[9] H. Tarver, M. E. Phillips, EPIC: A Proposed Model for Approaching Metadata Improvement, volume 1355, Springer, Cham, 2021, pp. 228–233. doi:10.1007/978-3-030-71903-6_22.

[10] H. Tarver, M. E. Phillips, A. Krahmer, EPIC: an iterative model for metadata improvement, International Journal of Metadata, Semantics and Ontologies 15 (2021) 244–253. doi:10.1504/IJMSO.2021.125885.

[11] H. Tarver, Better Guidelines, Better Functionality: How Metadata Supports the Cycle of System Improvement at UNT, International Conference on Dublin Core and Metadata Applications 2010 (2010). doi:10.23106/dcmi.952109784.

[12] University of North Texas Libraries, UNTL metadata documentation, 2022. URL: https://library.unt.edu/metadata/.

[13] University of North Texas Libraries, Github - unt-libraries/untl-metadata-documentation: Metadata documentation for the untl metadata format, 2022. URL: https://github.com/unt-libraries/untl-metadata-documentation/.

[14] M. E. Phillips, Metadata analysis at the command-line, Code4Lib (2013). URL: https://journal.code4lib.org/articles/7818.

[15] H. Tarver, Implementing a serial title database: A UNT case study, in: North American Serials Interest Group, NASIG Autumn 2023, 2023. URL: https://digital.library.unt.edu/ark:/67531/metadc2201684/.

[16] Internet Archive, Item metadata API: Write, 2022. URL: https://archive.org/developers/md-write.html.

[17] P. C. Bryan, M. Nottingham, JavaScript Object Notation (JSON) Patch, RFC 6902, 2013.

https://doi.org/10.23106/dcmi.952584682

URL: https://www.rfc-editor.org/info/rfc6902. doi:10.17487/RFC6902.

https://doi.org/10.23106/dcmi.952584682