EPIC: A Proposed Model for Approaching Metadata Improvement

Hannah Tarver^[0000–0003–2344–9268] and Mark Edward
Phillips^[0000–0002–9679–6730]

University of North Texas Libraries, Denton, TX, USA
{hannah.tarver,mark.phillips}@unt.edu

Abstract. This paper outlines iterative steps involved in metadata improvement within a digital library: Evaluate, Prioritize, Identify, and Correct (EPIC). The process involves evaluating metadata values system-wide to identify errors; prioritizing errors according to local criteria; identifying records containing a particular error; and correcting individual records to eliminate the error. Based on the experiences at the University of North Texas (UNT) Libraries, we propose that these cyclical steps can serve as a model for organizations that are planning and conducting metadata quality assessment.

Keywords: Metadata evaluation · Metadata quality · Processes

1 Background

The Digital Collections at the University of North Texas (UNT) Libraries comprise nearly three million records and represent more than fifteen years of digitization, web harvesting, and metadata activities. Publicly, users can access a wide range of materials – printed and handwritten text; photos and other images; large maps and technical drawings; audio/video recordings; physical objects; and other items – from three public interfaces: The Portal to Texas History (materials from partner institutions and collectors across the state), the UNT Digital Library (materials owned by UNT or created at the university), and the Gateway to Oklahoma History (materials from organizations in Oklahoma, managed by the Oklahoma Historical Society). Although many of the items have been digitized and described in-house, the Digital Collections also include items and metadata harvested from government databases, or provided by partner institutions; additionally, editors include a wide range of highly-trained specialists, trained students of various expertise, and volunteers who may have little-to-no experience.

Given the size and scope of the Digital Collections, we have conducted extensive quality-control and clean-up projects, as well as research related to methods of determining metadata quality system-wide (e.g., evaluating dates [9], overall change [10], and interconnectedness of values [7]). While there has been extensive research about metadata quality as a concept, most of it has focused on

frameworks and metrics for evaluating the quality of metadata in an individual record or in a larger collection; most notably the quality frameworks put forth by Moen, et al. in 1998 [3], by Bruce and Hillman in 2004 [1], and by Stvilia et al. in 2007 [8]. In each case, these papers outlined specific aspects of metadata quality – such as accuracy or completeness – and how these aspects ought to be defined as distinct and important when evaluating metadata records.

This case study outlines a model based on the experiences at UNT with practical metadata evaluation and correction, but has wide applications for other digital libraries. Rather than focusing on how to determine quality or the best methods of evaluation, the proposed model encompasses the overarching process of iterative metadata improvement, breaking assessment and mitigation into discrete steps in a replicable way.

2 EPIC Model

The model comprises four steps (EPIC): Evaluate (determine quality and identify errors), Prioritize (order errors according to resources and impact), Identify (connect an error to the affected record/s), and Correct (make appropriate changes to affected records and eliminate the error). These steps function as an ongoing, repeatable cycle (see Figure 1).

Fig. 1. Illustration of the EPIC Model.

The first two steps – evaluate and prioritize – occur primarily at a system or aggregated collection level, to determine what errors or issues exist, and to make decisions about which problems should be corrected immediately. The third and fourth steps – identify and correct – refer to individual records as issues are remediated. Although this is a cyclical process, not all steps necessarily take the same amount of time, or happen with the same frequency.

2.1 Evaluate

The key component for assessment is to evaluate the existing level of quality in records and to identify areas that need improvement. Organizations may employ other processes, e.g., “Methods of conducting assessment can include focus groups, surveys, benchmarking, observational analyses, interviews, and methodologies that we borrow from other disciplines such as business”[4]. Ideally, this would cover a range of quality types – e.g., accuracy, accessibility, completeness, conformance to expectations, consistency, provenance, timeliness – as well as content and structural elements. Realistically, some aspects are more easily evaluated or verified than others. Many repositories rely on manually checking records [2] and some problems can *only* be found by looking at a record – e.g., mismatches between the item and record values. However, as a primary method of analysis, manual checks do not scale to large collections and require a significant time commitment. Additionally, manual checks may not be the best method to evaluate overall consistency within a collection(s); however usage of system-wide tools depends on the digital infrastructure an organization is using.

At UNT, we do not have to evaluate structural aspects of metadata (e.g., whether XML is well-formed) because validation happens as part of our upload processes. Similarly, we use the same fields and qualifiers for all records, so it is easy to validate the metadata format. For descriptive content, we use tools (called count, facet, and cluster) that are integrated directly into our system [6]. These are based on similar operations in command-line functions and in OpenRefine, to rearrange metadata values in ways that assist editors to identify records needing correction or review for accuracy, completeness, and consistency.

Before these tools were implemented in 2017, we used a Python-based program (metadata breakers [5]) to evaluate values after harvesting records via OAI-PMH. Other organizations also use the metadata breakers as well as tools based in various programming languages or made to evaluate metadata in spreadsheets [2].

2.2 Prioritize

System-wide, there are likely to be a large number of identified issues, ranging from relatively simple fixes that affect only a handful of records (e.g., typos), to larger issues across entire collections (e.g., legacy or imported values that do not align with current standards). A number of possible criteria may determine which issues are fixed immediately or addressed in the future, based on local needs. In our system, we place higher priority on values that directly affect public-facing functionality or search results – such as coverage places and content descriptions – versus fields that are less-frequently used, or not browseable (e.g., source or relation). Other criteria may include the amount of expertise or time required, number of records affected, available resources/editors, or information known about the materials. At UNT, determining the most useful criteria for prioritization is an ongoing process.

2.3 Identify

This step requires connection between values or attributes and unique identifiers for the relevant records. When we used tools outside of our system, one challenge was determining which records contained errors found during evaluation. Although we had unique identifiers and persistent links in the harvested records, an editor would have to track down every record from a list of identifiers for a particular error. This made the process complicated if an error occurred in a large number of records. Organizations that can export and reload corrected records may not have this problem; however, not all systems have this capability, which means that a specialized process may be necessary to coordinate between value analysis and affected records. Once our tools were integrated into the system, editors had a direct link between attributes and records, although we do have to manually edit individual records to change values.

2.4 Correct

The final step is to correct records identified with a particular issue. Although we encourage editors to change other, obvious errors that they might see in the records, the priority is on fixing specific issues rather than more comprehensive reviews. This process promotes relatively quick, precise editing system-wide, resulting in iterative improvements. With tool integration, any editor in our system may access the tools (e.g., for self-review) and managers can link to records containing an error – or to a tool with set criteria for multiple errors – to distribute review and corrections. Other organizations could choose to approach correction more comprehensively, or according to methods that work within their systems; the essential purpose of this step is to ensure that records are improved, even if further edits are needed.

3 EPIC as a Cycle

We have described metadata editing in stages, but we approach it as an ongoing cycle, where some stages take longer – or happen more often – than others. Higher-priority issues or small issues affecting few records may be fixed immediately while others are documented for later correction. Additionally, we add thousands of items every month and complete most records after ingest, so the possibility of new errors is a constant concern. Organizations that employ more rigorous review prior to upload or that add new materials less frequently, might evaluate and prioritize issues periodically, or on a schedule (e.g., annually, biannually, or quarterly).

Evaluation can also inform managers about the work of less-experienced students and mitigate problems related to individual editors. Additionally, we start with the premise that a metadata record is never “done,” i.e., there may always be new information, updated formatting or preferred guidelines, or the identification of errors. Under this approach, a single record may be edited or “touched” many times and/or by many editors, but the overarching goal is always that metadata improves over time.

3.1 Benefits of EPIC

We use and suggest this approach for several reasons:

The cycle works at scale. Iterative metadata editing works for collections of any size; but, importantly, it is functional even as a system gets extremely large, when it is more difficult to determine record quality and to direct resources. Evaluating quality at the system level ensures meaningful changes that can improve consistency, findability, and usability across large numbers of records, even when quality in individual records may remain less than ideal.

Individual processes allow for planning. Although we are presenting this model as distinct steps in the cycle, in practical use some components may overlap or only occur occasionally – e.g., an editor evaluating records at a collection level may identify a small number of records containing a problem and correct them without “prioritizing.” However, it is difficult to plan ongoing work that would be required for system-level issues affecting thousands of records without breaking down the components. Some problems are more systemic or less important than others and we need a way to discuss and determine the best approach – i.e., prioritization – as well as methods within each step that are most appropriate for a particular digital library or institution. The model also provides a way to contextualize the experiences and research within the metadata quality domain from various organizations.

Division of labor. Fixing systemic problems requires a large number of people over time. By determining specific issues and documenting them – preferably with some level of priority – we can assign records to editors based on a number of criteria. For example, some errors are time consuming (e.g., they affect a large number of records) but are easy to correct without training or expertise by new editors, volunteers, or editors working outside their area. Other problems may affect fewer records but require familiarity with metadata or a subject area; those could be assigned to appropriate editors as time permits.

This has been particularly useful during the COVID-19 pandemic when we offered metadata editing tasks to library staff who could not perform their usual duties from home. Approximately 100 new editors were added to our system and assigned to metadata work with relatively little training. Those editors can make inroads on some of the simple, yet frequently-occurring problems (such as name formatting in imported records).

4 Conclusion

This proposed model explicates discrete steps in metadata improvement to disambiguate conversations around these activities. For example, discussions about how to evaluate metadata (or specific fields) is sometimes conflated with prioritization of issues or relative importance of individual fields. We have found these to be two distinct and equally important facets of metadata quality. Likewise, identifying records within a large collection that have a specific set of metadata characteristics can be challenging by itself, sometimes requiring specialized

tools, indexing, or analysis separate from other components. Organizing planning activities and methods within individual steps provides additional context and more productive discourse. At UNT, we are continuing our research around metadata evaluation and prioritization of quality issues, and also plan work to test and apply the EPIC model in a more structured way. By presenting this model, we hope to encourage other researchers and practitioners to explore these components individually as they create tools, workflows, and documentation that supports metadata improvement.

References

- [1] T. R. Bruce and Diane Hillmann. “The continuum of metadata quality: defining, expressing, exploiting”. In: *Metadata in Practice*. Ed. by Diane Hillman and Elaine L. Westbrook. ALA editions. Chicago: ALA Editions, 2004. ISBN: 9780838908822.
- [2] Steven Gentry et al. *Survey of Benchmarks in Metadata Quality: Initial Findings*. Tech. rep. May 2020.
- [3] W. E. Moen, E. L. Stewart, and C. R. McClure. “Assessing metadata quality: findings and methodological considerations from an evaluation of the US Government Information Locator Service (GILS)”. In: *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries*. Apr. 1998, pp. 246–255. DOI: 10.1109/ADL.1998.670425.
- [4] Rebecca L. Mugridge. “Assessment of Cataloging and Metadata Services: Introduction”. In: *Cataloging and Classification Quarterly* (Nov. 2017), pp. 435–437.
- [5] Mark Edward Phillips. “Metadata Analysis at the Command-Line”. In: *Code4Lib* (19 Jan. 2013).
- [6] Mark Edward Phillips and Hannah Tarver. “Experiments in Operationalizing Metadata Quality Interfaces: A Case Study at the University of North Texas Libraries”. In: *Proceedings of the International Conference on Dublin Core and Metadata Applications* (2018), pp. 15–23.
- [7] Mark Edward Phillips, Oksana Zavalina, and Hannah Tarver. “Using Metadata Record Graphs to Understand Digital Library Metadata”. In: *Proceedings of the International Conference on Dublin Core and Metadata Applications* (2019), pp. 49–58.
- [8] Besiki Stvilia et al. “A framework for information quality assessment”. In: *Journal of the American society for information science and technology* 58.12 (2007), pp. 1720–1733.
- [9] Hannah Tarver and Mark Edward Phillips. “Lessons Learned in Implementing the Extended Date/Time Format in a Large Digital Library”. In: *Proceedings of the International Conference on Dublin Core and Metadata Applications* (2013), pp. 60–70.
- [10] Hannah Tarver et al. “How Descriptive Metadata Changes in the UNT Libraries’ Collection: A Case Study”. In: *Proceedings of the International Conference on Dublin Core and Metadata Applications* (2014), pp. 43–52.