Enhancing the Quality of Metadata: Modular Approach to Digital Resource Lifecycle Management

Daniel Gelaw Alemneh and Mark Edward Phillips; Digital Projects Unit, University of North Texas; Denton, Texas

Abstract

Quality is a multidimensional concept. The two aspects of digital library data quality are the quality of the data in the objects themselves, and the quality of the metadata associated with the objects. Maintaining usable and sustainable digital collections necessitates maintaining high quality metadata about those digital objects. The University of North Texas Libraries recognize the strategic benefit of metadata as a means of ensuring long term access to its digital resources. This paper discusses issues related to digital resources management and describes how the University of North Texas Digital Projects Unit approaches metadata quality issues at various levels of the digital resources life cycle. It also suggests a number of metadata quality assurance procedures, tools, and associated quality assurance mechanisms.

Introduction

Digital libraries and supporting technologies have matured to the point where their contents and structures are incorporating complex and dynamic resources and services. The University of North Texas (UNT) Libraries have created an application framework for integrating diverse digital information resources from a multitude of participating institutions. The undertakings of the UNT Libraries include: the CyberCemetery, Congressional Research Service Reports Archive, the World War Poster Collection, Federal Newsmaps and other materials drawn from collections throughout the libraries.

One UNT Libraries digital libraries initiative, the Portal to Texas History (PTH) is a state-wide collaborative digital project that offers students and lifelong learners a digital gateway to the rich collections held in Texas libraries, museums, archives, historical societies, and private collections. It features digital reproductions of photographs, maps, letters, documents, books, artifacts, and more. In addition, Portal Primary Source Adventures that comply with TEKS (Texas Essential Knowledge and Skills) standards highlight relevant materials for young scholars and classroom teachers.

Considering the role of standardized metadata in digital resource life cycle management, the UNT Libraries actively promote metadata-based digital resource management. The existing metadata system empowers participating institutions to describe digital objects in a consistent way that provides for optimum searching, discovery, and retrieval, while ensuring long-term preservation of digital resources.

The UNT Libraries Metadata

The Portal to Texas History collaborators share a number of goals, including ensuring long-term, easy access to a wide variety of cultural heritage collections. The Portal provides a metadata framework that fosters a collaborative environment for participating institutions.

The UNT Libraries metadata element set comprises Dublin Core-based descriptive metadata along with detailed technical and preservation metadata elements that document how digital resources are created, formatted, arranged, identified, and sustained with application of appropriate preservation procedures. While promoting interoperability with widely accepted standards, the recommended UNT Libraries metadata elements allow flexibility at the local level to integrate existing and anticipated content, processes, and systems. The complete documentation is available at: http://www.library.unt.edu/digitalprojects/metadata/.

Figure 1. Metadata Records Added to the Portal to Texas History (from May 2004 to February 2007)

As can be seen from Figure-1 above, the number of metadata records added to the Portal to Texas History has seen great increase and continues to rise at a consistent rate. As the volume and complexity of resources increased, the need for highly-developed resource management tools that could ensure quality and consistency became apparent.

Metadata Quality

Metadata quality is a crucial issue for cultural heritage communities. Metadata errors occur in a variety of forms, but when errors exist, in whatever form, they block access to resources. Metadata quality has a profound impact on the quality of services that can be provided to users. The problem is particularly acute if there are multiple institutions participating in a collaborative digitization project such as the Portal to Texas History, where a high level of interoperability is an important element.

The metadata quality characteristics depend on various factors, including: user perspectives, needs, and priorities, which vary across groups of users. Metadata researchers have assessed metadata record quality by examining subject term specificity and exhaustivity, metadata record completeness, and other known substantive factors. The literature documents metadata quality in terms of:
• Error free, (such as adding/selecting wrong information in the wrong field/subfield, typographical errors):
  o Letter transposition, e.g., 0207 for 2007
  o Letter omission, e.g., Society for Society
  o Letter insertion, e.g., and for an
  o Letter substitution or mistrakes, e.g. anu for an any
• Noambiguous, (e.g., incomplete information)
• Non-ambiguous, (e.g., multiple spellings, multiple possible meanings, mixed cases, inconsistency, etc.)

Although no consensus has been reached on conceptual and operational definitions of metadata quality; all emphasize the importance of metadata quality. [4]. Errors, omissions, and ambiguities in the metadata affect the consistency of search results and high recall of available resources. [5]

In order for end users to benefit fully from the development of digital libraries, service providers and collaborators need to maintain a high level of consistency across diverse digital collections. The International Federation of Libraries Associations (IFLA) has identified four critical user information needs:

• **Find:** Various fields would be used as search criteria to find a specific resource.
• **Identify:** From the full record retrieved via a search, the most useful fields would display at the top level.
• **Select:** When multiple records result from a search, the short listing enables the user to quickly select the most useful records retrieved.
• **Obtain:** Obtain access to a resource.

**Factors influencing metadata quality**

Quality services depend on good metadata, but most metadata values are not very good [6]. Several metadata commentators identify factors that may influence metadata quality itself, and the effectiveness, efficiency, practicality, and scalability of the processes used to create it. [7]. Based on the literature review and our own UNT experiences, the following section summarize the major factors that affect metadata quality. [8].

**Local requirements**

In order to understand existing or local requirements, the following issues should be considered and addressed:

• What type of objects will the repository contain? [Heterogeneity]
• What functionality is required locally? What are the associated digital rights issues? [Content packaging, repackaging and repurposing]
• How will they be described? And used? And by whom? [Granularity, determining purpose and level of detail]
• What entry points will be used? [The type of access, templates, interfaces, etc.]

**Collaborators’ requirements**

Although collaborating institutions have much in common, they may have conflicting metadata requirements which may call for significantly different approaches. Library assumptions about metadata quality may not be appropriate in wider context. [7]. To come up with effective, practical, and sustainable metadata creation processes, the following metadata quality influencing factors should be considered:

• What is the nature of the institutions’ digital objects? [Museum objects, archives, historical documents, scholarly documents, etc.]
• How does the information-seeking behavior of their respective users differ? [Historians, genealogists, students, researchers, etc.]
• Does participation in the wider community impose specific requirements?
• What is required for interoperability? [Structure, semantics, and syntax.]
• Are requirements formal or informal?
• Will metadata be meaningful within aggregations of various kinds?
• Will access restrictions be imposed?

**Training Issues**

In most collaborative projects, non-professionals or volunteers create metadata, often working in isolation without adequate tools. Training issues greatly influence initial quality of metadata created. Some important considerations are:

• Who will be involved? What skills do they have?
• Are all actors qualified to produce the required metadata quality? [Very unlikely]
  o If not, what are the training needs?
• Are there adequate support mechanisms for those creating metadata? [Online tutorials, guidelines, FAQs, and other documentation]
• Is there sufficient supervision to ensure that actors receive regular feedback?

**Cost**

Creating and managing high quality metadata is an expensive endeavor [8]. Cost-effectiveness is an important factor that needs to be taken into account. Among other considerations:

• What resources are available locally?
• How can these resources be used to best effect?
• Are resources sufficient to produce the required metadata quality? [Very unlikely]
  o If not, what are the priorities? [Cost/Benefit]

Based on cost and benefit analysis, some lower metadata quality may be tolerated. The trade-offs between the various ways in which metadata quality can be improved and their costs can be considered.

All these issues significantly impact the quality of services including the consistency of search results and high recall of available resources. The impact of each factor, however, differs from institution to institution and even from project to project, depending on the type of repositories, economics, and the heterogeneity, size and scale of the collections and users.

**UNT metadata quality assurance mechanisms**

Responsible and viable metadata management activities should address a number of quality issues. The increase in the number and heterogeneity of digital resources has lead UNT to develop tools, workflows, and quality assurance mechanisms that allow for quick and effective metadata analysis and quality assurance.

The goal of the UNT Libraries metadata management team is to achieve metadata that is error free, without omissions, and non-ambiguous in order to enhance accuracy, relevance, accessibility, consistency, and coherence in our digital libraries. Accordingly, the UNT Libraries metadata management system provides cost
effective and scalable mechanisms to detect errors and clean up values to improve the consistency and overall quality of data. The following section describes very briefly some of the tools and quality control mechanisms used at the UNT Libraries.

The metadata template diagram in Figure-2 above shows one of the quality control tools in the UNT metadata creation workflow. It is a self-checking metadata entry template that ensures that all mandatory elements have values before the record is added to the system. In other words, no null value is allowed for mandatory elements.

Furthermore, an extensive suite of metadata analysis tools provide various analyses and reports. For example, as can be seen in figure-3 below, the null value analysis tool report confirms that all mandatory elements (Title, Subject, Description, Language, Coverage, Resource Type, and Format) are populated with metadata values.

In addition to the Metadata Template Creator and Null values analysis tools, the metadata system also provides other quality assurance mechanisms. For example, all values can be listed by element/field in aggregate and visually examined for errors and inconsistencies. The viewing tools are further enhanced by the use of additional refinements such as: Highlighter (On/Off), Qualifiers (Use/Ignore) etc. Furthermore, various graphical reports can be generated as needed. These include: Records Added over Time, Records Added per Month, Files Added over Time, Clickable Map of Texas, (by Collection, by Institution), etc.

Figure 2. Embedded quality assurance in metadata creation template

The metadata template diagram in Figure-2 above shows one of the quality control tools in the UNT metadata creation workflow. It is a self-checking metadata entry template that ensures that all mandatory elements have values before the record is added to the system. In other words, no null value is allowed for mandatory elements.

Furthermore, an extensive suite of metadata analysis tools provide various analyses and reports. For example, as can be seen in figure-3 below, the null value analysis tool report confirms that all mandatory elements (Title, Subject, Description, Language, Coverage, Resource Type, and Format) are populated with metadata values.

In addition to the Metadata Template Creator and Null values analysis tools, the metadata system also provides other quality assurance mechanisms. For example, all values can be listed by element/field in aggregate and visually examined for errors and inconsistencies. The viewing tools are further enhanced by the use of additional refinements such as: Highlighter (On/Off), Qualifiers (Use/Ignore) etc. Furthermore, various graphical reports can be generated as needed. These include: Records Added over Time, Records Added per Month, Files Added over Time, Clickable Map of Texas, (by Collection, by Institution), etc.

Figure 3. NULL value for visual inspection of UNT Mandatory elements

Human created and maintained metadata is expensive. As depicted in figure 4 below, metadata records are also created by automated means, usually importing from other databases or harvested from the Web. However, fully automated maintenance and quality assurance may not be feasible due to variability of crawling technologies, and quality issues with the source data.

Figure 4. General Workflow for UNT
When we speak of metadata quality issues, in addition to the metadata structure and the creation of the content of the metadata fields, it is also important to discuss the quality of the vocabularies and taxonomies used to describe heterogeneous digital resources within metadata records.

**The UNT Libraries Controlled Vocabularies**

Successful metadata must add value that exceeds the traditional static representations. High quality metadata does not rely solely on information contained within the resource itself. The UNT Libraries have developed a system for creating and managing hierarchical controlled vocabularies for use in digital library initiatives.

Controlled vocabularies draw different terms and concepts into one single word or phrase to enhance search and navigation. These vocabularies enable data enterers to easily select appropriate values and place them in metadata records. Selecting a value from a controlled vocabulary ensures metadata consistency. Consequently, precision across all digital resources will be maintained.

**Figure 5**

*Word Cloud for UNT Libraries Subject/Keywords metadata elements*

Figure 5 above is a visual depiction of frequently used words in our subject metadata field. The word cloud simply illustrates keyword density (alphabetically) using font size. The more often a word appears on our metadata field, the larger it appears within the word cloud. This is important in identifying the subject areas that are highly represented in our collections.

Considering the diversity of participating institutions and heterogeneity of the collections, all possible digital resources may not be described adequately using pre-determined or controlled terms. To overcome the limitations and balance the issues, the UNT Libraries are implementing a hybrid system that uses both controlled terms and free keywords in order to describe all possible resources adequately. This flexible approach of pre-defined and custom-generated vocabularies provides maximum flexibility to capture complete and high-quality metadata for all types of digital resources.

**Conclusion**

Maintaining usable and sustainable digital collections requires a complex set of actions. Quality metadata is crucial to implementing reliable, usable, and sustainable digital libraries. Metadata errors, omissions and ambiguities result in problems with recall and precision and affect interoperability.

The various quality control mechanisms applied at various levels of the metadata creation workflow facilitate improvements in metadata quality and optimize quality assurance processes throughout. Considering the complexities and multifaceted issues involved in determining the level of metadata quality required by all players, the UNT Libraries’ modular approach provides opportunities for continuous refinement in accordance with both local and wider contexts.

If the digital library community is to provide optimal access to the diverse information resources available across digital libraries and repositories, all stakeholders must give high priority to the task of creating and maintaining the highest possible level of metadata quality. Indeed, creation of good quality metadata requires a community-wide modular approach. By federating and utilizing quality assurance modules, we will be able to engage in scalable collaboration with the shared vision of building interoperable, usable, and durable digital libraries.
References


Authors Biography

Daniel Gelaw Alemneh is Metadata and Documentation Librarian in the Digital Projects Unit at the University of North Texas Libraries. He currently coordinates and leads metadata related activities of the department. Daniel received his BS in LIS from Addis Ababa University, Ethiopia (1994), his MA in Library and Information Management from the University of Sheffield, UK (1997), and his Post Masters in Digital Image Management from the UNT (2000). Daniel is a doctoral candidate in information sciences and also an adjunct faculty member at the UNT School of Library and Information Sciences teaching indexing, abstracting, and information retrieval courses.

Mark Edward Phillips is Head of the Digital Projects Unit at the UNT Libraries where he oversees digital library development and all work in the digital laboratory. He received his Master’s Degree in Information Science from the UNT School of Library and Information Sciences (2004) and a Bachelor in Music Performance from Oklahoma City University (2002). Mr. Phillips has been involved in statewide and national projects including the Portal to Texas History and the “Web-at-Risk” project—an NDIIPP funded project to capture and preserve digital information. The Web-at-Risk project is a 3-year collaborative effort with the California Digital Library, and New York University.