MAINTAINING QUALITY METADATA: TOWARD EFFECTIVE DIGITAL RESOURCE LIFECYCLE MANAGEMENT

DANIEL GELAW ALEMNEH;
University of North Texas;
Digital Projects Unit;
Denton, Texas-76201, USA
E-mail: daniel.alemneh@unt.edu

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. Because poor metadata quality can result in ambiguity, poor recall and inconsistent search results, the existence of robust quality assurance mechanisms is a necessary feature of a well-functioning digital library. Recognizing the strategic benefit of quality metadata as a means of ensuring long term access to its digital resources, the University of North Texas (UNT) Libraries employ a number of metadata quality assurance procedures, tools, and associated quality assurance mechanisms. This paper discusses issues related to digital resources management and describes how UNT approaches metadata quality issues at various levels of the digital resources life cycle.

1. 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 synergies of numerous emerging trends (individualized library portals, cross-discipline collaborations, social networking and collaboration tools, etc.) provide new ways to reach out to new audiences to enhance the content providers’ abilities to play an active role in today’s Web 2.0 environment.

1.1. The UNT Libraries Metadata

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 UNT Libraries metadata element set comprises Dublin Core-based locally qualified 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 UNT Libraries metadata elements allow flexibility at the local level to integrate existing and anticipated content, processes, and systems.
Among other UNT digital libraries initiatives, The Portal to Texas History℠, for example, features digital reproductions of photographs, maps, letters, documents, books, artifacts, and more. As can be seen from Figure 1, 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.

![Files added to portal](chart.png)

**Fig. 1. Metadata Records Added to the Portal to Texas History℠ (from May 2004 to May 2008)**

2. **Metadata Quality**

Maintaining usable and sustainable digital collections necessitates maintaining high quality metadata about those digital objects. 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.

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.
Metadata quality is a crucial issue for cultural heritage communities. Accordingly, various digital repository systems like ContentDM, DSpace, Fedora, ePrints, etc. try to provide workflow models that facilitate metadata management. 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, (Bui and Park, 2006 and Hughes, 2004) and other known substantive factors. Metadata quality researchers (Barton et al. (2003), Zeng and Mai, (2006), among others, documents metadata quality in terms of:

- Error free, (such as adding/selecting wrong information in the wrong field/subfield, typographical errors):
  - Letter transposition, e.g., 0208 for 2008
  - Letter omission, e.g., Socity for Society
  - Letter insertion, e.g., asnd for and
  - Letter substitution or misstrokes, e.g. anu for any
- No omissions, (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. Errors, omissions, and ambiguities in the metadata affect the consistency of search results and high recall of available resources. Human created and maintained metadata is expensive. 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 other quality issues with the source data. Barton et al. (2003) indicate that not all problems of metadata quality can be addressed effectively by machine solutions. The next sections outline the key factors influencing metadata quality issues.

2.1. **Factors influencing metadata quality**

Quality services depend on good metadata, but most metadata values are not very good. Metadata quality is influenced by so many factors. 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. As noted by Ochoa and Duval (2007) quality is not an absolute value, but it dependent on the community of use. Based on the literature review and the University of North Texas Libraries practical experiences, the following section summarizes the major factors that affect metadata quality.
2.1.1. 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?
  - 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.]
- What are the associated digital rights issues? [For content packaging, repackaging and repurposing]

2.1.2. 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. [9] 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?

2.1.3. Availability of Tools and Support

In most collaborative projects, non-professionals or volunteers create metadata, often working in isolation without adequate training, tools, and supports. Although most metadata quality evaluation focus on overall quality of records, (Moen et al (1998) and Shreeves et al (2005), a number of researchers (including Greenberg et al (2001), and Wilson, (2007) studied metadata quality focusing on quality of non-expert metadata in exponentially growing repositories. 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]
  - 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?
2.1.4. Cost

Creating and managing high quality metadata is an expensive endeavor. 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]
  - If not, what are the priorities? [Cost/Benefit]

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

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.

3. UNT Metadata Analysis Tool

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

Based on the UNT Libraries' descriptive metadata element set, which comprises Dublin Core and locally developed metadata element, the metadata analysis tools provide various analyses and reports. The Metadata Analysis Tool was developed to facilitate the day-to-day metadata management and quality assurance of metadata created for various digital projects at the University of North Texas Libraries. It provides metadata creators, editors and managers quick access to various representations of metadata housed in several digital library systems.

The Metadata Analysis Tool provides cost effective and scalable mechanisms to detect errors and clean up values to improve the consistency and overall quality of data. The tool allows metadata editors access to aggregations of records based on specific category, including the collaborators notion of collection or institution.

The following section describes very briefly some of the tools and quality control mechanisms used at the UNT Libraries to quickly identify inconsistent metadata entry within and across collections.
3.1. **Null Value analysis**

The metadata template diagram in Figure-2 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.

![Diagram of metadata template and quality assurance](image)

**Fig. 2.** Embedded quality assurance in metadata creation template

As can be seen in Figure-3, the null value analysis tool report confirms that all UNT Libraries mandatory elements (Title, Subject, Description, Language, Coverage, Resource Type, and Format) are populated with metadata values.
In addition to the aforementioned tools, (metadata template creator, detection of null values detector), the UNT Libraries Metadata Analysis Tool may also be used in situations where there is a need to interpret disparate metadata by providing a framework for data analysis and visualization. For example, all values can be listed by element/field in aggregate and visually examined for possible errors and inconsistencies.

3.2. **Visualization tools**

As the UNT Libraries digital collections increase, opportunities for creating value-added services emerge. 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 4 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.

Figure 5 is a visual depiction of Texas counties, also known as Clickable Map of Texas for visually surfing datasets of huge and diverse digital objects. If one click on a county, s/he can see all of the materials on the Portal to Texas History about and/or from that location, and if one rolls his mouse over a county, s/he can also see its name. Such enhanced and innovative functionalities (that improve users’ experiences) depend on quality metadata.
3.3. The UNT Libraries Controlled Vocabularies

High quality metadata does not rely solely on information contained within the resource itself. 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, ranging from museum objects, photographs, letters, and newspapers to audio, video and large format items. In addition to the variety in document types and formats, there is also a wide spectrum of metadata formats used to record descriptive information about these objects. Considering the diverse requirements, the UNT Libraries developed a system for creating and managing hierarchical controlled vocabularies for use in digital library initiatives such as the Portal to Texas History.
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. As can be seen from Figure-6, the UNT Libraries Browse-Subjects (UNTL-BS) provide a broad navigational tool for browsing through digital content. On the data entry side, users of the Portal can select a top level hierarchical search term, and then drill down through subordinate subject terms to find other content within that subject category. The UNTL-BS enables data enterers to easily select appropriate values and place them in metadata records. Selecting a value from a controlled vocabulary ensures consistency and enhances precision across all digital resources.

![Fig.6 UNT Libraries Controlled Vocabularies “Browse All” Alphabetical List](image)
4. Concluding Remarks

Digital life cycle management starts from the point an item is created or selected for digitization (if not born-digital) and continues through image cleanup, metadata capture, derivative creation, and ensuring long-term access. Maintaining high quality metadata about every digital object requires a framework that provides the appropriate context needed to carry out quality assurance measures. Metadata errors, omissions and ambiguities result in problems with recall and precision and affect interoperability.

As described in this document, 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. The team approaches metadata quality issues at various levels of the digital resources life cycle. The team continually reviews and refines the metadata creation processes and makes them up-to-date and useful in light of current requirements and developments in the field. Considering the complexities and multifaceted issues involved in determining the level of metadata quality required by all players, such a modular approach facilitates the flexibility and responsiveness required in such a diverse and collaborative environment.

In conclusion, creation of good quality metadata requires a community-wide approach. By federating and utilizing available quality assurance tools, we will be able to engage in scalable collaboration with the shared vision of building interoperable, usable, sharable and durable digital libraries!

References


The Portal to Texas History™ page: http://texashistory.unt.edu/