**Data Function Areas - Metadata Management**

Chase Rawson

Department of Technology, Granite State College

DATA 520: Data Mining, Cleaning, and Visualization

Dr. Tony Sulpizio

February 1, 2023

As our world adapts to a society based on information technology, the result comes with managing big data collected through consumers via the internet. However, because big data can be hard to interpret for some organizations, there can be tens of hundreds of terabytes collected that would need to be analyzed. Metadata is data collected from the dataset created by an organization, summarizing critical facts about collected data such as file extensions, date created, or the name of the document owner classifying the type of data such as photos, audio or video files, and logs. Metadata management ensures that these organizations categorize and organize all digital data efficiently and efficiently, ensuring that data can be found easily. Data analysts might have some knowledge of a dataset, but no single analyst will know everything about it; “metadata provides the primary means of capturing and managing organizational knowledge about data.” (Henderson et. al., 2017). Without the use of metadata, “an organization does not know what data it has, what the data represents, where it originates, how it moves through systems, who has access to it, or what it means for the data to be of high quality.” (Henderson et. al., 2017). As the size of all data increases, metadata becomes more critical, and a strategy must be developed to maintain the metadata architecture.

The use of metadata in the day-to-day operations of an organization would lead to the dataset becoming more valuable. More specifically, it can identify redundant data and measure the data quality; how can we use it to the fullest extent to make strategic changes in day-to-day operations? Metadata also prohibits data that may be deemed out-of-date, which can reduce the probability of incorrect data or misusing data for other purposes from being used for analysis. This, in turn, would provide better communication between data analysts and other departments. The process of successfully implementing metadata is based on an organization’s commitment to managing metadata which could either be from management or the allocation of funds towards developing metadata. Data analysts’ would also need to develop a strategy; “how Metadata will be created, maintained, integrated, and accessed and must align with business priorities.” (Henderson et. al., 2017). Successful metadata implementation must take an enterprise perspective, have context toward business development, have standards for appropriate use, and utilize consumer feedback to remove incorrect or out-of-date data. One example of metadata being used successfully comes from the United States Geological Survey, in which standard SM 502.7 of the USGS Survey Manual requires that “metadata must accompany all USGS scientific data and other information products. Metadata records are to be developed in a standardized way that enables users to understand the context and to evaluate the usefulness of the data or information product.” (Data Management, n.d.). This policy brings forth a rule that any strategy deemed acceptable by the USGS, and the collection of data that supports the priorities of the USGS, must include the use of metadata. Now that a strategy is developed, there are different tools and techniques to manage metadata successfully.

With the use of tools, primarily metadata repositories, information can be collected to show the transformation of data. These repositories allow an analyst to view the lineage of data as it has been processed from different applications and eventually collected and stored in the dataset. As described in the Business Intelligence Roadmap, a successful metadata repository should,

Store metadata in four classifications: ownership, descriptive characteristics, rules and policies, and physical characteristics. Ownership, showing the data owner and the application owner. The descriptive characteristics, define the names, types and lengths, and definitions describing business data or business processes. Rules and policies will define security, data cleanliness, timelines for data, and relationships. Physical characteristics define the origin or source, and physical location. (Moss & Atre, 2011)

One example is ordering an item from an online e-commerce business and how different aspects of data, such as the specific item, the address of where the item should be shipped, and the quantity of an item in stock, are handled by metadata repositories (Henderson et. al., 2017). Although it may seem that there is much freedom when using metadata and the tools supporting its implementation, metadata comes with appropriate use.

When implementing or strategizing uses for metadata, each organization should, “determine their specific requirements for the management of the Metadata lifecycle and establish governance processes to enable those requirements.” (Henderson et. al., 2017). The governance surrounding metadata should be employed to share information throughout the organization based on its appropriateness to the guidelines determined. Industry-based and sector-sensitive metadata standards and guidelines must be followed about how metadata should be stored and shared between consumers. With sharing data, the National Information Standards Organization advocates for using to support the exchange of data, quoting, “XML elements can take attributes, which typically also have their own values. An XML attribute and its value refine the meaning of the element in which they appear. XML supports multilingualism of metadata by providing a predefined attribute to indicate the language in which an element’s value appears.” (Riley, 2017). With these three principles in use, we can live easier lives by understanding more about the data we’ve collected.

One example of metadata in use comes from the National Institute of Allergy and Infectious Diseases. They use clinical metadata for many different purposes, including collecting generic patient data such as birth and death dates and admission and discharge dates, “Several required fields are dates (e.g., sample collection date) which should be supplied in the international standard (ISO) format (yyyy-mm-dd). However, these dates when combined may potentially lead to patient identification. To avoid this possible breach of patient health information, the dates should be collected but censored from public access.” (U.S Department of Health and Human Services, n.d.). The NIAID uses each of the three main points described above, the goals of using metadata, the process of using metadata, and applying standards for ethical metadata usage to ensure successful organizational operations.

In general, metadata has helped many organizations organize their collected data. Finding relevant information and eliminating incorrect and out-of-date information has given value to the use of metadata. Collecting metadata using metadata repositories has also helped maintain the value of data. Furthermore, using what has been collected, metadata has helped bring forth value to the day-to-day operations of an organization for the better.

**References**

Data Management. (n.d.). *Metadata Creation* . Metadata Creation | U.S. Geological Survey. Retrieved February 3, 2023, from <https://www.usgs.gov/data-management/metadata-creation>

Henderson, D., Earley, S., Sebastian-Coleman, L., Sykora, E., & Smith, E. (2017). *Dama-Dmbok: Data Management Body of Knowledge* (2nd ed.). Technics Publications.

Moss, L. T., & Atre, S. (2011). *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-support Applications*. Addison-Wesley.

Riley, J. (2017). *Understanding metadata: What is metadata, and what is it for?* NISO Press.

U.S. Department of Health and Human Services. (n.d.). *Clinical Metadata Standard*. National Institute of Allergy and Infectious Diseases. Retrieved February 3, 2023, from <https://www.niaid.nih.gov/research/clinical-metadata-standard>