METADATA II

8.1 Databases use Metadata: Metadata is "data about data". The term is ambiguous, as it is used for two fundamentally different concepts (types). Structural metadata is about the design and specification of data structures and is more properly called "data about the containers of data"; descriptive metadata, on the other hand, is about individual instances of application data, the data content.

Metadata is traditionally in the card catalogs of libraries. As information has become increasingly digital, metadata are also used to describe digital data using metadata standards specific to a particular discipline. By describing the contents and context of data files, the usefulness of the original data/files is greatly increased. For example, a webpage may include metadata specifying what language it is written in, what tools were used to create it, and where to go for more on the subject, allowing browsers to automatically improve the experience of users. Wikipedia encourages the use of metadata by asking editors to add category names to articles, and to include information with citations such as title, source and access date.

The main purpose of metadata is to facilitate in the discovery of relevant information, more often classified as resource discovery. Metadata also helps organize electronic resources, provide digital identification, and helps support archiving and preservation of the resource. Metadata assists in resource discovery by "allowing resources to be found by relevant criteria, identifying resources, bringing similar resources together, distinguishing dissimilar resources, and giving location information."

8.2 Metadata usage

Data virtualization

Data virtualization has emerged as the new software technology to complete the virtualization stack in the enterprise. Metadata are used in data virtualization servers which are enterprise infrastructure components, alongside database and application servers. Metadata in these servers are saved as persistent repository and describe business objects in various enterprise systems and applications. Structural metadata commonality is also important to support data virtualization.

Statistics and census services

Standardization work has had a large impact on efforts to build metadata systems in the statistical community. Several metadata standards are described, and their importance to statistical agencies is discussed. Applications of the standards at the Census Bureau, Environmental Protection Agency, Bureau of Labor Statistics, Statistics Canada, and many others are described. Emphasis is on the impact a metadata registry can have in a statistical agency.

Library and information science

Libraries employ metadata in library catalogues, most commonly as part of an Integrated Library Management System. Metadata are obtained by cataloguing resources such as books, periodicals, DVDs, web pages or digital images. These data are stored in the integrated library management system, ILMS, using the MARC metadata standard. The purpose is to direct patrons to the physical or electronic location of items or areas they seek as well as to provide a description of the item/s in question.

More recent and specialized instances of library metadata include the establishment of digital libraries including e-print repositories and digital image libraries. While often based on library principles, the focus on non-librarian use, especially in providing metadata, means they do not follow traditional or common cataloging approaches. Given the custom nature of included materials, metadata fields are often specially created e.g. taxonomic classification fields, location fields, keywords or copyright statement. Standard file information such as file size and format are usually automatically included.[22]

Standardization for library operation has been a key topic in international standardization (ISO) for decades. Standards for metadata in digital libraries include Dublin Core, METS, MODS, DDI, ISO standard Digital Object Identifier (DOI), ISO standard Uniform Resource Name (URN), PREMIS schema, Ecological Metadata Language, and OAI-PMH. Leading libraries in the world give hints on their metadata standards strategies.[23][24]

8.3 Metadata and the law

United States of America

Problems involving metadata in litigation in the United States are becoming widespread. Courts have looked at various questions involving metadata, including the discoverability of metadata by parties. Although the Federal Rules of Civil Procedure have only specified rules about electronic documents, subsequent case

law has elaborated on the requirement of parties to reveal metadata. In October 2009, the Arizona Supreme Court has ruled that metadata records are public record.

Document metadata has proven particularly important in legal environments in which litigation has requested metadata, which can include sensitive information detrimental to a party in court.

Using metadata removal tools to "clean" documents can mitigate the risks of unwittingly sending sensitive data. This process partially (see Data remanence) protects law firms from potentially damaging leaking of sensitive data through electronic discovery.

Metadata in healthcare

Australian researches in medicine started a lot of metadata definition for applications in health care. That approach offers the first recognized attempt to adhere to international standards in medical sciences instead of defining a proprietary standard under the WHO umbrella first.

The medical community yet did not approve the need to follow metadata standards despite respective research.

Metadata and data warehousing

Data warehouse (DW) is a repository of an organization's electronically stored data. Data warehouses are designed to manage and store the data whereas the business intelligence (BI) focuses on the usage of the data to facilitate reporting and analysis.

The purpose of a data warehouse is to house standardized, structured, consistent, integrated, correct, cleansed and timely data, extracted from various operational systems in an organization. The extracted data are integrated in the data warehouse environment in order to provide an enterprise wide perspective, one version of the truth. Data are structured in a way to specifically address the reporting and analytic requirements. The design of structural metadata commonality using a data modeling method such as entity relationship model diagramming is very important in any data warehouse development effort.

An essential component of a data warehouse/business intelligence system is the metadata and tools to manage and retrieve the metadata. Ralph Kimball describes

metadata as the DNA of the data warehouse as metadata defines the elements of the data warehouse and how they work together.

Kimball et al. refers to three main categories of metadata: Technical metadata, business metadata and process metadata. Technical metadata are primarily definitional, while business metadata and process metadata are primarily descriptive. Keep in mind that the categories sometimes overlap.

- **Technical metadata** define the objects and processes in a DW/BI system, as seen from a technical point of view. The technical metadata include the system metadata which define the data structures such as: tables, fields, data types, indexes and partitions in the relational engine, and databases, dimensions, measures, and data mining models. Technical metadata define the data model and the way it is displayed for the users, with the reports, schedules, distribution lists and user security rights.

- **Business metadata** is a content from the data warehouse described in more user-friendly terms. The business metadata tell you what data you have, where they come from, what they mean and what their relationship is to other data in the data warehouse. Business metadata may also serve as a documentation for the DW/BI system. Users who browse the data warehouse are primarily viewing the business metadata.

- **Process metadata** is used to describe the results of various operations in the data warehouse. Within the ETL process, all key data from tasks are logged on execution. This includes start time, end time, CPU seconds used, disk reads, disk writes and rows processed. When troubleshooting the ETL or query process, this sort of data becomes valuable. Process metadata are the fact measurement when building and using a DW/BI system. Some organizations make a living out of collecting and selling this sort of data to companies - in that case the process metadata becomes the business metadata for the fact and dimension tables. Collecting process metadata is in the interest of business people who can use the data to identify the users of their products, which products they are using and what level of service they are receiving.

8.4 Metadata on the Internet

The HTML format used to define web pages allows for the inclusion of a variety of types of metadata, from basic descriptive text, dates and keywords to further advanced metadata schemes such as the Dublin Core, e-GMS, and AGLS

standards. Pages can also be geotagged with coordinates. Metadata may be included in the page's header or in a separate file. Microformats allow metadata to be added to on-page data in a way that users do not see, but computers can readily access.

Interestingly, many search engines are cautious about using metadata in their ranking algorithms due to exploitation of metadata and the practice of search engine optimization, SEO, to improve rankings. See Meta element article for further discussion. Studies show that search engines respond to web pages with metadata implementations, and Google has an announcement on its site showing the meta tags that its search engine understands. Enterprise search startup Swiftype recognizes metadata as a relevance signal that webmasters can implement for their website-specific search engine, even releasing their own extension, known as Meta Tags 2.

Metadata in the broadcast industry

In broadcast industry, metadata is linked to audio and video Broadcast media to:

- *identify* the media: clip or playlist names, duration, timecode, etc.
- *describe* the content: notes regarding the quality of video content, rating, description (for example, during a sport event, keywords like *goal*, *red card* will be associated to some clips)
- *classify* media: metadata allow to sort the media or to easily and quickly find a video content (a TV news could urgently need some archive content for a subject). For example, the BBC have a large subject classification system, Lonclass, a customized version of the more general-purpose Universal Decimal Classification.

This metadata can be linked to the video media thanks to the video servers. Most major broadcast sport events like FIFA World Cup or the Olympic Games use these metadata to distribute their video content to TV stations through keywords. It is often the host broadcaster who is in charge of organizing metadata through its *International Broadcast Centre* and its video servers. Those metadata are recorded with the images and are entered by metadata operators (*loggers*) who associate in live metadata available in *metadata grids* through software (such as Multicam(LSM) or IPDirector used during the FIFA World Cup or Olympic Games).

Geospatial metadata

Metadata that describe geographic objects (such as datasets, maps, features, or simply documents with a geospatial component) have a history dating back to at least 1994 (refer MIT Library page on FGDC Metadata). This class of metadata is described more fully on the Geospatial metadata page.

Ecological and environmental metadata

Ecological and environmental metadata are intended to document the who, what, when, where, why, and how of data collection for a particular study. Metadata should be generated in a format commonly used by the most relevant science community, such as Darwin Core, Ecological Metadata Language, or Dublin Core. Metadata editing tools exist to facilitate metadata generation (e.g. Metavist, Mercury: Metadata Search System, Morpho). Metadata should describe provenance of the data (where they originated, as well as any transformations the data underwent) and how to give credit for (cite) the data products.

Digital music

Metadata is "information about information" and it is one of the really useful features of digital audio files. When audio went from analogue to digital, it became possible to label or encode audio files with more information than could be contained in just the file name. That descriptive information is called "metadata".

Metadata can be used to name, describe, catalogue and indicate ownership or copyright for a digital audio file, and its presence makes it much easier to locate a specific audio file within a group – through use of a search engine that accesses the metadata. As different digital audio formats were developed, it was agreed that a standardized and specific location would be set aside within the digital files where this information could be stored.

As a result, almost all digital audio formats, including mp3, broadcast wav and AIFF files, have similar standardized locations that can be populated with metadata.

CDs such as recordings of music will carry a layer of metadata about the recordings such as dates, artist, genre, copyright owner, etc. The metadata, not normally displayed by CD players, can be accessed and displayed by specialized music playback and/or editing applications.

The metadata for compressed and uncompressed digital music is often encoded in the ID3 tag. Common editors such as TagLib support MP3, Ogg Vorbis, FLAC, MPC, Speex, WavPack TrueAudio, WAV, AIFF, MP4 and ASF file formats.

Cloud applications

With the availability of Cloud applications, which include those to add metadata to content, metadata is increasingly available over the Internet.

Metadata administration and management

Metadata storage

Metadata can be stored either *internally*, in the same file or structure as the data (this is also called *embedded metadata*), or *externally*, in a separate file or field from the described data. A data repository typically stores the metadata *detached* from the data, but can be designed to support embedded metadata approaches. Each option has advantages and disadvantages:

- Internal storage means metadata always travel as part of the data they describe; thus, metadata are always available with the data, and can be manipulated locally. This method creates redundancy (precluding normalization), and does not allow managing all of a system's metadata in one place. It arguably increases consistency, since the metadata is readily changed whenever the data is changed.
- External storage allows collocating metadata for all the contents, for example in a database, for more efficient searching and management. Redundancy can be avoided by normalizing the metadata's organization. In this approach, metadata can be united with the content when information is transferred, for example in Streaming media; or can be referenced (for example, as a web link) from the transferred content. On the down side, the division of the metadata from the data content, especially in standalone files that refer to their source metadata elsewhere, increases the opportunity for misalignments between the two, as changes to either may not be reflected in the other.

Metadata can be stored in either human-readable or binary form. Storing metadata in a human-readable format such as XML can be useful because users can understand and edit it without specialized tools. On the other hand, these formats are rarely optimized for storage capacity, communication time, and processing speed. A binary metadata format enables efficiency in all these respects, but

requires special libraries to convert the binary information into human-readable content.

Database management

Each relational database system has its own mechanisms for storing metadata. Examples of relational-database metadata include:

- Tables of all tables in a database, their names, sizes and number of rows in each table.
- Tables of columns in each database, what tables they are used in, and the type of data stored in each column.

In database terminology, this set of metadata is referred to as the catalog. The SQL standard specifies a uniform means to access the catalog, called the information schema, but not all databases implement it, even if they implement other aspects of the SQL standard. For an example of database-specific metadata access methods, see Oracle metadata. Programmatic access to metadata is possible using APIs such as JDBC, or SchemaCrawler