

META DATA

1. Metadata is data that describes other data. Meta is a prefix that in most information technology usages means "an underlying definition or description." Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier.

Metadata Basics

2. In simple terms, metadata is "data about data," and if managed properly, it is generated whenever data is created, acquired, added to, deleted from, or updated in any data store and data system in scope of the enterprise data architecture.

3. Metadata provides a number of very important benefits to the enterprise, including:

(a) **Consistency of definitions** Metadata contains information about data that helps reconcile the difference in terminology such as "clients" and "customers," "revenue" and "sales," etc.

(b) **Clarity of relationships** Metadata helps resolve ambiguity and inconsistencies when determining the associations between entities stored throughout data environment. For example, if a customer declares a "beneficiary" in one application, and this beneficiary is called a "participant" in another application, metadata definitions would help clarify the situation.

(c) **Clarity of data lineage** Metadata contains information about the origins of a particular data set and can be granular enough to define information at the attribute level; metadata may maintain allowed values for a data attribute, its proper format, location, owner, and steward. Operationally, metadata may maintain auditable information about users, applications, and processes that create, delete, or change data, the exact timestamp of the change, and the authorization that was used to perform these actions.

4. There are three broad categories of metadata:

(a) **Business metadata**: includes definitions of data files and attributes in business terms. It may also contain definitions of business rules that apply to these attributes, data owners and stewards, data quality metrics, and similar information that helps business users to navigate the "information ocean." Some reporting and business intelligence tools provide and maintain an internal repository of business-level metadata definitions used by these tools.

(b) **Technical metadata**: is the most common form of metadata. This type of metadata is created and used by the tools and applications that create, manage, and use data. For example, some best-in-class ETL (Extract Transform Load) tools maintain internal metadata definitions used to create ETL directives or scripts. Technical metadata is a key metadata type used to build and maintain the enterprise data environment. Technical metadata typically includes database system names, table and column names and sizes, data types and allowed values, and structural information such as primary and foreign key attributes and indices. In the case of CDI architecture, technical metadata will contain subject areas defining attribute and record location reference information.

(c) Operational metadata: contains information that is available in operational systems and run-time environments. It may contain data file size, date and time of last load, updates, and backups, names of the operational procedures and scripts that have to be used to create, update, restore, or otherwise access data, etc.

5. All these types of metadata have to be persistent and available in order to provide necessary and timely information to manage often heterogeneous and complex data environments such as those represented by various Data Hub architectures. A metadata management facility that enables collection, storage, maintenance, and dissemination of metadata information is called a metadata repository. A diagrammatic representation of how Metadata is related to its data and its categories, is shown below:

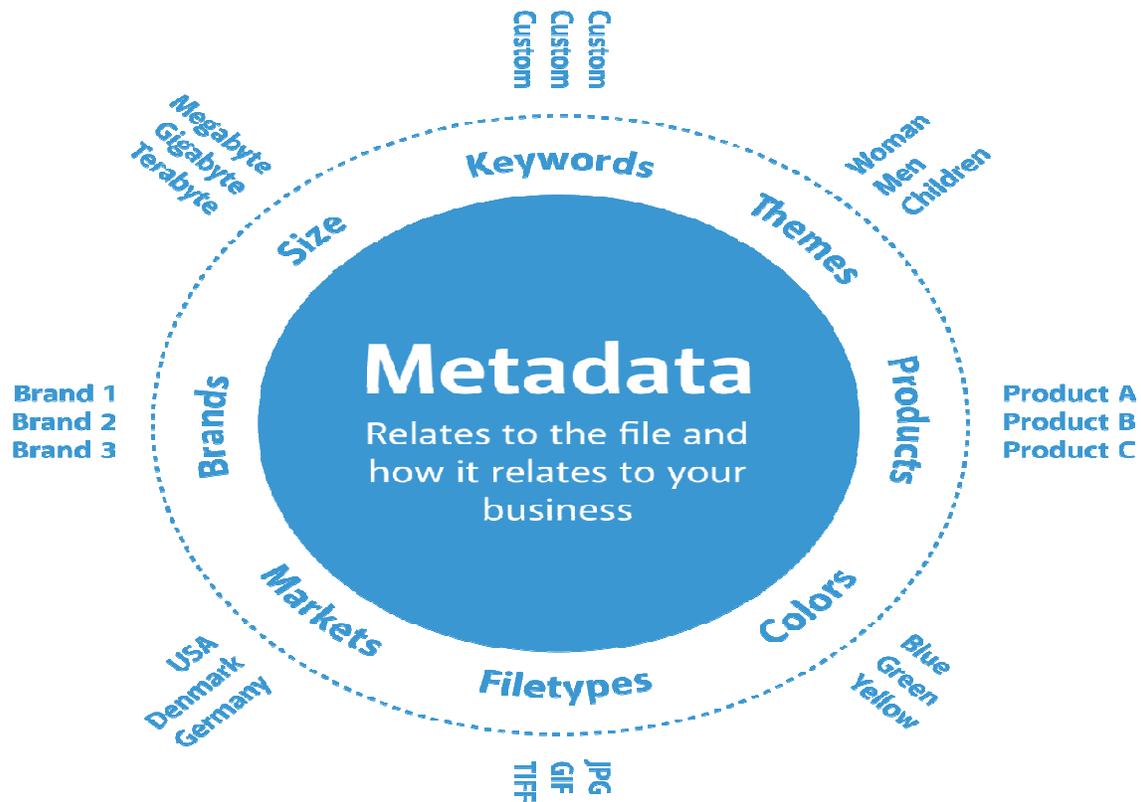


Fig.1 : Meta Data Relations With Files and Organisations

6. Topologically, metadata repository architecture defines one of the following three styles:

- (a) Centralized Metadata repository
- (b) Distributed Metadata repository

(c) Federated or Hybrid Metadata repository

7. The centralized architecture is the traditional approach to building a metadata repository. It offers efficient access to information, adaptability to additional data stores, scalability to capture additional metadata, and high performance. However, like any other centralized architecture, centralized metadata repository is a single point of failure. It requires continuous synchronization with the participants of the data environment, may become a performance bottleneck, and may negatively affect quality of metadata. Indeed, the need to copy information from various applications and data stores into the central repository may compromise data quality if the proper data validation procedures are not a part of the data acquisition process.

8. References:

(a) <http://www.merriam-webster.com/dictionary/metadata>

(b) Understanding Metadata (PDF). Bethesda, MD: NISO Press. ISBN 1-880124-62-9. Retrieved 2 April 2014.

(c) "A Guardian Guide to your Metadata". theguardian.com. Guardian News and Media Limited. 12 June 2013.