JOURNAL OF OBJECT TECHNOLOGY

Online at http://www.jot.fm. Published by ETH Zurich, Chair of Software Engineering ©**JOT, 2005**

Vol. 4, No.2, March-April 2005

# On Metadata Management Technology: Status and Issues

**Won Kim**, SamSung Electronics, Suwon, Korea

**Abstract**

Metadata captures the semantics of data in disparate data sources in an integrated enterprise information system. As such, there has long been a universal agreement on its importance. However, there are only a small number of vendors that offer metadata management systems as a separate product. In this article, I review the status of metadata management technology and vendors, and outline some of the key issues that are beyond the capabilities of metadata management products and are in the domain of consulting services.

## 1 INTRODUCTION

Metadata is loosely defined as "data about data" (i.e., descriptions of stored data). Although such a definition is not incorrect or inaccurate, it is too loose and vague when one has to organize, search, and manage metadata to support applications that drive one's business or organizational operations. Metadata management has had a long history. The first generation of metadata management system was file-based data dictionary systems. The second generation was metadata repositories based on relational database systems. There are several vendors of federated database systems, now being called enterprise information integration systems. A metadata management system is always an integral part of such systems. Today there really are very few satisfactory universal metadata management systems on the market. Enterprises that need metadata management in their information system infrastructure should adopt one of the systems on the market, and shore up the deficiencies of the system with system integration and consulting services.
In this article I discuss the following:

1. types of metadata
2. difficulties in metadata management
3. metadata management system functions and architecture
4. metadata management system vendors, and
5. metadata management issues that may require consulting services.

Cite this column as follows: Won Kim: "On Metadata Management Technology Status and Issues", in *Journal of Object Technology*, vol. 4, no. 2, March-April 2005, pp. 41-47
http://www.jot.fm/issues/issue_2005_03/column4

## 2 TYPES OF METADATA

In practice, I believe that there are several types of "data about data"; that is, it is useful and necessary to define different types of metadata.

1. **system catalogs metadata**
   Relational database systems automatically maintain a type of metadata typically named system catalogs. System catalogs are data descriptors, and include such tables as Relation Table, Column Table, Usage Table, etc. the Relation Table includes column names for each relation in the database, while the ColumnTable includes data type, length, integrity constraint (Null allowed or not; Unique or not), etc. the Usage Table includes information about when a compiled code becomes invalid and requires re-compilation.
2. **relationship metadata**
   Relationship metadata means information about the relationship between data entities (i.e., tables). Relationships include primary key-foreign key relationship between a column in one table and a column in another table; generalization/specialization relationship (i.e., IS-A relationship) between a class and its subclass in an object-oriented system or object-oriented database; aggregation relationship between an entity and its attributes; inheritance relationship between a class and its subclass in an object-oriented system or object-oriented database; and any other special semantic relationship which implies update or delete dependency.
3. **content metadata**
   Content metadata is descriptions of the contents of stored data at an arbitrary granule. Content data may be for an individual object (in the case of a textual document), a column in a table, or a table. Content metadata may be as simple as one keyword, or as complex as a business rule or a formula for computing tax or commission, or a link to an entire document. Content metadata is one of the most labor-intensive types of metadata with respect to its creation, reading, and updating. There are some products on the market, such as Interwoven's MetaCode, that use text-mining technology to automatically capture keywords or summaries of textual documents as content data about such documents.
4. **data lineage metadata**
   Data lineage metadata is lifecycle data about stored data. In particular, it includes information about the creation of data (when, who, why), subsequent updates (when, who, why), transformation, versioning, summarization, migration, and replication. It also includes transformation rules, and descriptions of migration and replication. Just as content metadata, data lineage metadata may be at an arbitrary granule. Data lineage metadata is, broadly speaking, a form of relationship metadata, since data transformation, migration, and replication imply dependency among different manifestations of the same original data. For example, when wrong data is found in one document, changes need to be made

not only to that document, but also all other documents from which the document was derived.

5. **technical metadata**
   Technical metadata is technical information about stored data. It includes such information as the format (e.g., .doc, .gif, .wav), compression or encoding algorithm used, encryption and decryption algorithm, encryption and decryption keys, software (including the release number) used to create or update the data, API used to access the data, etc.
6. **data usage metadata**
   Data usage metadata is descriptions of how and for what purposes the data is to be used by the users and applications. It is often called "business data", as the intended users are often business analysts.
7. **system metadata**
   System metadata is descriptions about the overall system environment, including hardware, operating systems, application software, etc.
8. **process metadata**
   Process metadata is descriptions of the process in which the applications operate, and any relevant outputs of each step of the process.

I note that although other authors and companies include the data usage metadata, system metadata, and process metadata as legitimate types of metadata, and I too include them herein, the case for including them as metadata is rather weak. These types of metadata are more accurately "data" rather than "metadata". Further, although other authors and companies tend to call such "data" as SQL code, design diagram, etc. as metadata, I think that these should be regarded more accurately as "data". "Legitimate" metadata is metadata that the customer needs in order to understand the semantics and lineage of stored data, and in order to properly run the applications in support of the business needs. In other words, it is not necessary to take as legitimate metadata all of the very broadly and vaguely defined "metadata" in various technical white papers or product brochures.

## 3 DIFFICULTIES IN METADATA MANAGEMENT

There are three types of difficulties in metadata management, including metadata definition and management, technology, and standards. Metadata definition and management is about defining, creating, updating, transforming, and migrating all types of metadata that are relevant and important to a user's objectives. As described in the previous section, most metadata, other than the system catalogs metadata and small parts of other types of metadata, requires diligent, timely, and disciplined manual data capture/gathering and update. Many organizations do not have the human resources or the discipline to identify, capture and manage comprehensive metadata.

Metadata management technology includes metadata design tools that allow users to model the schema of metadata across all data sources, and metadata repository systems that allow the users to extract metadata from various data sources, search and query

metadata, exchange metadata with other users, etc. I will discuss trends in metadata management technology in the next section.

Metadata standards include not only those for modeling and exchanging metadata, but also the vocabulary and knowledge ontology.

It is these difficulties that have stunted universal adoption of metadata management technologies. Most vendors of metadata management technology claim (to plan) adoption of Object Management Group's metadata modeling standards Meta Object Facility (MOF) and Common Warehouse Metamodel (CWM), and the metadata import and export standard XML Metadata Interchange (XMI). Further, there are efforts such as Dublin Core Metadata Initiative's Metadata Terms to standardize on certain metadata vocabulary. Standard knowledge ontology is also needed to organize such types of metadata as content metadata and data usage metadata.

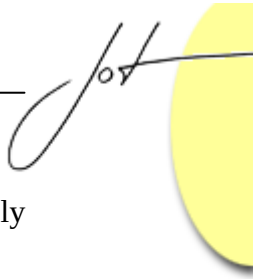
With respect to the vocabulary and knowledge ontology, where there are suitable industry standards, the standards may be adopted in full or in part. Where there is no industry standard or where the industry standard is too cumbersome or inappropriate, at least "standards" internal to an enterprise should be defined and used.

Further, appropriate procedures need to be defined and followed within the enterprise in documenting the capture, update, transformation, migration, replication of metadata and relevant transformation rules and business rules, etc.

## 4 METADATA MANAGEMENT SYSTEM FUNCTIONS AND ARCHITECTURE

The first and second generation of metadata management systems did not fare well. They did not provide adequate facilities for managing metadata and there were no standards. One of the major problems with these systems was the fact that all metadata was stored centrally, and when changes occur in data sources, the central metadata have to be updated manually. One trend in metadata management is for real-time access of distributed data sources. This means that a global metadata model is kept in a central repository, and metadata is extracted from distributed data sources on demand. This is called a federated metadata repository. The real-time extraction of metadata from a data source is performed through an adapter designed for that data source. However, a federated metadata repository can suffer in performance, since some types of metadata (e.g., data lineage metadata, technical metadata, data usage metadata, system metadata, process metadata) do not come from any data sources, and even the metadata that can be extracted from data sources (e.g., contents metadata, catalogs metadata, relationship metadata) should often be in the central repository for performance reasons. So, a hybrid approach of maintaining both the federated global metadata and some of the actual metadata in a central repository is the desired architecture.

Metadata management systems now on the market have become more powerful than the first and second-generation metadata repositories in terms of metadata management



facilities. The basic set of facilities in a metadata management system should really include

1. a metadata designer/modeler with a graphical user interface
2. a query manager (with query formulation, index creation and management facilities) and metadata and query results browser with a graphical user interface
3. security and access control (either by an access control list or group and role-based access control)
4. backup and recovery (of metadata)
5. adapters to allow extraction of data from a very wide variety of modern enterprise applications, such as ERP, CRM, SCM, and ECM systems, and a wide variety of data types, such as relational databases, indexed sequential files, legacy hierarchical databases, message middleware, HTML, XML, multimedia data, etc.
6. support for application development in Java, XML, and web services.
7. adopt such standards as XMI, MOF and CWM.

Typically, vendors of metadata management systems offer their own adapters and adapters provided by third-party adapter vendors, such as iWay. Further, the trend is to make the adapters bi-directional, in that the metadata management systems receive data (and metadata) from the data sources, and also push updated data back to the data sources.

Beyond the above "basic" facilities, metadata management systems really need to provide facilities to automatically manage impact analysis, data lineage analysis, and support for terminology and ontology standards. There is almost no metadata repository system that supports these "advanced" facilities. Data Advantage Group's MetaCenter, although lacking in some key areas, does provide some impact analysis and data lineage analysis support. Further, automatic means of capturing "contents" metadata can help the users of metadata repository systems. Interwoven's MetaCode is one example; using text-mining technology, it extracts key phrases and summaries (i.e., contents metadata) from textual documents.

As remarked earlier, much of the metadata can only be created or updated manually, and such metadata needs to augment the part of metadata that can be automatically extracted and updated via adapters to data sources. For this reason, metadata repository systems now tend to emphasize metadata extensibility, that is, provide facilities to accommodate adding new types of metadata. However, this important facility is not easy to provide. The reason is that as a new type of metadata is added, such considerations as data lineage, data dependency based on semantic relationships, vocabulary and knowledge ontology related to the new type of metadata must be accommodated in a manner that is consistent with those for the existing types of metadata.

## 5 METADATA MANAGEMENT SYSTEM VENDORS

There are at least five categories of vendors that possess metadata management technology: traditional database system vendors, enterprise application vendors, content management system vendors, metadata repository vendors, and enterprise information integration vendors. I will discuss only the latter three types of vendors below.

Vendors of metadata repository systems (along with the names of their products) include Data Advantage Group (MetaCenter), ASG/ViaSoft (Roche), Informatica (SuperGlue), Microsoft (Repository), Server Enterprise (Saphir), Computer Associates (Platinum), etc.

Enterprise Information Integration (EII) technology is basically federated database technology that integrated multiple data sources (e.g., database systems, file systems, applications that store and manage their own data in databases and files) while keeping all data in their native data sources. EII systems must create a global view of all the metadata that describe data that reside in the external data sources. The global view is the metadata that unifies all metadata. Just about every EII vendor uses XML as the data model for the global metadata, and provides an adapter to each of the external data sources. EII vendors focus on the "information integration" aspects, rather than metadata management aspects of the EII technology. There are many EII vendors. The list (along with the names of their products) includes MetaMatrix (MetaBase), IPEDO (Information Hub), XAware (XA-Suite), Actuate/Nimble (Integration Engine), Attunity, DataMirror (iFederate), TIBCO (Canon Developer), Certive, Venetica (VeniceBridge), etc.

Another category of vendors that have metadata management component in their products is ECM (enterprise content management) vendors. ECM systems include sub-products such as document management systems, website management systems, records management systems, digital asset management systems, collaboration systems, etc. Website management systems in particular need to bring together different types of data from a variety of data sources, for distribution and presentation to portals. ECM vendors include FileNet, OpenText, Interwoven, Vignette, Stellent, Documentum, etc. Vignette, in particular, creates an object-oriented global view of different data sources as metadata. However, ECM vendors do not offer metadata management component of their product suite as a separately supported product.

## 6 METADATA MANAGEMENT ISSUES THAT MAY REQUIRE CONSULTING SERVICES

Enterprise that depend on one or more complex enterprise applications that manage large volumes of complex data invariably need to manage metadata. Given today's metadata management technology, the state of metadata management standards, and the fact that much of metadata management cannot be automated means enterprises require not only

metadata repository systems but also consulting services. The following are areas (or topics) of metadata management in which consulting services need to complement metadata repository systems.

1. identification of metadata that is relevant and important to an enterprise's data management objectives (this will require interviewing business analysts and technical managers)
2. metadata design and modeling (using a particular metadata repository system)
3. definition of metadata vocabulary (this should be done in stages; further, this will require interviewing business analysts and technical managers)
4. definition of metadata knowledge ontology (this too should be done in stages; further, this will require interviewing business analysts and technical managers)
5. adapter development (using an adapter development SDK that comes with a metadata repository system)
6. determination of metadata and data prefetching (into the metadata repository) strategy.

## REFERENCES


1. Overview of Metadata Management Architecture, a technical white paper from Data Advantage Group. *(* This provides a good insight into metadata management architecture. *)*
2. Metadata As An IT Platform, a technical white paper from Data Advantage Group *(* This provides a good insight into metadata management architecture. *)*
3. Metadata Management for Data Warehousing: An Overview, Anca Vaduva and Thomas Vetterli, International Journal of Cooperative Information Systems, vol. 10, no. 3 (2001), pp. 273-298 *(* This provides a good overview of the general scope of metadata management in a large enterprise. *)*
4. The Changing Face of Repositories, Lana Gates, Application Development Trends, December 2001, pp. 25-30. (* This is a good short article about the scene of repository technology, vendors, and standards, as seen in 2001. The article is now dated, but still sheds some useful insight into metadata management. *)


## About the author

**Won Kim** is Senior Advisor at SamSung Electronics, Korea. He is Editor-in-Chief of ACM Transactions on Internet Technology (www.acm.org/toit), and Chair of ACM Special Interest Group on Knowledge Discovery and Data Mining (www.acm.org/sigkdd). He is the recipient of the ACM 2001 Distinguished Service Award. He can be reached at wonkim@austin.rr.com