



What is Metadata?

Metadata is any information that describes the data managed by an organization and the processes that affect that data. In other words metadata is organized documentation.

The process of gathering and managing Metadata can also be simply described as cataloging. A metadata repository is akin to a card catalog in a library.

Metadata not only applies to data in an organization but can also be descriptive of processes, code, configurations, and virtually any part of an Information Management System.

Metadata is used to describe and provide the knowledge necessary to manage, maintain and enhance our technology environments.

Why Metadata?

- Help new employees learn technical systems.
- Aid developers and technical people in sourcing information.
- Allow business users to navigate, find and analyze information.
- Protect the knowledge in your organization.
- Identify potential security issues.
- Reduce project analysis time.
- Determine the impact of making system changes.
- Collect metadata to assist in systems migrations.
- Use the metadata to help you perform major system cleanups.
- Manage change control for code deployment.
- Capture metadata from legacy systems having little or no documentation.
- Implement a configuration management repository.
- Automatically provide documentation to other business units.

The Business Challenge

Most data exists in many different databases, systems and even formats. For example, some data may be on the mainframe, while other information could be in SQL databases and yet other sources will be in the form of various flat file systems or event source code that affects the data, spreadsheets and desktop databases.

Business users find it difficult to query the complex data. Writing reports against the data usually requires time consuming and expensive resources. In most cases the reports do not even answer the real questions managers ask.

Managers want to see the global picture. They want to have the ability to drill down to the detail, drill up to more summarized views and pivot.



The Technology Challenge?

“We need to consolidate the islands of information in our company!”

In an effort to fulfill business user’s needs, technology departments spend countless man-hours sourcing, mapping and sifting through the reams of code and processes that load and manipulate data. They may also be involved with building an enterprise data warehouse solution or simply getting “some perspective of the data”. According to John Zachman, the world’s leading expert on Enterprise Architecture, organizations spend nearly 20-40% of their IT budgets evolving their data.

- Migrations from one system to another.
- Cooking data for reporting.
- Restructuring data to make it more useful.

Even more frightening are these statistics...

KPMG Canada Survey 1997

- Over 61% of all IT projects were deemed to have failed.

The Chaos Report

- 75% blew their schedules by 30% or more.
- 31% of all projects were cancelled before completion.
- 53% of Projects will cost over 189% of their original estimates.
- 16% of all projects are completed on time and on budget.

Robbins-Gioia Survey (2001)

- 51% of respondents viewed their large IT implementation projects as unsuccessful.

OASIG Study

- 7 out of 10 IT projects “fail” in some respect

What does this mean and what is the solution?

It appears that the number one issue is “Scope and understanding” - determining what is truly involved when embarking on any kind of technology integration project. Experience tells us that in most cases the single most difficult task facing any technology initiative is to determine the scope of the changes needed and what unforeseen challenges await us as we embark on any new project involving integration with an existing system.

Enterprise Metadata can help relieve this pain.

What is truly needed is a Repository like a card catalog in a library by which business users, project managers and technical people can look up by subject the elements of information, view descriptions and see what processes actually load and manipulate the data. Additionally, we need the ability of seeing documentation showing the interrelationships between the different parts of the systems we maintain, regardless of the complexities within.

...However, gathering Metadata can be a very daunting task.



Key requirements of a metadata solution

What to look for:

- A central, vendor neutral, repository to store the metadata.
- Powerful tools to discover as much metadata as possible automatically.
- Not be cost prohibitive.
- Provide immediate value.
- Provide a search engine.
- Ability to perform impact analysis.
- Simple automated maintenance and scalability.
- Should allow you to add descriptions to artifacts.
- Provide a historical perspective.
- Retain history and allow you to compare versions.
- Should not retain your metadata in a proprietary system, but rather make it available to you to use with any other tools you may choose to add on.

A well-designed metadata solution should show:

- What you have.
- What the meanings are.
- Where it is located.
- Show what created it.
- Show how to retrieve it.

Types of Metadata

Technical Metadata

Describes the technical systems it represents.

Such as: Data Dictionaries

Source Code

Process Flow descriptions.

Diagrams – Network Application Architectures etc

Flow charts

Element Descriptions

System Diagrams

Other related documentation

Business Metadata

Describes the perspective of the information infrastructure from a non-technical point of view.

Such as: Orders

Sales

The Data Warehouse

Web Development

Projects

Accounting

The business metadata layer is key to providing a global perspective to the metadata and allows all business users to drill through to the level of detail they require. It delivers more than just technical descriptions, but presents information that is useful to the business.



Centralized Repository

One key to a successful metadata effort is the establishment of a centralized repository and information portal.

Currently there is no real metadata repository standard in the industry. Therefore, the key is to capture and maintain the metadata in a format you know can always be used going forward.

Build Your Own

Build a Custom Metadata repository using technology and resources available to you. This requires a tremendous effort with high costs, low success rates and often is too difficult to achieve.

Typical custom metadata scenario

(Small to medium sized organization 100-500+ employees)

Sell it to the business	2 years
Design the repository	1.5 years
Implementation	8 Months

Resources required

Database administrators	130K
Architects	80K
Developers	250K
Project Managers	90K
Software	60K
Hardware	45K

Total cost of \$500K dollars or more

Things to Consider when trying to implement a custom metadata solution.

- What architecture will you use?
- How will you automate the metadata?
- What resources will you need?
- Do you have the resources and the time?
- What skills are needed?
- Do you have people with the required skills?
- How much will it cost?

Statistics prove that a homegrown metadata project is more likely to fail than to succeed.

Fundamental Design Features of a Data Warehouse

- Read Only (Never updated) snapshots in time.
- Historical.
- Summarized.
- Subject oriented.
- Consolidated and consistent. Key Technical Architecture Features.

Metadata driven

The fact that metadata drives the data warehouse is the literal truth. The metadata catalog plays such a critical role in the architecture that it makes sense to describe the architecture as metadata driven. The metadata catalog could be considered synonymous with a card catalog in a library.

The holy grail of data warehousing is to have a single source for metadata. In that case, a single change made in the metadata catalog would be reflected throughout the architecture- it would be available to all services at once. In most cases it's not practical to bring all the information to one central repository. Metadata lives in the various tools, programs, and utilities that make the data ware-



house work. We will be well served if we are aware of what metadata we keep, centralize it where possible, track it carefully, and keep it up to date.

Metadata is needed to explain and present the meaning of the data to end users through query and reporting tools, as well as the description and extraction or “Audit” of the data warehouse contents.

Caveats

Without metadata – The process simply will just not work!

Business Intelligence

Business intelligence (BI) is a broad category of application and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions. BI applications include the activities of decision support systems, query and reporting, online analytical processing, statistical analysis, forecasting, and data mining.

Business intelligence applications can be:

- Mission-critical and integral to an enterprise’s operations or singular to meet a special requirement
- Enterprise-wide or local to one division, department, or project
- Centrally initiated or driven by user demand

This term, BI, was used as early as September, 1996, when a Gartner Group report said: “By 2000, Information Democracy will emerge in forward-thinking enterprises, with Business Intelligence information and applications available broadly to employees, consultants, customers, suppliers, and the public. The key to thriving in a competitive marketplace is staying ahead of the competition. Making sound business decisions based on accurate and current information takes more than intuition. Data analysis, reporting, and query tools can help business users wade through a sea of data to synthesize. Gathering metadata is integral to a successful Business Intelligence effort.”

Knowledge Management

Knowledge management is the name of a relatively new concept in which an enterprise consciously and comprehensively gathers, organizes, shares, and analyzes its knowledge to further its aims. Until the late 1990’s, few enterprises actually had a comprehensive knowledge management practice (by any name) in operation. Instead, many companies were focusing on existing processes and striving to bring them together. Some aspects of knowledge management such as data mining and push information to users are new; others, such as data entry and OCR are very familiar. Some vendors are now offering products that address the newer ideas. Since the process is complex, involving many stages and addressing many different needs, industry experts agree that no vendor provides a comprehensive suite of products. The consensus is that an enterprise’s knowledge management plan can only be implemented with a meld of different products and cannot be accomplished without a metadata repository.

OLAP (Online Analytical Processing)

OLAP (online analytical processing) enables a user to easily and selectively extract and view data from different points-of-view. For example, a user can request that data be analyzed to display a spreadsheet showing all of a company’s products sold in California in the month of July, compare revenue figures with those for the same products in September, and then see a comparison of other product sales in California in the same time period. To facilitate this kind of analysis, OLAP data is stored in a “multidimensional” database. Whereas traditional relational database systems can be thought of as two-dimensional, a multidimensional database considers each data attribute (such as product, geographic sales region, and time period) as a separate “dimension.” OLAP software can locate the intersection of dimensions (all products sold in the Eastern region above a certain price during a certain time period) and display them. OLAP engines accomplish this by using vector arithmetic in contrast to physically reading data pages in a typical relational DBMS. Attributes such as time periods can be broken down into sub attributes. OLAP can be used for data mining or the discovery of previously



indiscernible relationships between data items.

An OLAP database does not need to be as large as a data warehouse, since not all transactional data is needed for trend analysis. OLAP Servers are typically dedicated and data is imported from existing data warehouses to create a multidimensional database for OLAP. OLAP usually sits on top of a data warehouse to provide advanced data analysis capabilities.

Online Transaction Processing (OLTP) Versus OLAP

OLTP applications are characterized by many users creating, updating, or retrieving individual records. Therefore, OLTP databases are optimized for transaction updating. Analysts and managers who frequently want a higher-level aggregated view of the data, such as total sales by product line, by region, and so forth use OLAP applications. The OLAP database is usually updated in batch, often from multiple sources, and provides a powerful analytical back-end to multiple user applications. Hence, OLAP databases are optimized for analysis.

While relational databases are good at retrieving a small number of records quickly, they are not good at retrieving a large number of records and summarizing them on the fly. Slow response time and inordinate use of system resources are common characteristics of decision support applications built exclusively on top of relational database technology. Because of the ease with which one can issue a “run-away SQL query,” many IS shops do not give users direct access to their relational databases.

Many of the problems that people attempt to solve with relational technology are actually multidimensional in nature.

For example, SQL queries to create summaries of product sales by region, region sales by product, and so on, could involve scanning most if not all the records in a marketing database and could take hours of processing. An OLAP server could handle these queries in a few seconds. OLAP applications tend to be “subject oriented,” answering such questions as “What products are selling well?” or “Where are my weakest sales offices?”

Most companies already collect and refine massive quantities of data.

OLAP Provides for the following:

- Data mining (extract hidden predictive information from large databases).
- Advanced decision support.
- Multidimensional data analysis.
- Data visualization.
- Data navigation.
- Prospective and retrospective data analysis.
- Drill down data.
- Classification and regression analysis.

Classifications of OLAP Databases

MOLAP Storage (Multidimensional OLAP Databases)

MOLAP is a high performance, multidimensional data storage format. With MOLAP, data is stored on the OLAP server. MOLAP gives the best query performance, because it is specifically optimized for multidimensional data queries. MOLAP storage is appropriate for small to medium-sized data sets where copying all of the data to the multidimensional format would not require significant loading time or utilization of large amounts of disk space.



ROLAP Storage (Relational OLAP Databases)

With ROLAP data remains in the original relational tables. A separate set of relational tables is used to store and reference aggregation data. ROLAP is ideal for large databases or legacy data that is infrequently queried.

HOLAP (Hybrid OLAP Databases)

HOLAP combines elements from MOLAP and ROLAP. HOLAP keeps the original data in relational tables, but stores aggregations in a multidimensional format. HOLAP provides connectivity to large data sets in relational tables, while taking advantage of the faster performance of the multidimensional aggregation storage.

An Overview Of Data Mining

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Conclusion

Many tools and technologies are available to solve the problem of providing better ways to store, analyze and present information. However, what will remain the number one factor in determining whether or not you are successful is “do you have a well defined layer of metadata to help describe your systems in place today”. In most cases, companies are compelled to jump into business intelligence tools, data warehouses and OLAP technologies without consideration for an abstract layer of metadata. For instance, most BI tools have very impressive graphical interfaces but don’t utilize a well defined understanding of the systems in place and therefore can offer unexpected results. Some BI tools provide repositories for metadata, but you still have to collect and build it. Without well defined metadata these BI tools become yet another island of information in the organization.

Companies spend hundreds of thousands of dollars mapping existing data and structures to new systems. In that process they will still need to go to the source and understand what is there in order to make use of it.

The power of InfoLibrarian™ Metadata Integration Framework is being used by fortune 100 companies accross many industries including finance, retail, manufacturing, high tech and health care.

InfoLibrarian Corporation provides best in class consulting services to help drive your projects to success by providing you with a proven world class solution and expertise in leveraging your metadata.