Paper 125-26

Metadata: Everyone Talks About It, But What Is It?

John E. Bentley, First Union National Bank

Abstract:

Everyone agrees that metadata is important. But then why do so many data warehouse and data mart users complain about the unavailability or completeness of their metadata? At least part of the reason is that there's no agreement on just exactly what is "metadata". The simple definition—"data about data"—is too fuzzy and in most contexts simply inadequate. Unfortunately, most contextual-specific definitions are unworkable or inappropriate for business users. This paper provides a definition that can be incorporated into a metadata solution for those who often need it most—the business users.

Disclaimer: The views and opinions expressed here are those of the author and not those of First Union National Bank. First Union National Bank does not necessarily subscribe to any philosophy, school of thought, approach, definition or process the author describes.

Why is Metadata a Problem?

Although metadata is widely acknowledged as critical for getting the most business value out of a data warehouse (in this paper the term "data warehouse" includes data marts), most data warehouse managers actually do little than talk about metadata. In a 1998 survey of 154 data warehouse managers by The Data Warehousing Institute, only 25 percent of respondents had deployed or were deploying a metadata solution. Twenty-one percent reported having developed a plan but had not yet implemented it, and 54 percent of respondents reported that they had "no plans" to implement a metadata solution.

Although these statistics are over two years old, there is no hard evidence to show that the situation has improved dramatically. Most anecdotal evidence suggests that little has changed. It still appears likely that less than half the companies that have a data warehouse are willing to invest the time, money, and resources needed to implement a comprehensive metadata management system, leaving users with only partial, *ad hoc* solutions.

Part of the reason for this is false assumption that the return on investment (ROI) for a metadata management system is difficult to quantify in terms of increasing revenues or decreasing expenses, but the costs in terms of people, software and time are clear. (See Marco, 2000.) Data warehouse projects are usually on an "aggressive" project schedule, but collecting metadata is time consuming when documentation for the legacy systems that feed the warehouse is unorganized, sketchy, or simply unavailable.

Another reason is that the mindset and priorities of both the technical staff and the business sponsor are different. Most frequently, the business side has authority over the project—they control the budget—and so the technical group has to implement the project as directed. Because of up-front cost or time constraints, the business sponsor may minimize the metadata effort even over the advice of the technical staff. As a result, metadata is assigned a "do it later" priority, but later never quite comes.

At the same time that the business sponsors may be neglecting the business metadata, an aggressive project plan forces the technical staff to concentrate on the system design, build, and initial load phases and not on maintenance and use. The IT staff then quite naturally focuses on the metadata that they need to do their job.

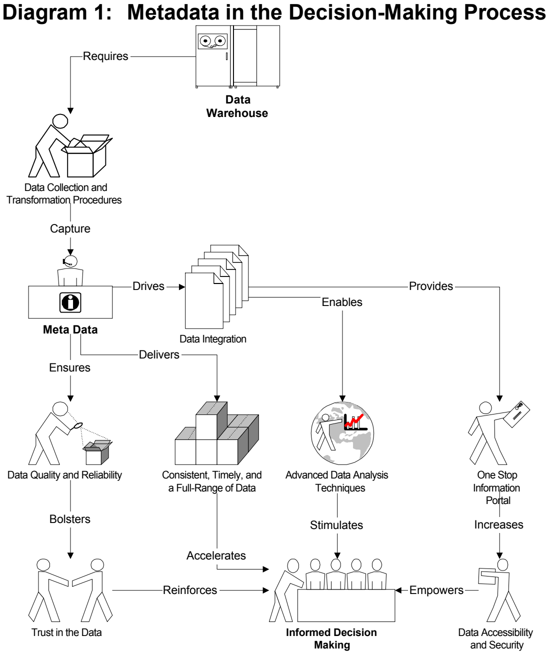
Because of these "organizational pathologies" and for other reasons, most companies poorly document the source and nature of the data they are warehousing. Even fewer produce metadata that can be leveraged to automate or improve extract-transform-load processes, data quality, or production reporting tasks.

Why is Metadata Important?

An underlying, often-ignored truth is that a data warehouse is only as good as its metadata. A data warehouse does not generate real value until it is exploited to provide information that supports business decision-making. Metadata is critical to exploitation because it tells users (and programmers!) where to find the exact data they need and helps them understand what it means. Put simply, good metadata makes it easier to use the data warehouse by allowing faster turn-around for information requests. Faster turn-around equals higher productivity, and ease-of-use gives users confidence in the information retrieved.

Business users must have confidence in the data in the warehouse and the answers it provides. Otherwise, they will be disinclined to use any data except that with which they are already confident—they will revert to using their own islands of parochial data where they know and trust its lineage. If that happens, the business value of the data warehouse is lost. The next diagram shows the impact metadata has on the decision-making process.

Diagram 1: Metadata in the Decision-Making Process


Requires
Data Warehouse
Data Collection and Transformation Procedures
Capture
Drives
Provides
Enables
Meta Data
Data Integration
Delivers
Ensures
Data Quality and Reliability
Consistent, Timely, and a Full-Range of Data
Advanced Data Analysis Techniques
One Stop Information Portal
Bolsters
Stimulates
Increases
Accelerates
Reinforces
Empowers
Trust in the Data
Informed Decision Making
Data Accessibility and Security

Adapted from Fletcher and Pinner, "Navigating the Data Warehouse Paradox Zone."

Metadata provides real value to a data warehouse and increases the ROI. David Marco, an industry-recognized expert on metadata, has developed formulas for calculating metadata ROI and estimates that good metadata can add 1% to a corporation's revenues. (Marco, 2000.) Metadata contributes measurable value by:

- Improving decision making accuracy;
- Reducing new employee training costs;
- Increasing user confidence in the data warehouse, which results in higher usage and productivity;
- Allowing sophisticated data quality processing; and
- Identifying mistakes and problems with source IT systems

An Example of Why Metadata is Important for Business Users

Consider the following scenario. Among other things, a data set extracted from a warehouse contains a variable with a business name of "Customer On-Line Access Category." The distribution is

A -	540,000
D -	161,000
I -	2,690,000
O -	1,859,000
P -	126,000
X -	18,930,000

The metadata shows that

A = Active, 3 month on-line average > 6 times
I = Inactive, not on-line in 3 months
O = Occasional, 3 month on-line average between 1 and 6 times
P = Pending, on-line account applied for but not assigned
X = Not an on-line customer

Our project is to provide contact information for Inactive customers who are most likely to move into the Occasional category. But notice that there's no listing for category D—what are we to make of that? Is D a valid category that hasn't yet made it into the metadata? Is it an old category of "Dormant" that is no longer valid and should have been recoded to "Inactive"? Does D mean "Dropped, had an on-line account but never used it"? Is it a legacy system default value?

The point is, we don't know and until we get an answer these 161,000 records—about 6 percent of the number of Inactive customers--are useless in our project. Do we ignore them and possibly not contact customers we should be contacting, include them in the selection at the risk of contacting inappropriate customers, or spend time figuring out who these customers are and why the data is like this? In effect, we have 161,000 possibly relevant customer records in the data warehouse that we can't immediately use because the metadata is incomplete. (This also points to a failure of data quality assurance because metadata isn't incorporated into the QA process, but that's a subject for another paper.)

Metadata Contexts and Views

There is a third reason for the disinclination to properly address the metadata issue, and that reason has its roots in how metadata is defined. The term "metadata" doesn't has a simple, universal meaning. "Data about data" no longer works and different functional groups within IT have applied their own definitions. As a result, "metadata" has become a technical code word that, depending on the context in which it is being used, generates conflicting definitions containing a variety of sometimes vague and often misleading messages that are hard for non-technical managers to decipher. As Table 1 shows, "data about data" has evolved into a number of things, depending on its IT context.

Table 1: Metadata in Context

Context	Description
Data Administration	Properly documented logical and physical data models and entity-relationship diagrams for both source and target systems. Usually controlled by the IT staff and usually a high priority.
Data Warehouse Back-End	Documentation tracking the extract-clean-transform-load process. Source system, staging area, and data warehouse data structures, copybook names, column mapping translation tables, and other ETL documentation. Controlled by IT staff and receives a high priority.
Application Development	Documentation about processes that access data via an application. Presented as process models or decomposition diagrams. Also includes pseudo-code and the final program code with internal documentation. Due to time and resource constraints, typically is ignored or given low priority.
Data Warehouse Front-End	Documents the meaning of data for the benefit of those running queries against the warehouse and interpreting the results. Imperative to ensuring an accurate, consistent interpretation of the data, but often is assigned low priority.

In the list of contexts presented above, the data warehouse front-end context is clearly the one most referenced by and important to business users because it helps understand the data. In an article about optimizing data warehouse usage, one data warehouse guru used an analogy about pioneers exploring the Wild West (Devlin 1988). Like the Wild West, a data warehouse is a vast territory. Without maps, the Western pioneers were often. Likewise, data warehouse users will be lost without the map that metadata provides.

The same data warehouse expert categorizes metadata as build-time, usage, and control, and suggests two "views" of metadata: builder and end-user.

Table 2. Metadata Views

View	Description
Builder	A blueprint. Key component is the enterprise data model. The ultimate definition of what the warehouse is.
End-User	A flexible, easy to use "route map". Defining feature is that the warehouse contents are presented in a business context.

So, What is Metadata for Business Users?

Some definitions of metadata appropriate for business users do exist, but they range from broad to narrow in their scope.

- “Metadata is high-level data that describes low-level data. ... [It] maps the data to business concepts that are familiar and useful to end-users.” (Korzybski, March 1996.)
- “In the context of data warehousing, the term refers to anything that defines a data warehouse object, such as a table, query, report, business rule, or transformation algorithm. … It also supplies a blueprint that shows how one type of information is derived from another.” (Gardner, November 1997.)
- “Metadata describes the information in the data warehouse: what is means, where it came from, how it was calculated, when it was loaded, who owns it.” (Ekerson, March 2000.)
- “Metadata is the definitions, sources, rules, and thresholds used to constrain the business data you are collecting, validating, transforming, reconciling, loading, and reporting.” (Fletcher and Pinner, March 2000.)

Metadata for Business Users

Technical metadata is used by IT professionals in the planning, design, creation, and maintenance of the data warehouse. This is the Data Administration, DW Back-End, and Application Development contexts. But as is pointed out in a SUGI25 paper, “[b]usiness users require more descriptive information, which will assist in translating codified information into the business concepts relevant to their domain. This would include the content and purpose of the data, related business rules, ownership and administration, and location.” (Stevens, 2000)

With this in mind, here is a clear, concise definition of “Business Metadata” based on the Data Warehouse Front-End context description:

> Business metadata shows non-technical users where to find information in the data warehouse, where it came from and how it got there, describes its quality, and provides assistance on how to interpret it.

Technical (back-end) metadata and business (front-end) metadata are mutually reinforcing. This definition is purposely broad enough to include technical metadata because many users want or need a deeper understanding of the origin and evolution of the data.

Implicit in this definition is the assumption that metadata will be dynamic, thereby helping ensure overall data quality and reliability which will, in turn, bolster users trust in the data. Dynamic metadata provides the ability to review the lineage of the data. Decision-makers will know where the data came from, how it was transformed, and what it really means.

What does Metadata Contain?

A complete metadata solution requires a lot of information. Ralph Kimball bases his categorization of metadata on the intended user base and distinguishes between “back-room metadata” that guides the extraction, cleaning, and loading processes and “front-room metadata” needed by query tools and report writing. As mentioned earlier, although there is a lot of crossover between back-end and front-end metadata, it’s the front-room metadata that makes the data in the warehouse really meaningful to business users. The next table shows some specific examples of metadata grouped into Kimball’s categories.

Table 3: Examples of Metadata

Category	Example
Back-End	✓ Ownership descriptions of each source schemas ✓ Source file layouts and target schema designs ✓ Definitions and characteristics of tables and columns ✓ Primary/foreign key assignment scheme and relationships ✓ Database partition and disk striping specifications ✓ Index and view definitions and specifications ✓ Mainframe or source system job specifications ✓ File/copy book descriptions and specifications ✓ COBOL/JCL, C or Basic code to implement extraction ✓ Update frequencies of the original sources ✓ Job specifications for joining sources, stripping out fields, and looking up attributes. ✓ Data cleaning, enhancement, and transformation rules, specifications, and mappings ✓ Data audit records and transformation run-time logs ✓ Access methods, access rights, privileges and passwords for source access
Front-End	✓ Process flows, e.g., BPwin ✓ Presentation graphics, e.g., PowerPoint ✓ Flowcharts and program code for accessing source system data ✓ Ownership and Business descriptions of the source systems ✓ Ownership and Business name of data elements ✓ Business-rule based definitions of data elements ✓ Descriptions of the valid values in categorical fields ✓ Descriptions and flowcharts of aggregation and transformation processes ✓ Physical data models ✓ Table join guidance, including cautions and restrictions ✓ Validation statistics for quality control ✓ Legal limitations on usage ✓ User login profiles and security/access controls

With these examples in mind and considering the different contexts and views of metadata and the back-end/front-end distinction, here is a list of some metadata items that are essential for business users:

Table 4. Critical Business Metadata

All variables

- Variable name
- Variable business name
- Variable definition (short)
- Variable description (long)
- Data set name
- Data set business name
- Data set description
- Legacy system and Quality Assurance contacts
- Update frequency
- Date of last update
- Special missing values
- List of variable names used if the variable is a created or calculated variable
- Business logic, algorithms, and pseudo-code used in cleaning, transforming, creating, summarizing, or calculating the variable
- Special cautions, legal limits, tips and clues on usage

Categorical variables

- List of valid values and their definitions
- Frequency distribution including number of missing values

Interval variables

- Formula used in calculating the variable
- Descriptive statistics including the number of records, mean, standard deviation, median, number of missing, and range.

Creating and maintaining metadata will require both up-front and on-going investments of time and resources. For the most part, though, the metadata will be static or slowly changing and it's creation can be automated, so the resources needed for maintenance will comparatively small. Calculating or generating the specific items needed for categorical and interval variables should be included as part of the warehouse load process. For this, when a very large number of records are loaded a weighted sample may suffice, depending on the accuracy needs of the users.

Metadata updating processes can easily be incorporated into an automated or metadata-driven quality assurance process. For example, PROC COMPARE makes it easy to compare the current month's list of valid values for a categorical variable with last month's list as a quick check for new categories. Likewise, comparing the current month's summary descriptive statistics such as mean, standard distribution and percentiles to last month's will quickly identify anomalies in the data.

Competing Metadata Standards and the SAS Metadata Architecture

Until September 2000, companies working to implement strong metadata management processes and procedures were handicapped by competing metadata standards within the IT industry. There are two competing groups, both spearheaded by industry heavyweights. These groups are The Meta Data Coalition (MDC) and The Object Management Group (OMG). Competition going forward, however, should be minimized because in September 2000 the two groups agreed to merge, retaining the OMG name, and work together to develop a single best-of-breed standard built around the OMG's Common Warehouse Metamodel (CWM).

The SAS Institute is a member of the 53-member Metadata Coalition, as is Microsoft. The MDC sponsors the Open Information Model (OIM) as a comprehensive source of standards and specifications for business engineering, knowledge management, and databases in addition to data warehousing. The OIM schema consists of standard object types and relationships described in Unified Modeling Language. The OIM is vendor-neutral and uses SQL as a query language and Extensible Markup Language (XML) as an interchange format between data repositories. In October 2000 the MDC endorsed the Microsoft Repository as it's preferred metadata storage database.

Major members of the Object Management Group are IBM and Oracle, among others. The OMG provids a standard called the Common Warehouse Metamodel (CWM) to enhance metadata sharing and interoperability in data warehousing environments. The CWM is based on the OMG's UML™ (Unified Modeling Language), XML™ (XML Metadata Interchange) and MOF™ (Meta Object Facility) and incorporates the Common Object Request Broker Architecture (CORBA) as the basis for interoperability and application integration.

The merger of MDC into the OMG marks a major agreement of data warehousing and metadata vendors to converge on one standard, incorporating the best of the MDC's Open Information Model with the OMG's Common Warehouse Metamodel. When the work is complete, the resulting specification will be issued by the OMG as the next version of the CWM. A single standard will allow users to freely exchange metadata between different products from different vendors.

SAS Software uses a layered metadata architectural approach that maximizes flexibility. Four distinct functional layers combine to minimize the need to continually develop and revise the metadata architecture to account for changes in metadata sources, application needs, and hardware platforms.

Table 5. SAS Metadata Architecture

Layer	Description
Facility	The Common Metadata Facility (CMF) provides tactical control. It controls the creation, deletion, update, and persistence of metadata objects.
Model	The Common Metadata Model (CMM) ensures that applications that access the metadata all interpret it the same way. It defines the objects and relationships.
API	The Application Program Interface (API) is the presentation layer and, in Version 8, may be written in either C or SCL.
Repository	The Repository stores the metadata. Users can choose from many different physical formats, and it can reside on almost any platform.

Adopted from Vernee Stevens, "SAS Metadata Architecture and Current Industry Metadata Trends."

Summary

The value of a data warehouse doesn't come from having a lot of data in one place. A data warehouse by itself has only potential value. The return on investment is realized when the data is converted to information and then used to make decisions that solve problems. Without comprehensive metadata, though, much of the data in a warehouse may never become information; at best, converting it to information will take much longer and even then decision-makers may have doubts about its accuracy.

Unfortunately, metadata can be costly to compile and make available to users. Metadata experts, however, have developed guidelines for calculating metadata ROI, showing that business and technical metadata can both increase revenues and reduce expenses. For business users, metadata improves the accessibility, quality, credibility, and usability of a data warehouse in multiple ways:

- By documenting flows of data used to populate the warehouse;
- By documenting the creation, calculation, and summarization processes;
- By providing dynamic quality-control metrics;
- By allowing the data warehouse to be navigated using meaningful business terms; and
- By allowing the use of advanced data analysis techniques needed for data mining.

The need for a complete metadata solution becomes increasingly apparent as organizations begin to deploy second and third generation decision support databases that propagate data from the data warehouse into specialized data marts. As the lineage of data increases and the numbers of users grow, the need (and demand) metadata multiplies.

The metadata needs of business users are different from the technical staff, but there are numerous overlaps. A common set of shared metadata that includes the critical items in Table 4 will benefit both groups. For business users though, the basic metadata they need is that which shows them where to find information, where it came from and how it got there, describes its quality, and provides assistance on how to interpret it.

References and Resources

Devlin, Barry (1998) "Metadata: The Warehouse Atlas." DB2 Magazine Online, Spring. www.db2mag.com/98spWare.htm

Ekerson, Wayne W. (2000) "Ignore Meta Data Strategy at Your Peril." Application Development Trends, March.

Fletcher, Tom and Jeff Pinner (2000) "Navigating the Data Warehousing Paradoz Zone." dmDirect, March 3. www.dmreview.com

The Hurwitz Group (1988) Enterprise Metadata Management, December. www.hurwitz.com

Kimball, Ralph (1998) "Meta Meta Data Data." DBMS Magazine, March. www.dbmsmag.com/9803d05.html

Korzybski, Alfred (1996). "What is Metadata?" Data Warehousing Tools Bulletin, March 1. www.computerwire.com/bulletinsuk/212e_1a6.htm

Marco, David (2000) Meta Data ROI 3-part series. DM Review, September, October, November.

Marco, David (2000) Building and Managing the Meta Data Repository: A Full Life-Cycle Guide. New York: Wiley.

Meyers, Rachel (1998) Metadata series. The Data Warehousing Career Newsletter, July.
www.softwarejobs.com/dataware/7-10-98.html
www.softwarejobs.com/dataware/7-17-98.html
www.softwarejobs.com/dataware/7-31-98.html

Stevens, Vernee (2000). "SAS Metadata Architecture and Current Industry Metadata Trends." SUGI25 Proceedings.

Wiener, Jerry (2000) "Meta Data in Context." dmDirect, February 11. www.dmreview.com

Contact Information

John E. Bentley 704-383-2686
First Union National Bank John.Bentley2@FirstUnion.Com
201 S. College Street
Mailcode NC-1025
Charlotte NC 28288

Certified Professional V6, Advanced

About the Author

John Bentley has used SAS Software for fourteen years in the healthcare, insurance, and banking industries. For the past three years he has been with the Enterprise Information Group of First Union National Bank with responsibilities of supporting users of First Union's data warehouse and data marts and managing the development of SAS client-server applications to extract, manipulate, and present information from it. John is regularly presents at national, regional, and local SAS User Group Conferences and is on the Executive Committee of the Data Mining SAS User Group.