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 quality 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 no longer adequate. 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 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 or data mart (in this paper the term "data warehouse" includes data marts), few administrators or managers actually do more than talk about the issue. In a 1998 survey of 154 data warehouse managers by The Data Warehousing Institute, only 25 percent of respondents have deployed or are deploying a metadata strategy. Twenty-one percent reported having developed a strategy but not yet implementing it and 54 percent of respondents reported that they have "no plans" to even develop a metadata strategy.

Even though these statistics are two years old and we can assume that the situation has improved at least somewhat, it appears that less than half the companies involved in data warehousing are willing to invest the time, money, and resources required to implement a robust metadata management system. In fact, more than half the respondents appear unwilling to invest in *any* formal system.

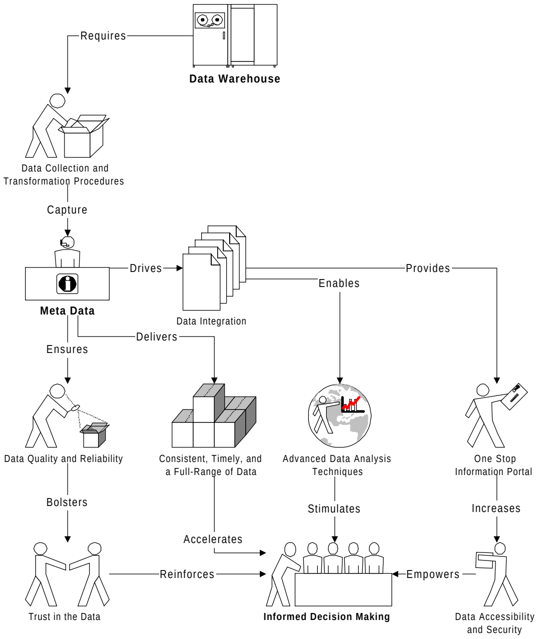
Part of the problem is that metadata's benefits are hidden and difficult to quantify even though the costs in terms of people, software and time are clear. Furthermore, designing and implementing a data warehouse is too often managed and driven by IT staff and not business users, so the focus may be on building and loading the warehouse and not maintaining and using it. As a result, most companies poorly document the source and nature of the data they are warehousing. Very few leverage this information to automate extract-transform-load processes or production reporting tasks.

Why is Metadata Important?

An underlying truth is that a data warehouse is only as good as its metadata. A warehouse is generates value when it is exploited to provide information and support business decision-making. Metadata tells users (and programmers!) where the data they need resides and helps them understand what it means. Good metadata makes it easier to use the data warehouse by allowing faster turn-around for information requests. Ease of use gives users confidence in the contents and 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


Data Warehouse
Requires
Data Collection and Transformation Procedures
Capture
Meta Data
Drives
Data Integration
Enables
Provides
Ensures
Delivers
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."

An Example of Why Metadata is Important for Business Users

Metadata is fundamental to getting business value out of a data warehouse. Consider the following scenario. Among other things, a data set in a warehouse contains a variable with a business name of "Customer On-Line Access Category." The distribution is

A -	540,000
I -	2,690,000
O -	1,859,000
P -	126,000
N -	161,000
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

The project is to do an incentive mailing to the Inactive customers who are most likely to move into the Occasional category. But notice that there's no listing for category N—what are users to make of that? Is N a valid category that hasn't yet made it into the metadata? Is it an old category that is no longer valid and should have been recoded? Does N mean "Never, has an on-line account but never used it"? Is it a legacy system default value?

The point is, we don't know and, as a result, these 161,000 records—about 6 percent of the number of Inactive customers--are useless in our project. Do we ignore these customers, include them in the scoring at the risk of mailing to inappropriate customers, or do we spend time figuring out why the data is like this and who these customers are? In effect, we have 161,000 possibly relevant customer records in the data warehouse that are of no value for our project because the metadata is incomplete and/or data quality assurance may have failed because the metadata isn't incorporated into the QA process.

Metadata Contexts and Views

There is another cause for the disinclination to properly address the metadata issue, and that cause has its roots in the definition. The term "metadata" no longer has a simple 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 the table below shows, "data about data" has evolved into a number of things.

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 end users 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. 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 lost and ended up where they didn't want to be. Likewise, data warehouse business users will be lost without the map that metadata provides.

The same data warehouse guru 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, I suggest a definition of "Business Metadata" based on the Data Warehouse Front-End context description:

Metadata shows non-technical users where to find information, 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 the users trust in the data. Dynamic metadata provides business users with the ability to review the lineage of the data they are using to make decisions. They 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 audience 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, this author suggests that the following metadata items are essential metadata 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 contact
- Quality Assurance contact
- 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.

Clearly, maintaining metadata will require an up-front and on-going investment of time and resources. For the most part, though, the metadata will be static or slowly changing. For the specific items neded for categorical and interval variables, their calculation should be included as part of the warehouse load or quality assurance process. Where a very large number of records are loaded, a clearly documented weighted sample may suffice depending on the accuracy needs of the users.

Metadata updating processes can easily be incorporated into an automated quality assurance process. For example, PROC COMPARE makes it easy to compare the current month's list of valid values for a categorical with last month's list as a quick check for new categories. Likewise, comparing the current month's 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

Companies working to implement strong metadata management processes and procedures are 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).

The SAS Institute is a member of The 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.

Major members of The Object Management Group are IBM and Oracle, among others. The OMG offers a standard called the Common Warehouse Metamodel (CWM) to enhance metadata sharing and interoperability in data warehousing environments. The standard complies with the OMG's Meta Object Facility (MOF) for defining meta models, uses the Unified Modeling Language, and incorporates the Common Object Request Broker Architecture (CORBA) as the basis for interoperability and application integration.

Although some members of each group are extreme competitors, others (such as the SAS Institute) have realized that the IT industry as a whole will be best served by some cooperation and not total confrontation, and some official cooperation does in fact exist. Some members of both groups are working to achieve a level of integration between the OIM and CWM models so that the same set of interfaces and interchange formats could be used regardless of the data repository.

SAS Institute has implemented a layered metadata architecture 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. Until it's exploited, a data warehouse has only intrinsic value. The real value comes when that data is converted to information and then used to make decisions that solve problems. Without metadata, much of the data in a warehouse can be useless as information.

For business users, metadata helps actualize a warehouse's intrinsic value by improving the accessibility, quality, credibility, and usability of the data

- 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 even more apparent as organizations begin to deploy second and third generation decision support databases that propagate data from the data warehouse into a specialized data mart. As the lineage of data increases and the number of users grows, the need and demand metadata multiplies

The metadata needs of business users are different from the technical staff, yet 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

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
First Union National Bank
201 S. College Street, 5th floor
Mailcode NC-1025
Charlotte, NC 28288
704-383-2686
John.Bentley2@FirstUnion.Com

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 an Officer with the Enterprise Information Group at First Union National Bank with responsibilities for the development of SAS client-server applications to extract, manipulate, and present data from data warehouses and data marts. John is a SAS Certified Professional in Data Management V6, regularly presents at SAS User Group Conferences, and is on the Executive Committee of the Data Mining SAS User Group.