BILLINMON.COM

TOWARD A UNIFIED THEORY OF METADATA

by

W. H. Inmon

March 2002

www.billinmon.com

Table of Contents

There is a simple definition for metadata: *"metadata is data about data".*

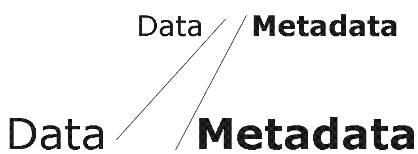
The problem is, this definition of metadata is so simplistic and general that it is almost worthless. If the definition is the starting point, then it is no wonder that metadata has become the wasteland of information processing. No other aspect of information is so important and holds so much promise as metadata. Yet for the most part, the reality is, it has been an industrial monument to underachievement. Only a few companies have ever introduced a successful and sustainable metadata product.

NO THEORETICAL FOUNDATION

One of the reasons for the travails of metadata is that there is no theoretical understanding of it; metadata is complex, ghastly complex. Another reason is that other arenas of computing have always had more glamour and more marketplace appeal than metadata. Many corporations see minimal value in investing in a metadata infrastructure. It is only over time and as the information processing environment ages and grows that the investment value in metadata becomes obvious.

This white paper will address some metadata complexities and will put a theoretical underpinning for some of those more important aspects of metadata.

In many regards the road for metadata and its underlying infrastructure has been the same as the road for data. Figure 1 depicts the parallelism of those two roads.



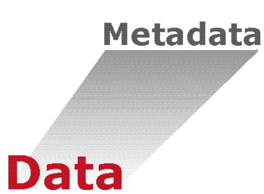
In A Number Of Aspects,
Data And Metadata Have Followed Parallel Paths

Figure 1

While data as a subject and a discipline is far advanced over that of metadata, there nevertheless is a similar treatment in terms of technology, market place awareness, and emphasis on data and metadata.

METADATA – THE SHADOW OF DATA

Metadata is the shadow for data. Figure 2: The classical positioning of data and metadata. Where data goes, so goes metadata. Data blazes the trail and metadata follows in lock-step with the data in many respects.

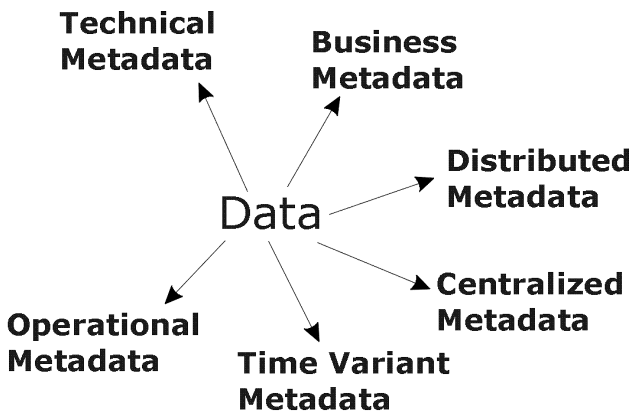


Metadata, In Many Regards,
Is The "Shadow" Of Data

Figure 2

MANY FORMS OF METADATA

One of the most confusing aspects of metadata is that it's found in many forms and in many places. Figure 3 suggests the complexities and varieties of metadata.



There Are Many Forms Of Metadata

Figure 3

It is because of these many manifestations of metadata that a unified theory has not been attempted. In fact, there are many more than those suggested in Figure 3. In some cases, there is one form of metadata but not another. In other cases, one form of metadata is present under one set of circumstances, but at a later point in time is not present. Some times, two forms of metadata will apply to the same data. It is no wonder that the complications of metadata make any unified theory very difficult.

METADATA IS NOT FLAT

When different units of data are all considered to be equivalent, the data can be said to be "flat". A good starting point for understanding metadata is the notion that it is not flat. One unit of metadata is not the same as, nor equivalent to, another unit. There is texture among different units of metadata, as suggested in Figure 4.

Figure 4

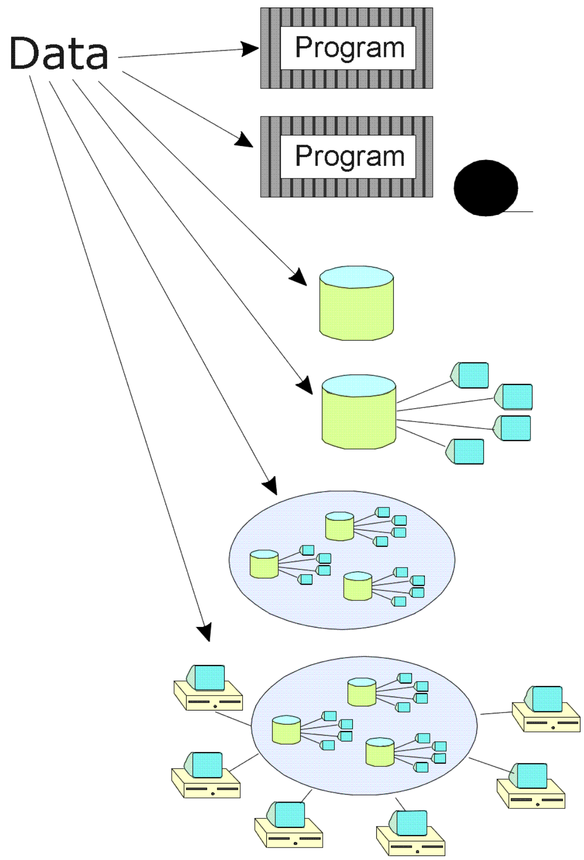
Here metadata is shown to have texture and depth. Not all metadata is the same and not all metadata is equal. Because of the types of metadata and their complex interrelationships, the texture and depth is not at all obvious.

In order to understand what is meant by metadata having texture, consider the parallel alter ego of metadata – data. Once upon a time, data was considered to be flat; however, today we recognize the many different types of data and their positioning in the world.

THE PROGRESSION OF DATA

There was a day when data was not considered to be especially important or interesting. As time has passed and the industry of information processing has matured, data has taken on an increasingly important role. This understanding did not happen over night.

Consider the sequence of events and milestones illustrated in Figure 5.



The Path That Data Has Walked

Figure 5

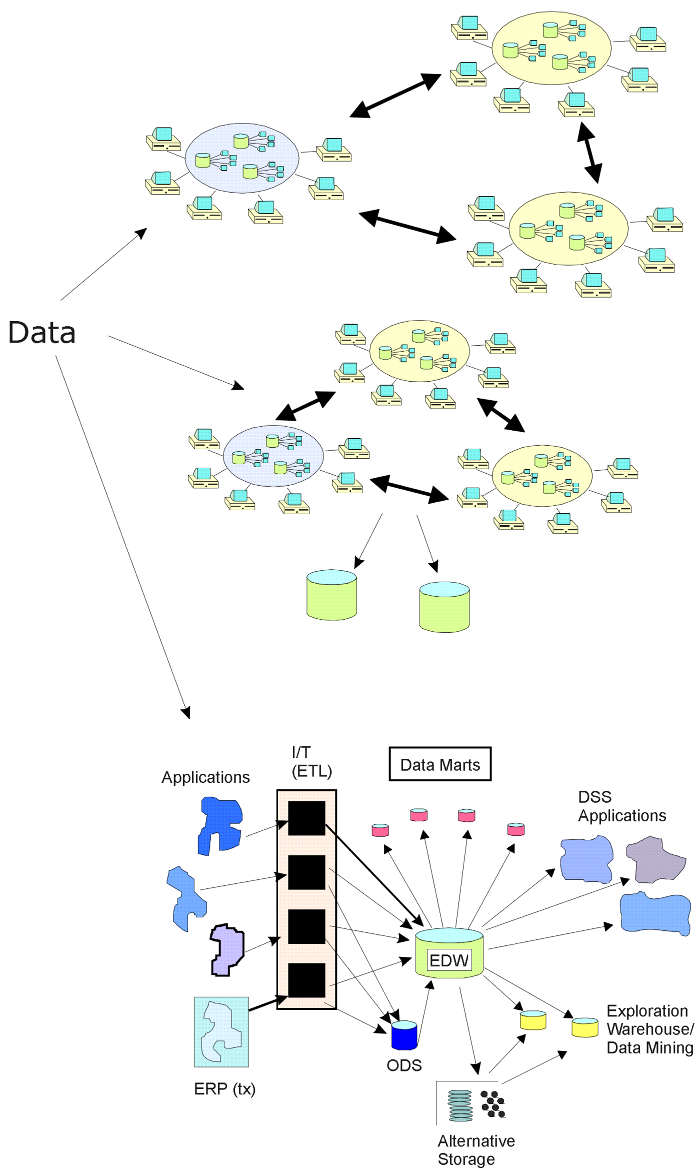
In the beginning were programs. In the very early days the algorithms and the logic that were at the heart of programs consumed the attention of the information analyst. Before long it was recognized that programs needed data on which to operate. Very few programs were self-generating when it came to data.

Data was thought of as an adjunct to processing, when it was thought of at all. Data was stored on sequential media that had many limitations; the most onerous was the necessity to access 100% of the unit of data when only a small fraction was needed.

Soon there came disk storage and data base management systems. With disk storage data could be accessed directly. Then online transaction processing became a possibility. With the advent of data base design, data was beginning to be treated as its own subject.

Next came applications. Online processing was so powerful that it found its way into many places in the corporation. In a short time, there was a desire to access data that was contained inside the online applications. Access of data by the personal computer became the norm.

Figure 6 continues the progression of data.



The Path That Data Has Walked (continued)

Figure 6

In Figure 6 many online applications started to appear, all of which had data being spun off onto personal computers. The result was called the “spider’s web” environment. There are many problems with this, the chief issues being the lack of data integrity and the environment impermeability. The frustrations with the spider’s web environment gave way to the concept of the data warehouse. Almost immediately upon arrival, the data warehouse began to cast off different related architectural entities (or components). These components formed a framework called the Corporate Information Factory.

The emphasis on the role of data and data management has changed drastically. In the early days data was thought to be a byproduct of processing; now it’s to the point where the focus of the environment is data. Data becomes the medium of exchange between different architectural components.

Data’s role changed from being unnoticed, to being considered to be “flat”, to a central role in a large and sophisticated framework. Metadata is undergoing a similar progression, although it lags considerably behind data.

A FRAMEWORK

Today metadata is either unnoticed or is flat. Neither role is appropriate. In order to bring out its texture, it is necessary to create a framework just for metadata. In many ways, the framework is a taxonomy of the most important aspects of metadata.

As with any framework, there are parameters which form the basic structure.

DIRECT DESCRIPTIVE METADATA

The first of those parameters is the notion that some metadata directly describes the data that it shadows. Figure 7 shows this type of metadata.

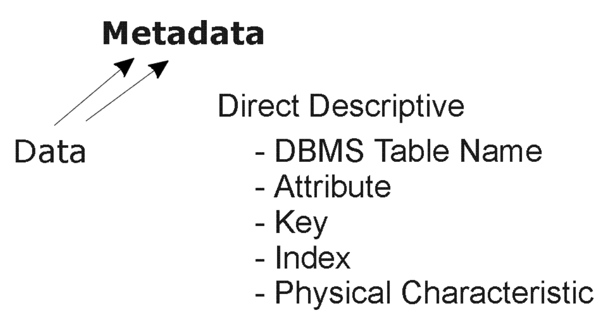


Figure 7

This direct descriptive type of metadata is the simplest form of metadata. There are many forms including:

- Table name
- Attribute name
- Key
- Index
- Physical characteristic, and so forth.

Direct descriptive data is the most straightforward kind of metadata.

INDIRECT DESCRIPTIVE METADATA

The second parameter for the framework of metadata is indirect descriptive metadata. Figure 8 suggests some forms:

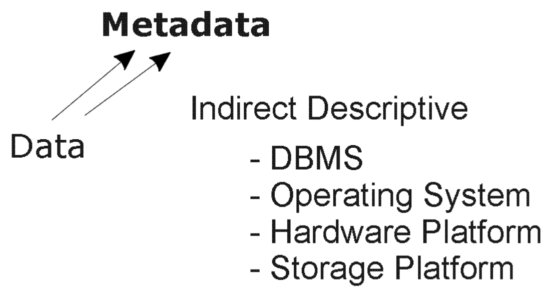


Figure 8

Indirect descriptions include:

- The DBMS under which the data operates
- The operating system under which the data operates
- The hardware platform on which the data operates
- The storage platform on which the data operates, and so forth.

Indirect descriptive metadata are descriptions which apply to data.

DIRECT RELATIONSHIP METADATA

The second major classification of metadata which belongs in the framework is data relationships. Perhaps the most important metadata is metadata about data relationships. As in the case of descriptive metadata, relationship metadata is divided into two subclasses – direct relationships and indirect relationships.

Figure 9 suggests some direct data relationships that might be described by metadata.

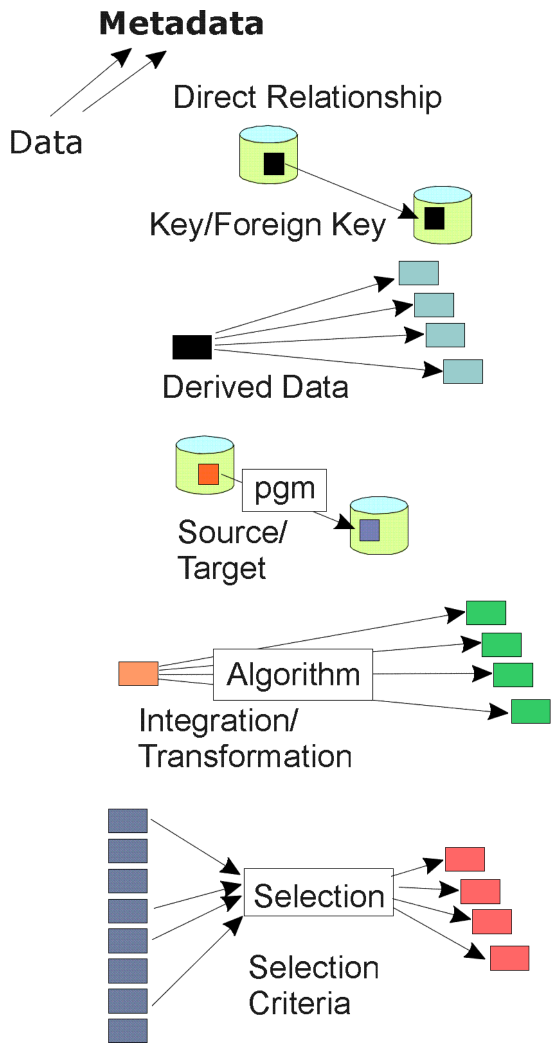


Figure 9

Examples of direct data relationships include:

- Key/foreign key relationships
- Derived data relationships
- Source/target relationships
- Integration/transformation relationships
- Selection criteria, and so forth.

Just as data has relationships, so metadata has its own relationships which shadow and mimic those of the data that it relates to.

INDIRECT RELATIONSHIP METADATA

The second kind of metadata is indirect relationship metadata. Some simple examples of indirect relationship metadata are shown in Figure 10.

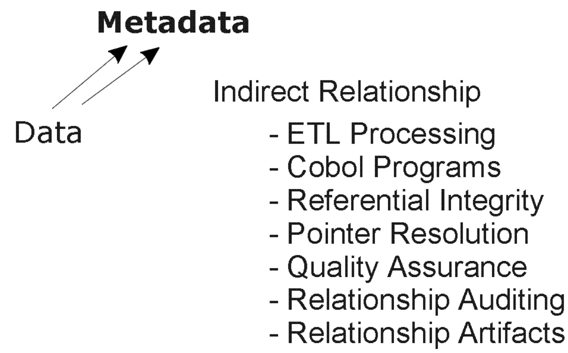


Figure 10

There can be indirect relationship metadata, such as:

- ETL processing description, including ETL tool
- COBOL programs
- Referential integrity specifications
- Pointer resolution utilities
- Quality assurance procedures
- Relationship auditing utilities
- Relationship artifacts, and so forth.

There are then any number, of indirect relationships of metadata.

BUSINESS METADATA AND TECHNICAL METADATA

The other parameters that belong in the framework describing metadata include the division between technical and business metadata. Technical metadata relates to technology and includes such things as:

- Tables
- Attributes
- DBMS, and so forth.

Figure 11: Examples of technical metadata.

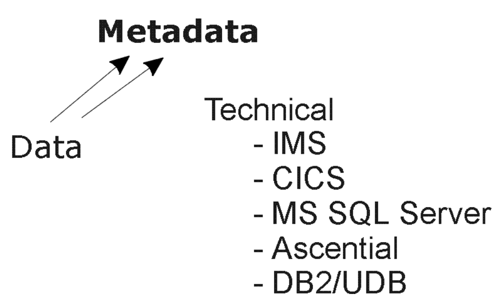


Figure 11

Business metadata is useful to the business person and typically includes:
- Report name
- Report column description
- Formula
- Summarization calculation, and so forth.

Figure 12: Examples of business metadata.

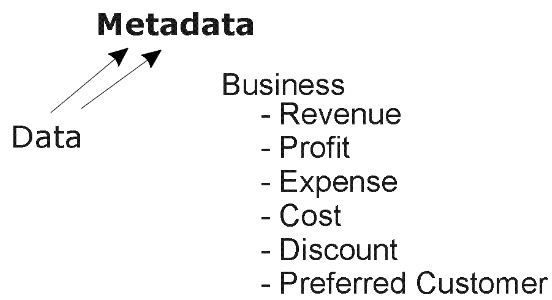


Figure 12

Business metadata is very important in that it is through business metadata that the business person relates to and understands information being placed in front of him/her. Unfortunately, nearly all of the products and efforts of metadata to date have related to and supported technical metadata.

STATIC METADATA AND TIME VARIANT METADATA

The second sub classification of metadata that belongs in the framework is that of static metadata and time variant metadata. Static metadata does not change - it is set once during the life of data that is being described by metadata. Typical static metadata might include:
- Table name
- DBMS
- Operating system, and so forth.

Time variant metadata is subject to change. There are many forms of time variant metadata, such as:
- Date of ETL processing
- Date attribute changed
- Date of pointer resolution verification
- Date of quality assurance audit
- Date of structural change of definition, and so forth.

In fact, there are many more time variant forms of metadata then there are static forms. As in the cases of time variancy found elsewhere, time variant records are stored in snapshots.

THE FRAMEWORK FOR THE MOST BASIC DIVISIONS OF METADATA

The result of mixing these parameters together is a matrix, depicted in Figure 13.

	Business		Technical	
	Static	Time Variant	Static	Time Variant
Direct Descriptive				
Indirect Descriptive				
Direct Relationship				
Indirect Relationship				

A Framework for Classifying Metadata

Figure 13

The matrix is useful for several reasons:

- It suggests a framework for classifying metadata. This is useful from the standpoint of recognizing that all metadata is not flat, but has texture. The framework starts to demonstrate some of that texture.
- It can be used as a framework for assessing your organization’s need for metadata. Cells that are fully populated will tell you what you already have. Cells that are not populated will tell you what you need.
- It suggests that there are ways to organize metadata.
- It can be used as an evaluation tool in comparing one vendor’s product against another’s, and so forth.

In a word, the matrix is a first step at starting to classify and understand the interrelationships of metadata items throughout the organization. The world of metadata as seen through the matrix starts to take on the character and texture that, in fact, has been there all along.

As useful as the matrix is, it is worth noting that there are other aspects of metadata that are not addressed at all by the matrix. These dimensions include:

- Distribution and centralization
- Meta processing
- Ownership
- Business rules
- Architectural structure, and so forth.

However, the matrix does address the most visceral and basic aspects of metadata.