The Bloor Group

Mastering Metadata

Robin Bloor, Ph.D.
& Rebecca Jozwiak

WHITE PAPER

The Metadata Malaise

Metadata may not be the sexiest topic, nor may it be at the forefront of everyone's mind. But it needs to be.

The operational culture of the enterprise has gradually shifted towards a more business-centric model, one where business users and front line information consumers do not just desire better access to enterprise data assets, but need to have a comprehensive view of the organization. This means that data, as well as the metadata that describes it, must be shared, both at the developer level and the BI level. This is easier said than done.

The importance of metadata within the data warehouse context cannot be understated. In many ways it is the backbone of enterprise data warehousing and information management. Metadata serves as a directory for everything that is in the data warehouse environment and most other data stores, providing detailed information about data and objects. In large organizations, metadata is typically dispersed and fragmented across departments and business units, and some of it resides in spreadsheets, text files and multimedia files. These files are rarely shared and accessible throughout the enterprise.

You cannot share data effectively without managing the metadata resource. Here's why.

Metadata defines the structure of data in files and databases. Often this definition data is buried in program code, and only the program knows the precise data structure. That is not in and of itself the problem, as the program is quite able to use the data appropriately. The problem is that the metadata is meaningful to that program, but not to the human user or possibly even to any other program. If only one suite of business applications were in use throughout the enterprise, this might not be so constraining. However, almost no organizations have such a homogeneous software environment.

Consider the following situation. Within a particular database the "Person" table is defined to consist of: Person-code, Title, First Name, Last Name, Job-Title, along with some data type information. Even if you are told that the data is from a database used by an HR application, there is not enough context to determine exactly what the data refers to. A specific row in the table might refer to a staff member, or a contractor, or a former employee. If this is the case, sharing this record would require an additional explanation as to what the reader is looking at.

In practice this imprecision will be present for many tables defined in many databases across the business. As a consequence, when there are no implemented standards or methods for sharing data, it is practically impossible to ensure accuracy and consistency of reporting, analytics or decision making.

The main reason for sharing data is to populate the diverse array of business intelligence and analytics applications within an organization. And yet application data is frequently defined in a contextual way – for a particular use case within a particular business unit – because the application is not used anywhere else within the organization. These applications could be homegrown or cloud-based. All these operational applications help drive the business, and yet they all carry different metadata that only the program code truly understands.

Metadata needs to be organized and managed if it is to be leveraged effectively. Most data analysts and business analysts are very familiar with reports within their own data marts and focus areas, but when they step outside that comfort zone, that familiarity vanishes. Today's

expectation – that business and data consumers understand how data is used across the organization – means users need to know what the data is, what it means, where it came from, etc.

Big Data Means Big Metadata

Why is there so much metadata? Quite simply, because there is so much data. Businesses have seen data volumes grow at roughly 55% per annum for some time now, and that growth is unlikely to slow.

Recent years have seen an explosion in the number of data sources as well, which in turn increases the amount of metadata.

The main sources of this data are:

- **Business data:** Think of this as traditional data from business applications that we tend to collect in traditional data warehouses.
- **Log file data:** Think of this as operational data (database logs, network logs, OS logs, etc.) that we generated in the past but rarely looked at.
- **Mobile data:** Think of this as location data. Many companies don't collect much of this, but some do. In the future we will collect more of it.
- **Social network data:** This data is available, useable and frequently used.
- **Public data:** There are many sources of publicly available data, such as census data.
- **Commercial databases:** There is now a burgeoning market in companies selling data.
- **Streaming data:** This is commercial data sold as a continuous stream.
- **IOT data:** The "Internet of Things" includes sensor data of various kinds. This is currently relatively small as a source of external data, but predicted to grow very large, very soon. It probably will.

When data is flying in from all possible directions, in varying formats, from various unrelated sources, metadata management becomes much more complex. Even if an organization has a strong grip on internal data definitions, the addition of external data can spell disaster for the coherence of the metadata resource.

One way to meet this challenge head on is to create and maintain a searchable, extensible metadata registry that allows for data enrichment and facilitates sharing. The idea is to establish a sort of map that includes all enterprise data sources, internal and external, and allow users to add business information to the metadata. Essentially, those who use the data become responsible for providing useful business definitions of what the data is and does.

Few software vendors address this specific need. There are plenty of Master Data Management (MDM) offerings, where the goal is to establish a well-defined and consistent master model of enterprise data. However, MDM projects tend to lead to political in-fighting over data and business definitions, data ownership and policies. Such projects are also long-running and tedious, and can easily become overwhelmed by the continual growth in the number of data sources.

Fortunately, IDERA Inc. offers a metadata collaboration platform that delivers on all fronts.

The IDERA Environment

IDERA provides a collaborative metadata platform called ER/Studio Team Server, and it is aimed squarely at solving the problem of disparate data sources and definitions. When combined with IDERA's data modeling tool, ER/Studio Data Architect, it becomes a data enrichment platform which delivers an enterprise-wide big picture of how data is spread around the organization, and what that data means.

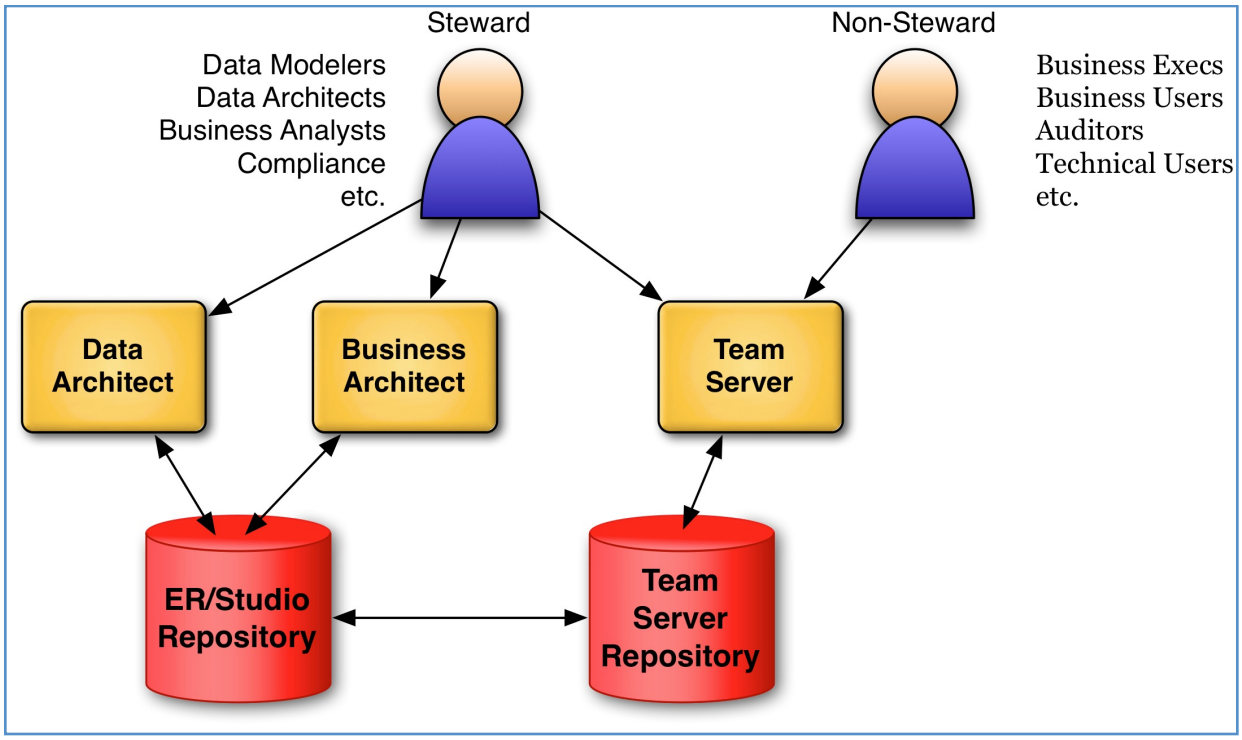

Figure 1: ER/Studio Team Server in Overview

The above diagram depicts how ER/Studio Team Server is normally deployed with IDERA's other products. IDERA thinks of the user community as consisting of data stewards and non-stewards. The stewards are professionals such as data architects, developers and business analysts, whose job involves database design and maintenance. They use IDERA's Data Architect and Business Architect tools to define and update data definitions stored in the ER/Studio Repository and, collectively, possess considerable knowledge of the business's data at a design level.

The non-stewards are users who can contribute to enriching the metadata definitions by virtue of having a deep knowledge of how data is used within the business. Business executives, auditors, line of business staff and in fact, anyone not involved in data design activity will fall into this category. A data analyst can, for example, go into ER/Studio Team Server, click on a certain metric, and get a broader picture of where that data is and how it is stored across the organization.

It is worth noting that ER/Studio Team Server caters to different comfort levels when it comes to views. As such, a technical user will get a complex view of the data landscape, whereas a business user, unless he chooses otherwise, can be presented with a high level representation.

Differences of Opinion

It is fairly common – in fact, it is expected – that various stakeholders within an organization will espouse different definitions or opinions about definitions of terms or business rules. This can have many implications, particularly in the areas of compliance and regulations.

The ER/Studio Team Server addresses this in two ways. First, it employs an enterprise glossary of business definitions and data elements. Second, it integrates the glossary with data management tools, allowing users across the enterprise to access a single repository of business definitions and data sources. Each term is linked to the metadata, which means a user can look at any item within the glossary and locate the reports that include the term in the metadata layer.

ER/Studio Team Server additionally provides social collaboration within the tool. Users can have discussions around areas of interest, be it the model, the glossary, a subset of data, and so on. If, for example, User A adds a new term to the glossary and User B disagrees with that term or its definition, the latter can begin a collaborative discussion to work on improving or clarifying the term. These conversations are persisted within the application, and a full audit trail of the discussion remains available in perpetuity.

The benefit to this is obvious. Any one, at any given point, can look at the record between User A and User B to determine how and why they reached a consensus. If User C comes along and disagrees with the term or definition, he can immediately look back at the previous conversation and potentially mitigate a second round of discussion. This doesn't just foster collaboration – it fosters *understanding*.

Users at any level can also create their own custom metadata to associate with the models without changing the master glossary. Such metadata might be an extended definition of the term "supplier," or it might provide additional context around a policy or business rule. The point is, users are empowered to further their understanding of business terms and policies.

The two main impacts of ER/Studio Team Server are that it enables users to make better use of the data available to them because it surfaces the meaning of the data, diminishing misunderstandings and ambiguity, and it supports the implementation of standards and regulatory compliance.

In effect, ER/Studio Team Server helps a company to develop and maintain an information map of the business – a single searchable registry of all data sources used by the business – that can be leveraged by both BI users and developers.

Bridging the Gap

No one would dispute that an enterprise needs to manage its data. In turn, no one should dispute that an enterprise also needs to manage its metadata.

Today's organizations know the gravity of effective data management and the challenges that lie therein. Data is diverse, and not just because we are seeing more of it. We are seeing more diversity with the growth of unstructured or semi-structured data. And we are seeing new sources, including streaming data, social media data, sensor data and so on.

If organizations are to leverage data as an asset, they need to manage both the data and the metadata, and they need to manage the metadata both at the software developer level and at the business level.

However, there is often a disconnect between data managers, or stewards, and information consumers. Typically, data managers focus on creating and maintaining data models for the purpose of building applications, feeding reports and dashboards and enabling analysis, while business users focus on organizational improvements and decisions. There are some exceptions to this, but by and large, most organizations are segmented in this manner.

Some organizations have tried to improve communication by forming cross-departmental teams. This is helpful, but only in a limited way. Although there might be detailed meeting minutes, there is no true audit trail of discussions and decisions, and there is unlikely to be any record of side conversations. And what of questions that arise in the course of the month in between meetings? Emails and instant messages are frequently used but are unreliable as a means for a long term discussion, especially when the discussion might be around several items of interest and among many different contributors.

Only a meaningful, managed, collaborative environment can produce the level of investment that organizations expect. Business definitions and glossaries play a critical role in determining how the business is run, and when data stewards and business users can identify, track and manage that data together, it can lead to a solid foundation of quality, consistent business practices.

In our view, IDERA is one of just a few vendors who provides a comprehensive solution to a very specific problem. By treating metadata as an asset, IDERA can set businesses on a path to greater enterprise-wide understanding and growth.