IDERA® | ER/Studio Whitepaper

MAKE THE MOST OF YOUR METADATA

BY DAVID LOSHIN of KNOWLEDGE INTEGRITY, INC

Using Collaborative Metadata for Structure, Content, and Semantics to Enhance Enterprise Knowledge Sharing

INTRODUCTION

Metadata management is not a new concept, although it is gaining new adherents as the value of enterprise data governance is increasingly recognized. The success of metadata projects may have been hampered in the past due to misguided attention on the deployment of metadata tools and population of the repository.

This white paper addresses considerations for collaborating on metadata within and across an organization. First, the paper considers how the organic and distributed evolution of the enterprise application ecosystem has allowed inconsistencies to be introduced that create challenges as the data sets from those applications are repurposed for reporting and analytics. Next, the question of the value of metadata management is raised, suggesting that a strategy for metadata management and use must clearly direct the types of activities for metadata discovery and capture.

TABLE OF CONTENTS

THE EVOLUTION OF ENTERPRISE DATA CONFUSION

In many organizations, there are recurring challenges in meeting business process requirements that are driven by gaps in effective communication and information sharing among the data practitioners and the business domain owners. The scale of these gaps can range from the macro level (such as multiple yet variant definitions of key business terms such as “customer” or “material”) to the micro level (such as variances in field lengths for commonly-used data elements).

Understanding the root cause of variances among value formats, data element structures, and business term definitions requires a little bit of a history lesson, examining the organic development of the enterprise application ecosystem. During the early days of computing, the risk of data variation was relatively small. All applications were run in batch on a single mainframe system. Records were stored in flat files accessed by the batch applications.

However, as workgroup computing came into vogue, business areas began to invest in their own infrastructure accompanied by their own custom-developed applications. Each business area’s set of applications were intended to support transactional or operational processing, and the functional requirements reflected the specific needs for initiating business process workflows and completing the corresponding transactions.

As an example, consider a typical bank deposit transaction in which the customer presents a check for deposit in a named bank account. This specific process is focused on the operational aspect of logging the transaction by finding the named account and increasing the stored balance by the amount of the deposit. The identity of the customer, her behavior patterns, and other analytical aspects were irrelevant to achieving the outcome of the transaction, and so there was no demand for high fidelity in capturing or even documenting that information.

Nevertheless, as operational reporting and business intelligence have evolved, there has been a corresponding growing interest in repurposing data created or captured as side-effects of the transaction. For example, the bank might want to know which accounts are associated with which individual customers, and what the total transaction volume is by customer, not just by account.

As data sets collected from different functions within the business are repurposed for reporting and analytics, issues of data semantics emerge. Isolated design and development allowed different business function areas to use similar concepts in their data models and applications in slightly different ways – different definitions, data sizes, types, formats, and allowable values. The concentration on satisfying functional requirements for transaction or operational processing needs has led to inconsistency and low fidelity across uses of the same or similar business terms and corresponding data element representations.

These differences do not matter significantly as long as the applications continue to operate within the functional context and within their own domains. But when the data sets are accumulated for reporting and analysis, small differences suddenly have significant impact. A very simple example involves sharing records containing unique identifiers that have been assigned by different authorities. This happens frequently when data records created in different environments are exchanged across domains, such as in healthcare identifiers, banking account numbers, social service case identifiers, or where similar sets of records are accumulated for reporting and analysis.

As these data sets are forwarded to analytical environments (data marts, data warehouses, or even streamed into desktop self-service BI tools), the types of variances begin to unfold – differences in the values (such as names being spelled differently), differences in the structure (a last name field in one data set is 25 characters long while a last name field in another data set is 30 characters long), and differences in the definitions ("last name" refers to a residential customer's family surname in one data set while "last name" means the previous name used for a commercial customer in the other data set).

THE METADATA CONUNDRUM

While data professionals don't necessarily need to shoulder the blame for these variances, in a modern information-aware enterprise they are entrusted to seek out ways to mitigate the impacts of legacy variation. A reasonable approach is to understand where the variations occur, what the root causes are, and what can be done to address those root causes. This can be facilitated through a strategy for metadata management – collecting and managing the "data about the data" that provides the context for the formats, standards, definitions, and meanings associated with reference data sets, data element concepts, data elements, records, and tables.

A metadata tool provides a repository for collected information, but if it is only used to store metadata without considering how that knowledge can be made actionable, the tool's value is significantly diminished. That being said, the potential resource demands for researching, documenting, and validating data element metadata can be surprisingly large, raising questions about the investment of capital and time. At the same time, in the absence of any previously identified discrepancies or inconsistencies about conflicting formats, value domains, or definitions, there may appear to be few reasons to invest the energy in populating the repository.

In other words, we have a conundrum – when there is a need for the metadata, without the prior investment in populating a tool's repository, there won't be any of the critical information available to help the consumers. However, without a need, it may be difficult to acquire the resources necessary to capture and document the metadata in the first place!

DEVELOPING THE METADATA STRATEGY

That is the reason to establish a strategy for metadata that encompasses the finer points of policies, processes, and expectations for ensuring enterprise data usability. Focusing on a tool alone effectively defers the analysis of the metadata until the tool evaluation and acquisition process has completed. However, energizing tool acquisition within a more complete strategy allows you to develop guidelines for metadata collection and harmonization. At the same time the strategy will drive the determination of business needs to be addressed by corresponding technology. In turn, the objective morphs from initiating short-term projects intended to storm through data models and note the data element types and lengths into an ongoing program for aligning semantics for information consumption in ways that are consistent and instill a level of trust in analytical results.

Collecting information about the definition, structure, origination points, and lineage of data concepts provides a road map for any downstream data consumers interested in using the data for alternate purposes. A proper metadata strategy can guide the users in knowing when entity concepts and their corresponding data elements are compatible and can be subjected to reporting and analysis in ways that are consistent and that make sense.

More importantly, the strategy introduces processes that bring data consumers and data practitioners together to enumerate potential issues, assign priorities in relation to the business demands, and organize communities of interest around particular areas of content. That enables the ecosystem surrounding business metadata to evolve into a framework through which the data professionals and business data users communicate and more effectively articulate the needs and expectations of the members of the business data consumer community.

STRUCTURE, CONTENT, MEANING

The foundation of actionable metadata is the ability to expand beyond logging the typical structural representation metadata that is either managed within a database catalog or captured in a generic “data dictionary.” This is particularly true as the demand for data repurposing grows, such as within reporting and analytics systems like data warehouses, data exposed to end-user analysis and visualization via data virtualization, or streamed to big data platforms, for example.

A good starting point links business terms with standardized definitions, the origination points, and data lineage associated with reference data domains (such as Country Codes or Currency Codes), entity concepts (such as Customer or Vendor), and data element concepts (such as CustomerCity, CustomerState, or CustomerZIP) that are manifested as entity data attributes. This would help in designing and developing new applications that reuse data in ways that are consistent with the underlying semantics.

In spite of reasonable business drivers for collecting and managing corporate business metadata, it is important to realize the potential exploding scope of metadata harvesting. The wealth of systems, their associated databases, integrated tables, and corresponding data attributes create an overwhelming number of data artifacts whose metadata characteristics need to be analyzed and documented. Given the perception of the virtually limitless amount of work that could be expended in metadata analysis, what is the best way to allocate resources for metadata analysis that will capture the actionable knowledge that best benefits the business users?

Because the most significant issues that impact data repurposing arise from metadata variance, a reasonable first step involves assembling an inventory of data artifacts along with their basic metadata. This allows you to compare and contrast the artifacts to see which are similar and which ones differ. That provides the baseline knowledge to identify those variances that have the greatest potential for incurring undesired business impact. This suggests three metadata facets on which to focus:

- **DATA STRUCTURE** Documenting data element names, their types, and their lengths or sizes will provide a solid foundation for metadata organization that will facilitate identification of similar data element concepts. For example, two data elements that share the same name, type, and size will be prime candidates for further evaluation of similarity.

- **CONTENT** What is the authoritative description of the data values that are valid for the data element? And what are the constraints on the data element values? Even if two data elements share the same name, type, and size, if their content domains are not identical, it would appear that they truly refer to different data element concepts.

- **MEANING** What are the authoritative definitions, and how are the levels of “authority” prioritized? As a matter of completeness, can the metadata analysts review the ways the data elements are used and verify that those uses are consistent with the defined semantics?

Resource allocation and effort priority can be focused on accumulating these three pieces of information for critical data sets, and establishing a foundation for analyzing data element concepts and their instantiations will more easily enable data harmonization.

HARMONIZE OR DIFFERENTIATE?

Given an inventory of metadata entries associated with a broad set of data elements pulled from across the enterprise, the natural thing to do is to try to organize them into buckets that represent similar attributes or characteristics of commonly-used entities. A straightforward example is attributes of customer records: data elements with names such as FirstName, first_name, FIRSTNAME, and first might be lumped together, while those with names like LAST, LASTNAME, LastName, CustomerLastName, or Last_Name might be put into the same group.

In turn, we might also want to look at entity models from different sources to see if they share a general structure that would suggest that they represent the same real-world entity. **This is the concept of harmonization – eliminating the variance in data concept definitions and structures so that they are aligned across different operational environments.**

Alternatively, one might consider the point at which two models are distinct enough to definitively state that they don't represent the same entity type. Ultimately, the question boils down to this: if we have two different structures or definitions for what appear to be a single concept, should we harmonize the definitions and structures into one and eliminate the redundant data element?

In some cases this will be a good idea because it will increase consistency, but only as long as the two concepts really refer to the same real-world idea. Alternatively, the two concepts may be similar but not exactly identical. As an example, consider the term "customer," which might have slightly different semantics depending on the context. From the sales perspective, the customer might be the individual who pays for the product. However, at the same time and in the same organization, the customer service department might deem anyone using the product as a customer, and this group may be much larger than the set of "paying" customers that the sales department cares about. Merging these two entity types into a single definition and structure would probably significantly disrupt the business!

In this case, one would want to make sure that the definitions and structures are differentiated, not harmonized. Differentiation means resolving the semantic or structural ambiguity by qualifying the difference between two data elements. In our case, we might choose to rename the customer entity in the sales department as a paying customer and change the name of the customer service department's entity to support customer to indicate that **they are not the same thing.**

But how can one make that determination? Establish some common-sense criteria and basic rules for consideration when harmonizing data elements and entity models. This will help find the "low-hanging fruit" candidates for harmonization. An example might be if two data elements in different tables:

- Share the same name,
- Share the same definition,
- Share the same data type,
- Have the same size/length, and
- Take their values from the same domain of valid values

then the pair of data elements is a clear candidate for harmonization. Two data elements that share none of those characteristics are clear candidates for differentiation.

ENGAGING THE BUSINESS USERS IN ENTERPRISE SEMANTICS

The challenges emerge when two similar data elements do not reflect 100% symmetry. Two data elements with the same name, data type, and length that have different definitions would raise a red flag as to whether they can be harmonized. In this case, one must engage the business users as well as those who are accountable for ownership over the data to review the semantic differences in the use of that data element.

From a practical perspective, the mechanical processes of metadata harvesting, assessment, collation, review, and harmonization are only relevant when business users are paying attention to the outcomes. Their attention must be grabbed so that they will help in the process to review business glossary terms and data elements and their corresponding definitions to determine where there is sufficient similarity among metadata concepts to harmonize their definitions or if their definitions are distinct to take the steps to ensure their differentiation. More succinctly, the harmonization strategy must leverage the metadata management environment to encourage collaboration between the data practitioners and the business data consumers regarding harmonization and differentiation.

The value and criticality of an accurate and synchronized metadata repository is only apparent when there is a business need for business term reconciliation to address perceived inconsistencies, such as when two different reports show completely different numbers of current customers. However, the absence of any inconsistencies hides the criticality. This suggests some degree of tension regarding the perception of the value of high quality actionable metadata; while the need for maintaining accurate and current metadata is constant, its relevance is limited to those situations where the business is using the metadata.

The task for the data practitioner is to build awareness among the business data consumers to highlight the real motivations behind effective metadata management: continuously satisfying the needs of the business data consumers without forcing them to become experts in data management techniques.

Engaging the business users begins with education in demonstrating that business metadata is the lingua franca for enterprise communication. Reports, presentations, spreadsheets, and documents are rife with business terms that are ripe for misinterpretation. Customer, product, part, site, vendor, and employee – these are all examples of business terms used almost universally yet are seldom (if ever!) clearly defined. It would be surprising to review even a small handful of corporate reports without finding discrepancies in the presumed semantics of commonly-used business terms. Pointing these out to key business data consumers provides the foundation upon which awareness is built about semantic inconsistency, data harmonization, and the need for metadata management.

Reach out to the business users and engage them on their own terms:

- Select a sample of enterprise information products: reports, presentations, or documents;
- Scan through the artifacts and select the most frequently-used business terms;
- If possible, provide the most reasonable definitions of these business terms as they are used within the context of each artifact; and
- Present the variant or discrepant definitions to the owners of those information products;

The last step is the most important:

- **Ask the data users about the business impacts of the indeterminate semantics.**

Ideally, the business users will immediately provide feedback about potential issues, or they will be intrigued and will investigate the possible impacts on their own. Either way, highlighting the risk of business impact motivates business involvement in establishing processes for researching the candidate terms and concepts for a business term glossary. This initial engagement helps to foster a cultural shift in establishing collaboration among the business users and information professionals to synchronize critical business terms, their definitions, and their uses across the application landscape.

USING COLLABORATION FOR SPREADING ENTERPRISE KNOWLEDGE

Fostering collaboration across the enterprise community enables the introduction of the types of standardization that support an enterprise data governance program. However, without providing a means and a platform for continuous engagement, interaction and commitment, there is a risk that the harmonized metadata can become stale (at best) or obsolete (at worst).

Fortunately, semantic metadata can become the language through which business users and data practitioners communicate and continue to share enterprise knowledge. This communication requires a medium through which the different players interact, which is where the informed selection of the right kind of metadata tool comes in.

A typical metadata "repository" is just that: a repository. While it can be a good tool for storing metadata, in many cases it is difficult for the stored metadata to be accessed and made actionable. What is needed are tools that are not just limited to storing the metadata but instead facilitate knowledge sharing within the context of metadata and collaboration in both definition and use.

That being said, a tool that enables the right kind of collaboration should enable the following types of activities or provide these key capabilities:

- Define business terms.
- Define conceptual domains and reference data sets. Define data element concepts.
- Capture and review data element structural metadata.
- Allow users to subscribe to indicate interest in a metadata concept.
- Provide automatic notifications to subscribed individuals when some aspect of a metadata item has been updated.
- Allow users to search business terms, data element concepts, and definitions for key words.
- Map data elements to their use within data artifacts such as databases or reports.
- Map data elements to their use within business processes and applications.
- Present a broad perspective on data lineage across the enterprise.
- Capture discussions and interactions about definitions.
- Allow for contextual business term definitions that may differ depending on use.
- Provide a means for reviewing definitions as part of a harmonization process.

The common thread that links these capabilities together is enabling interaction to motivate discussion and agreement. And collaborative management for semantic metadata implies a need for tools and technology that not only capture information about business terms, data elements, definitions, structure, lineage, and value domains and other content criteria, but also provide a means for interaction among the individuals who have self-identified their stake in maintaining consistency and precision of definition and use.

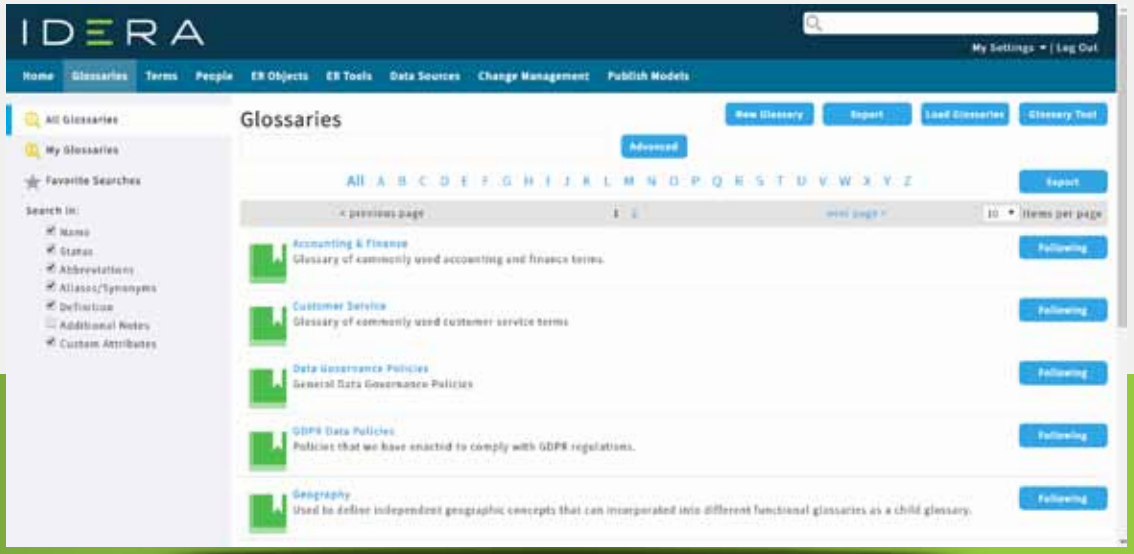
To make the most of your metadata, seek out an environment that allows for comments about definitions, interactive debate, and processes for collaboration and resolution of identified discrepancies. The creation of a semantic metadata community facilitated through technology helps to transform the "black hole" metadata repository into a living data artifact that synchronizes knowledge across the enterprise. Collaborative methods breathe life into the metadata management processes, and through awareness and information-sharing will elevate metadata into an actionable asset that influences behaviors that lead to increased semantic consistency, reduced confusion, and streamlined and clear communication.

ER/STUDIO ENTERPRISE TEAM EDITION

ENTERPRISE ARCHITECTURE MODELING AND METADATA

- Document and enhance data and metadata from multiple database platforms
- Implement naming standards and a data dictionary to improve consistency
- Share and collaborate on global business glossary terms and definitions
- Effectively communicate models and metadata across the enterprise
- Build a foundation for data governance and compliance programs

Request a Demo



ABOUT THE AUTHOR

David Loshin, president of Knowledge Integrity, Inc., (www.knowledge-integrity.com), is a recognized thought leader and expert consultant in the areas of analytics, big data, data governance, data quality, master data management, and business intelligence. Along with consulting on numerous data management projects over the past 15 years, David is also a prolific author regarding business intelligence best practices, with numerous books and papers on data management, including the recently published "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph," the second edition of "Business Intelligence – The Savvy Manager's Guide," and other books and articles on data quality, master data management, big data, and data governance. David is a frequently invited speaker at conferences, web seminars, and sponsored web sites and channels including www.b-eye-network.com.

IDERA understands that IT doesn't run on the network – it runs on the data and databases that power your business. That's why we design our products with the database as the nucleus of your IT universe.

Our database lifecycle management solutions allow database and IT professionals to design, monitor and manage data systems with complete confidence, whether in the cloud or on-premises.

We offer a diverse portfolio of free tools and educational resources to help you do more with less while giving you the knowledge to deliver even more than you did yesterday.

Whatever your need, IDERA has a solution.

IDERA