# The Role of Metadata in a Data Governance Strategy

Prepared by:

David Loshin
President,
Knowledge Integrity, Inc.
(301) 754-6350
loshin@knowledge-integrity.com

Sponsored by:

## Introduction – Recognizing the Criticality of Data Governance

Since the early 2000s, there has been a combination of factors that has challenged the business world to engender a greater awareness about the quality and usability of information. Examples include recovering customer trust in the wake of financial scandals, instituting systems in reaction to increasing legislation about data privacy, as well as facilitating the creation of public data sets resulting from government mandates for data transparency. In turn, these activities that must be contrasted with organizations seeking to adopt big data management platforms to analyze massive data volumes to create corporate value.

In essence, there is an exploding interest in managing information, and correspondingly, there is a need to institute policies and practices for overseeing the proper management and stewardship, quality, and usability of data managed within the enterprise. The enforcement of data policies and the management of best practices for data management are collectively referred to as "data governance," and its criticality is increasingly recognized in proportion to the perceived value of information to the organization.

An illustrative example in the health care industry is the HIPAA (Health Insurance Portability and Accountability Act of 1996) Breach Notification Rule, 45 CFR §§ 164.400-414, associated with the aftermath of a data breach. This rule requires covered entities (including providers, health plans, health care clearinghouses, and health insurance companies) to notify the U.S. Department of Health and Human Services (DHHS) following a breach of unsecured protected health information[1] if the breach has impacted more than 500 individuals. This is a mandated business policy regarding protecting health information, and there are penalties if the covered entity does not provide that notification.

## Data Governance Defined

There are many perceptions as to what is meant by the phrase "data governance," but for our purposes we will assert that data governance as a managed program for developing, agreeing to, and deploying the policies, practices, and processes for ensuring that enterprise data sets are properly used to enforce compliance with business policies. In turn, data governance transforms business policy specifications into discrete data management directives implemented using a range of data management tools.

To continue our example, in the event of a health care data breach, the HIPAA Breach Notification rule directs the affected organization to assemble a report to be delivered to the DHHS providing specific data: the organization's name, its state, the type of organization, the number of individuals affected, the date reported, the type of breach, the location of the breached information, and a description of how the breach occurred.[2]

[1] U.S. Department of Health & Human Services, Health Information Privacy: Breach Notification Rule http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/
[2] U.S. Department of Health & Human Services, Office for Civil Rights Breach Portal https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf

## Making Data Meaningful

When enforcing data policies, precise definitions of the information concepts employed are critical. The requirements for enforcing the business policy in our health care example seem straightforward – the information concepts seem to be relatively straightforward and easy to collect. Yet we often believe that there is a common understanding of the meanings associated with business terms, even though there are no agreed-to standards, and this complicates the presumption of how those business terms are mapped to discrete data elements.

Consider the data requirements associated with the international banking regulatory framework, Basel III, which requires banks to demonstrate that they manage their funding structure in a way that reduces the risk of failure due to the ways that variations in funding affect liquidity. The Basel II framework discusses numerous business terms (such as "Net Stable Funding Ratio," "funding tenor," "funding type," "funding counterparty," "bank behavior," "asset tenor," "asset quality," and "liquidity value"[3]) that not only may be calculated differently by different banks, there may be different definitions and methods of calculation within the same organization!

To enable compliance with business policies, there must be acknowledged ways of ensuring that data is meaningful, and there are two implications. First, there is a need for a semantic context for how data is to be interpreted, managed, and used. Consequently, the organization must support formal methods for providing the practical oversight of data consistency and quality that are necessary for instituting and enforcing data governance. Both of these facets are reflected in terms of documenting the details of entity representations (such as "customers" or "suppliers") and how those representations are aligned with real-world business policies. In other words, we need to document the enterprise *metadata*: the business terms and their definitions, critical data element definitions, the hierarchical relationships among and between different types of entities, and data lineage.

## Is the Tail Wagging the Dog?

Clearly, metadata is an important part of data governance, and consequently, most nascent data governance programs are rife with project plans for assessing and documenting metadata. But in many scenarios, it seems that the underlying driver of the metadata collection projects is the presumption that it is just one of the "things you do" for data governance. As a result, most early-stage data governance managers kick off a series of projects to profile data, make inferences about data element structure and format, and store the presumptive metadata in some metadata repository. But are these rampant and often uncontrolled projects to collect metadata properly motivated, or are they just a case of the tail wagging the dog?

Upon scratching the surface, one may find that despite the resources invested in metadata assessment and capture, there is seldom a clear directive about how metadata is used. That means that prior to

[3] Basel Committee on Banking Supervision Consultative Document, "Basel III: The Net Stable Funding Ratio" http://www.bis.org/publ/bcbs271.pdf

launching the metadata collection tasks, it is important to specifically direct how the knowledge embedded within corporate metadata is to be used to:

- Assure levels of data quality.
- Ensure consistent use of business terms and associated data elements.
- Monitor observance of defined data policies.
- Enforce conformance with business policies.

Managing metadata should not be a subsidiary goal of data governance. Assembling a repository of data elements and their structural metadata is not an end in itself, but rather the means to a more powerful objective. It is through metadata management that the data requirements are collected, business rules are defined and shared, and by which the directives for ensuring compliance with data policies are manifested.

## The Role of Metadata: Actualizing Data Governance

Metadata is not just one subsidiary component of data governance. On the contrary – proper management of corporate metadata empowers the data governance and data stewardship teams to monitor conformance to corporate data expectations and enforce alignment with corporate data policies.

The breadth of the use of metadata to enforce data policies spans operational and business-facing applications and activities.

- **Data Protection** – the Children's Online Privacy Protection Act (COPPA), which took effect in 2000, is a law mandating protection of private personal information collected online from children under the age of thirteen. Metadata management is necessary to ensure that your organization is compliant with COPPA. First, you need to note the business terms provided by the definitions section of the law (in 15 USC § 6501, section 8)[4] that defines "personal information:

  *(8) **Personal information** The term "personal information" means individually identifiable information about an individual collected online, including—*
  - *(A) a first and last name;*
  - *(B) a home or other physical address including street name and name of a city or town;*
  - *(C) an e-mail address;*
  - *(D) a telephone number;*
  - *(E) a Social Security number;*
  - *(F) any other identifier that the Commission determines permits the physical or online contacting of a specific individual; or*
  - *(G) information concerning the child or the parents of that child that the website collects online from the child and combines with an identifier described in this paragraph.*

[4] See https://www.law.cornell.edu/uscode/text/15/6501

Second, your metadata management capability must provide visibility to all data elements in any managed data asset in which data is collected online and may contain any of the aforementioned variations of personal information. Third, you must define rules associated with the business processes indicating that prior to sharing any user records that may contain personal information, that the user's birthdate is checked to determine whether that user is over the age of thirteen.

- **Data quality** – Metadata platforms provide a means for documenting data quality expectations, such as specifications for data element completeness, observance of defined structure formats (such as email addresses, telephone numbers, or identifiers like social security numbers), and use of values from a standard value domain (such as ISO 3166 country codes[5]).
- **Data model migration** – When an aging system is slated for renovation, a data task requiring oversight is the movement of the data from the system to be retired to its replacement. There is always a risk of variation in the discrete definitions and uses of data domains between the old and new systems, so data element metadata can be employed to verify that the migrated value are valid within the renovated environment.
- **Schema-on-Read** – A data utilization approach that is gaining momentum in the context of big data and data lakes is based on the notion that acquired data should be captured in its original, raw format and that structure would only be imposed when the data is accessed (referred to as "schema-on-read"), as opposed to the conventional practices of transforming acquired data into a predefined tabular structure. While the schema-on-read approach provides more flexibility to a broader community of downstream data consumers, there is added complexity in enabling different consumers to view the data in different ways. Metadata management techniques can be used to document consumer views as well as allow for review of the variant views to make sure that there are no explicit conflicts in data interpretation.
- **Enforcing business policies** – There are business policies that are orthogonal to those associated with regulatory compliance. For example, some large ecommerce companies are introducing same-day delivery to their preferred customers whose orders conform to certain circumstances. Some of these qualifications insist that the items being purchased are available for same-day delivery, the customer's location is within those certain metropolitan areas that are included, and that the order was made prior to noon within the local time zone. These business rules will be translated into data rules to be applied, either when the product's web page is displayed, when the customer adds an item to the shopping cart, or when the order is executed. Of course, these data rules are an emerging form of corporate metadata representing the validation rules that those parts of the order can be fulfilled and that delivery can be arranged for the same day.

## Summary

These are just a few examples of how metadata management provides the foundation for ensuring that enterprise data sets are properly used to enforce compliance with business policies. And while in the past the initiation of metadata "busy-work" might have curtailed the efficiency of the data governance

[5] See the Wikipedia entry https://en.wikipedia.org/wiki/ISO_3166 for a good overview.

council to define data policies and standards, our metadata-centric approach leverages semantic knowledge to simultaneously define and enforce policies and standards.

That suggests clearly considering the key features and characteristics that are required in a metadata product to actualize data governance, such as:

- An integrated **business glossary** used as a enterprise resource for marshaling the collection of business terms and phrases and documenting their authoritative definitions and semantics.
- **Collaboration and visualization tools** that enable publication of definitions, standards, models, and data policies for review, interpretation, discussion, and agreement.
- **Notification methods** to alert interested subscribers as data element metadata is debated, agreed-to, and potentially modified.
- **Role based access** to the metadata repository enabling individuals taking on the different data stewardship and governance roles to facilitate enterprise information management best practices.
- Specific metadata **extensions around governance** (such as "user defined properties") enabling data consumers and managers to define reusable properties that can be standardized and used across the enterprise.
- **Forward engineering** metadata to guide assessment of aging systems and applications, the interpolation of inherent structure and architecture, and the ability to automate the generation of evolved data models (and corresponding data definition language) targeted at specific databases.

When launching the data governance program, recognize that the ability to enforce compliance with enterprise data policies is rooted in the ability to formalize how compliance can be specified as enterprise metadata. Consider acquiring metadata management technology and devise the practical metadata management processes and procedures to accompany data governance to improve the probability of a successful program.

## 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 data quality, master data management, and business intelligence. David is a prolific author regarding BI best practices, with numerous article published at searchBusinessAnalytics.techtarget.com, and numerous books and papers on Big Data, Analytics, Business Intelligence, Data Warehousing, data governance, and data quality. Visit http://dataqualitybook.com for more of his insights. David can be reached at loshin@knowledge-integrity.com.

## About the Sponsor

*Sponsored by CA ERwin® Modeling — providing a collaborative data modeling environment to manage enterprise data through an intuitive, graphical interface. With a centralized view of key data definitions, you can have a better understanding of corporate data, managed in a more efficient and cost-effective way — placing ERwin at the Center of Data Management. For more information visit www.ERwin.com*