DATA MODELING IS A FORM OF DATA GOVERNANCE

BY ROBERT S. SEINER

TABLE OF CONTENTS

INTRODUCTION

Data Modeling is a form of Data Governance. Let me explain. I define data governance as "the execution and enforcement of authority over the definition, production and usage of data and data-related assets." The management of data begins with governing the definition of the data. This is also the most basic goal of delivering quality data models — delivering high-quality data definition that meets organizational requirements. Therefore, if you agree with my definition of data governance, you will likely also agree that data modeling itself is a form of governing data — specifically governing the definition of your data.

Data modeling is all about data definition but has a much wider impact on the data of your organization. The quality of the data definition has a direct impact on many other aspects of the data lifecycle. Quality data definition impacts how data is produced and directly impacts how the data is or will be used throughout your organization. These statements are the basis for this white paper. Let me restate these as facts — the quality of data definition has a direct impact on the quality of the production and usage of the data.

If quality data definition is THAT important, then we better make certain that we execute and enforce authority over how we define data. That means that we must govern the process of how we define data. Again, data modeling is a discipline that must be governed, making data modeling a form of data governance.

THREE ACTIONS OF GOVERNING DATA

I am a firm believer that there are only three actions that can be taken with data. I have challenged many people to come up with an additional action or two that does not fall under one of my three. Everything people do with data falls into these categories.

The actions are: **people define data; people produce data; and people use data.**

The main premise of the non-invasive approach to data governance says that every person that either defines, produces and/or uses data must be held formally accountable for their behaviors associated with each action. Therefore, the Non-Invasive Data Governance approach recognizes that EVERY person who participates in one or more of these three actions becomes a steward of the data; that is, they will be held formally accountable for their behaviors. The truth is that everybody is a data steward.

This paper will look at the data modeling discipline as a form of data governance from the perspectives of how modeling impacts all three actions that people take with data.

GOVERNING THE ACTION OF DEFINING DATA

The action of defining data may be the most important of the three actions. The act of governing the definition of data leads to advances in the other two actions— namely improvements in quality data production and data usage.

Organizations that follow strict processes around defining data also seem to have less data to manage. Okay, that may not always be the truth— but it does make sense. Organizations that prevent duplicate data sets from being created have less data to manage. Organizations that have knowledge about all of the data sets being defined across the organization usually have a handle on their data inventory as it grows naturally or through acquisition or merger.

I propose that we look at how we are governing data definition by answering a few simple questions:

1. **Does your organization model its data as part of the process of defining data?**
2. **Is there a process including the steps that must be followed to define data?**
3. **Do the steps include gaining validation of the data definition with the customers?**
4. **And most importantly: are the steps that you defined being followed?**

These questions focus on the most basic actions of governing data definition. If you answered "no" to the first question, the chances are that your un-modeled data is either completely ungoverned or it is defined using internally developed tools (such as a spreadsheet or data dictionary) to record the basic qualities of the data. The attributes of data definition include business name, business definition, valid values, and common attributes of data design including data type, positions, etc. The technical data definition qualities are often handed off to the database administrators to build the database.

In this era of unstructured data, big data and the plethora of non-traditional data sources, it is important to govern the definition of these sources and keep an inventory of the sources from a stewardship perspective. Many organizations consider the definers of the non-traditional data sources to be the "data owners." In data governance, these people are typically referred to as Data Stewards or Data Resource Stewards.

RELATING DATA DEFINITION TO DATA MODELING

Data modeling is described as a series of processes used to define data requirements that support business processes. Data modeling often focuses on conceptual, logical and physical data definition — each representing the informational components of the organization at differing levels of abstraction. The models most often result in databases and data resources that become integral to the organization's information system landscape.

Besides following standard data modeling conventions, data modeling best practices include 1) getting the right people involved in defining requirements, 2) recording appropriate qualities of the data (metadata), and 3) resolving differences in opinion and business understanding. These three practices are all forms of data governance.

The first practice of involving the right people requires that you can recognize and involve the appropriate people in the steps of your modeling processes. I often describe involving the right people as the lead statement of my Data Governance Bill of Rights:

Getting the **"Right"** People

Involved at the **"Right"** Time

In the **"Right"** Way

Using the **"Right"** Data

To Make the **"Right"** Decision

Leading to the **"Right"** Solution

The word "Right" is in quotations to let you know that I mean the "right" thing to do (which can be different in different situations) rather than the rights of the people.

The second practice focuses on recording the appropriate metadata as part of the data modeling process. Data model metadata typically includes the core definitional qualities of the data including business terminology, definition, sensitivity, and rules as well as the physical attributes of the data in the database.

The third practice requires a governed process to resolve differences in business opinions about how the data should be defined. Business areas and important individuals often have contrary or dissimilar ideas of "how the data should look" or how the data should be defined based on their experience of what they believe to be best for the organization. These need to be addressed proactively to establish a common understanding across the team.

GOVERNING THE ACTION OF PRODUCING DATA

The action of producing data is directly related to the action of defining data. Data can only be produced as well as the data is defined. Data can be produced manually or through data acquisition. Quality data definition leads to improved understanding of manual data production requirements. It is difficult or impossible to meet manual data production requirements when the people responsible for producing the data do not understand how the data is defined.

Data that is manufactured or produced from other data is often the organization's most critical data while, at the same time, the data that is least understood. It is a best practice to assure clear definition of how this data is produced, derived, calculated, matched, sorted, rolled up, and broken down. It is best practice to govern how manufactured data is defined and to make that definition available to the people that consume this data directly through the databases or the reports they receive. The most valuable business intelligence data is the data that is manufactured and defined for user purposes.

Individuals that produce data as part of their job must be held formally accountable for the data they produce. This requires the governance of the processes for producing data. The governance of these processes makes certain that everybody that produces data is aware of and follows the rules associated with producing the data. In many circumstances, governance also includes sharing the knowledge of how the data will be used.

RELATING DATA PRODUCTION TO DATA MODELING

To the casual business person observer, Data Modeling may not appear to have a direct impact on the production of data. However, the evidence points to the contrary. Organizations that govern the modeling of their data have a better chance of producing higher quality data.

Organizations often depend on data that moves throughout the enterprise. As data moves, from data store (system) to data store, or from data store to a business intelligence platform, there are specifications for how the data must look and the quality of that data. If data modeling lies at the core of quality data definition, the definition of the data in the data model, including physical attributes, valid values, and business definitions must be used to make certain that the supplier of the data understands how the data must be produced. Without the detailed definition of the data, the data producer manufactures the data to the best of their knowledge, which may or may not be what the business needs. Basically, governed data definition leads to better data quality of data that moves within the enterprise.

Organizations also depend on data that is produced externally to meet specific business requirements or meet specifications required by the organization to receive and absorb that data. Let's address these two ways that improved quality of external data will provide benefit to the organization.

Organizations either have authority over their external data sources or they do not. When the organization has the authority to demand quality external data or they have significant influence over the quality of the external data, the quality of the data that is received benefits greatly from governed business data definition provided by the receiver to the producer of the data.

Organizations that acquire data from sources where they have no influence over the quality, typically are tasked with aligning that data with their data specifications for the acquisition. The ability to align the acquired data with internal data specifications also benefits greatly from having high-quality governed business data definition.

High-quality governed business data definition, and therefore data production, begins with data that is modeled. The process of modeling the data is a form of data governance.

GOVERNING THE ACTION OF USING DATA

The action of using data is directly related to both the actions of defining data and producing data. Data usage is dependent on people understanding the data they are using. This understanding comes from quality data definition that takes place during data modeling or other data definition processes.

Data usage includes the risk management components of 1) protecting sensitive data and 2) following compliance and regulatory requirements. Let's address each of these risk management components separately.

Protecting sensitive data is a requirement that impacts all businesses. Sensitive data includes: Personally Identifiable Information (PII), Protected Health Information (PHI) or Intellectual Property (IP).

- **PII** data is data that can be used on its own or with other data to identify, contact, or locate a person.

- **PHI** data is any data about health status, health care, or health care payment that is collected by an entity and can be linked to a specific individual.

- **IP** data include trademarks, copyright, patents, design rights, and in some jurisdictions trade secrets.

Quality data definition includes the defining of rules associated with protecting sensitive data. These rules focus on the handling of data that is classified as confidential or sensitive. Handling rules include how to share, print, distribute, transmit, use, and discuss sensitive data. Data governance involves formal execution and enforcement of the rules and processes associated with the protection of sensitive data.

Auditable compliance and regulatory reporting begins with the people with these responsibilities being provided with a thorough understanding of the rules they are expected to follow. In general, compliance implies that the organization must conform to rules, policies, standards, and laws. Data governance is the execution and enforcement of these rules.

Data governance requires that the rules are collected, recorded in a digestible manner, approved, communicated, and enforced. These actions require that an organizational entity, a Data Governance Office per se, be given the responsibility and authority to execute and enforce these actions. The governance of the definition has a direct impact on quality usage of data.

RELATING DATA USAGE TO DATA MODELING

We have already stated that data modeling has a direct and positive impact on the actions of defining and producing data. The same can be said for the action of using said data. Data modeling, and the metadata collected during the process of modeling data, can result in several key improvements when it comes to how data is used across the organization. Three of these improvements include:

1. **Organizations can improve people's knowledge of which data to use.**
2. **Organizations can improve people's understanding of the data itself.**
3. **Organizations can improve people's knowledge of data quality requirements.**

All three of these improvements require governed processes associated with data modeling. As stated earlier, the governance of these data modeling processes requires that the right people are involved at the right time to define the data in the right way. The right way means that the correct metadata is collected within the modeling environment.

FOCUS A DATA GOVERNANCE PROGRAM ON THE THREE ACTIONS

Non-invasive data governance operates on the premise that all people in the organization that define, produce, and/or use data must 1) be held formally accountable for the quality of these relationships to data and 2) follow the rules associated with the relationship. The two important words in that last statement are relationship and rules.

"Non-invasive data governance operates on the premise that all people in the organization that define, produce and/or use data, must be held formally accountable for the quality of these relationships to data and following the rules associated with the relationship."

Relationships to data are directly connected to the activities of a person's job. Following this reasoning, separating the jobs into positions associated with the three actions makes perfect sense.

DATA DEFINERS

Data Architects, Data Modelers, Data Owners, Data and System Integrators, Transformation Leaders, Program and Project Managers, and Business Architects and Analysts and representatives on projects are just a few of the roles that are associated with defining data. These people work diligently to make certain they define data that will meet business requirements. Data governance can help these people.

For data governance programs that focus on improving how the organization defines data, the program must provide guidelines and supervision for how data should be defined. That includes guidelines for the development and enforcement of data standards, business terminology, data models, metadata, and data dictionaries.

DATA PRODUCERS

Data and System Integrators, people that acquire data, and people that take the data they have access to and manipulate that data for their purposes and the purposes of others in any way are just a few of the roles that are associated with producing data. These people work diligently to make certain they produce quality data to meet business requirements and achieve business goals. Data governance can help these people.

For data governance programs that focus on improving how the organization produces data, the program must provide guidelines and supervision for how data should be produced. That includes guidelines for the development and enforcement of data quality, data acquisition and big data management.

DATA USERS

Report Writers, Analysts, Super Users, Data Scientists, and people that use data to answer questions and make decisions at all levels of the organization are just a few of the roles associated with using data. These people work diligently to make certain that they use data to meet their team and corporate needs and requirements. Data governance can help these people too.

For data governance programs that focus on improving how the organization uses its data, the program must provide guidelines, rules and supervision for how data should be used. That includes guidelines for the development and enforcement of data classification, protection, compliance, and regulatory reporting concerns.

APPLYING NON-INVASIVE DATA GOVERNANCE TO THE THREE ACTIONS

There are two basic categories of how to apply data governance to the three actions. There are many methods associated with each category. The two categories are proactive and reactive applications of data governance. Ideally, both will be implemented within an organization.

Proactive data governance involves building the action of governing data into process. One example of proactive data governance is the addition of thorough data considerations into a system development methodology. By inserting data governance focused activities into the steps of a system development methodology, it is assumed that these steps will be completed as the methodology is being followed.

Reactive data governance involves the development of repeatable processes and designated responsibility for specific roles to respond when there is a specific need to resolve a data-related issue. Examples of reactive data governance include the development of a data issue collection and resolution process, the development of process to address requests for access to sensitive data, and process to acquire packages or tools that will enable data capabilities.

USING IDERA TOOLS TO MODEL AND MANAGE DATA

Data governance can only be effectively accomplished in an organization that models their data and processes. It is not a one-time activity; data governance is an on-going initiative that must respond to changes while demonstrating compliance. In order to establish compliance to regulations such as GDPR, HIPAA, SOX, PCI DSS, and others, businesses need to know which data is sensitive and who has access to it, and be able to provide detailed reports on any changes made throughout the data lineage. An organization needs to effectively plan, manage, monitor, and control access to their data, whether sourced internally or externally.

IDERA offers the feature-rich ER/Studio Enterprise Team Edition suite, which includes tools for logical and physical data modeling (Data Architect), business process and conceptual modeling (Business Architect), and a shared model and metadata repository along with a collaborative portal for business glossaries and terms (Team Server). The ER/Studio solution empowers users to easily define models and metadata, track changes made to models and business glossaries, define an enterprise architecture to effectively manage data across the whole organization, and establish a solid foundation for data governance initiatives.

By leveraging collaborative capabilities and simplifying access to data models and glossaries, ER/Studio encourages communication, expedites decision-making, and improves data quality across the organization. Business analysts and architects can define business processes, collaborate with data professionals,

participate in the metadata definition workflow, and access information on models and metadata at the right level for their needs. Data modelers and architects can easily document and share models, metadata, and reports, as well as collaborate with business stakeholders on a unified enterprise glossary with metadata terms and definitions that can be used consistently across multiple database platforms and applications. The ability to share models, metadata, data sources, and glossaries across the organization increases confidence that high value and complex decisions are based on a common understanding of the corporate metadata.

The addition of unstructured data can complicate the data landscape. ER/Studio Data Architect makes this easier by providing native round-trip engineering support for Hadoop Hive tables and MongoDB document stores. MongoDB includes arrays which are not readily viewable in most modeling tools. ER/Studio provides the ability to view these relationships, including nested objects, using a special 'is contained in' notation within an organized model layout. For other big data platforms, ER/Studio leverages the MetaWizard import and export bridges, which provide the ability to integrate numerous data sources including ETL into modeling diagrams.

Data lineage shows the movement of data through the organization. It captures the source of truth as the data moves through the org, and describes the relevant sources, targets, and transformations. ER/Studio can create a diagram to show these transformations within the models, which includes relational, unstructured, and ETL data sources. By incorporating diverse data sources and enabling data lineage to trace the data movement, ER/Studio enables data professionals to effectively document and understand their data landscape and establish a useful enterprise architecture that will enable them to achieve business goals.

WHY DATA MODELING IS A FORM OF DATA GOVERNANCE

The truth is that Data Modeling by itself is not Data Governance – in its entirety. But Data Modeling is a form of Data Governance.

Data modeling is a data discipline. Through that discipline we design our organization's data, reduce redundancy, follow standards, and build business-useful definitions for the data. Data modeling does much more than that. Ask any data modeler of David Hay, Steve Hoberman, Karen Lopez, and Len Silverston's stature. They can tell you the value data modeling brings to the organization much better than I can.

Data modeling can be done well, or … less well. Some data models include redundant "cheeseburger" definitions (What is a cheeseburger? A burger with cheese.) and some have well thought out and validated business descriptions of data that make that data production and usage infinitely more valuable. The use of data modeling varies widely from organization to organization.

"Data modeling can be done well, or ... less well."

Some organizations have Enterprise Data Models (EDM) that are built to design the entirety of data for the organization. Let me write that again for emphasis – design the entirety of data for the organization. Developing the EDM is often a monstrous task that requires the involvement of a plethora of business and technical people discussing the detailed data and information needs of the organization. Some people view the enterprise model as the place to start the improvement of data and data quality across an organization. Other people view the EDM as a step towards defining and addressing the overall data needs of the enterprise. Still others view the development of an EDM a big waste of time (no telling for some people's line of thinking!).

Some organizations model data for their internally developed information systems and/or for the data that resides in their data warehouse or business intelligence environment. Often these models are smaller than an EDM and are built for specific purposes — although many organizations select to reuse components of existing models to create new models.

Other organizations purchase industry data models, follow described patterns for producing data models, and otherwise take immediate steps to acquire and place discipline around the design phase of defining, producing, and using data. Data modeling is, or has been in the past, viewed as the basis of data management activities for the organization. Again, data modeling is all about data discipline.

There are many reasons to create a data model. These reasons include following data standards, reducing redundancy, putting business definition to data, and coming to grips with how to define data better or manage the definition of data as an important asset. There is no doubt that data modeling is both an art and a science, but the primary reason to model data is to instill discipline around defining data for the organization.

"There is no doubt that data modeling is both an art and a science."

Industry definition tells us that data modeling is a process used to define and analyze data requirements needed to support the business processes within the information systems in organizations; the process of data modeling involves professional data modelers working closely with business stakeholders, as well as potential users of the data and information systems.

According to Data Modeler extraordinaire Steve Hoberman, data modeling is the process of learning about the data, and the data model is the result of the data modeling process.

So why do I say that Data Modeling is a form of Data Governance? Data governance is the execution and enforcement of authority over the management of data. Data modeling can be considered the execution and enforcement of authority over the definition of data. The discipline of data modeling involves the "right" people at the "right" time to define the "right" data for the organization. This is the essence of data governance.

Data stewardship is the formalization of accountability for the management of data. If you subscribe to the idea that everybody is a data steward because of their relationship to data (a core tenet of the Non-Invasive Data Governance approach) then certainly the people providing information and assisting the data modelers must also be data definition stewards. And to think, the people that the data modelers work with have been playing the data steward role way longer then the term "data steward" has been trendy.

CONCLUSION

Data Modeling is a form of Data Governance — or at least a piece of Data Governance — because it takes discipline, which is necessary to make certain the design of data is the way it needs to be. Organizations that do not model their data have more difficulties improving the value they get from their data because their data becomes riddled with inconsistency and misunderstanding. Ask any organization that does not model their data if their data is being governed. The answer will surely be "no."

ABOUT THE AUTHOR

Robert (Bob) S. Seiner is the President and Principal of KIK Consulting & Educational Services (KIKConsulting.com) and the Publisher of The Data Administration Newsletter (TDAN.com). In 2017, TDAN.com celebrated its 20th anniversary. In 2016, KIK Consulting celebrated its 15th anniversary focusing on knowledge transfer and consultative mentoring in the areas of Data Governance, Metadata, and Information Quality. Bob was awarded the DAMA Professional Award for significant and demonstrable contributions to the data management industry. Bob specializes in Non-Invasive Data Governance™, data stewardship, and metadata management solutions, and has successfully assisted and mentored many notable organizations.

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