SAS® GLOBAL FORUM 2020 #SASGF

Paper SAS4615-2020

Getting from Governance Practice to Data Awareness

Chris Replogle, SAS Institute Inc.

ABSTRACT

If you want to have a good practice of governance on assets with your organization's employees, then it is vital to put good policies, procedures, measures, and rules in place to manage those assets. Good management of assets alone is not enough in the connected and information-rich world we engage in; we must strive to have good practices for people and processes to follow in order to ensure compliance. These policies and practices need to be fed and managed by good information sources and tools. These assets are typically described or managed with a system such as a catalog. The catalog enables your systems and people to collect and manage information about your information or about the metadata of your systems. This catalog enables you to add information, relate assets, assign responsibilities, and find the information necessary to uphold and implement the policies you designate in your governance practice. The practice of cataloging assets such as data, processes, analytical models, digital locations, and intellectual property gives you the ability to not only inventory the assets, but, through their relationships, you get insight into how they are connected. This connection or relationship among assets can provide you with knowledge in how assets are created, consumed, produced, processed, classified, or are influential within your systems. This knowledge or awareness of your assets leads to more insight into how relevant information is—in other words, you have data awareness.

INTRODUCTION

There are a several considerations impacting organizations in today's digital landscape.

Governance is about the people and processes we use to maintain, onboard, and organize our data. Governance practices define how data is assessed, categorized, organized, and secured. The data has influences such as regulations, policies, compliance, profit and company secrets set the standards we abide by. Those influencers need to be organized into a structure with responsible parties and owning/managing organizations.

Most governance comes down to an organization or unit trying to manage the data they have or ingest. Data may be internal corporate data or from an external entity and it may be leased from another organization, division or group. Regardless of the source, a Data Use Agreement should exist and dictate the usage of that data. Use of data is a privilege and as such should be treated with respect and transparency. We should be driven with this sense of social responsibility when we consider the sharing and use of data.

As we collect data of interest, the data should be cataloged or the information about that data (metadata) including the information related to that data should be captured. This enables us to search, organize, categorize, link and relate, or otherwise collect information such as metrics and quality that are related to that data. Once collected, we need to ensure we can allow for search and discovery of this data. This connected data landscape allows users to be much better informed about the data they have available and interact with.

Data interactions come from applications, tools, processes, and consumers who want to build analytical models, reports, and gain knowledge from data within our organizations.

The consumption of the data is used to solve the business problems we have within our organizations or business units. This the reason for governance around information, the sharing of information should be controlled.

The big take-away here is once you have data on-boarded, you should have an understanding about the lifecycle of the data, and how it should be managed or governed. Setup your management practices, business rules and decisions, metrics and thresholds to allow proper management within your organization. This will allow the organization to define and relate the business terms about the data to the data itself.

This document describes the moving parts and pieces used in governance. This is not a complete how-to on the governance of information, but an introduction to the moving parts of the system an organization should have when implementing a governance practice.

ASSETS

An asset is used to describe any object, element or structure that has value to the organization. These assets are what we need to govern. Examples of assets:

- data sources (databases, file-systems, etc.)
- infrastructure and applications
- APIs and services that interact or provide information
- analytic models, processes, and intellectual knowledge within the organization
- buildings and natural resources
- physical resources that have unique identity

The assets of your organization such as data need to have assigned stewards, teams, or responsible parties to be the champion of that asset and information related to it. These assets are typically described and/or managed with a system such as a catalog.

ASSET MANAGEMENT

The management of assets requires an organization to maintain information about those assets and management events related to changes of those assets. This enables notifications or information updates as assets update or change status.

ASSET LIFECYCLE

The lifecycle of an asset is the management of an asset throughout its usage within the organization.

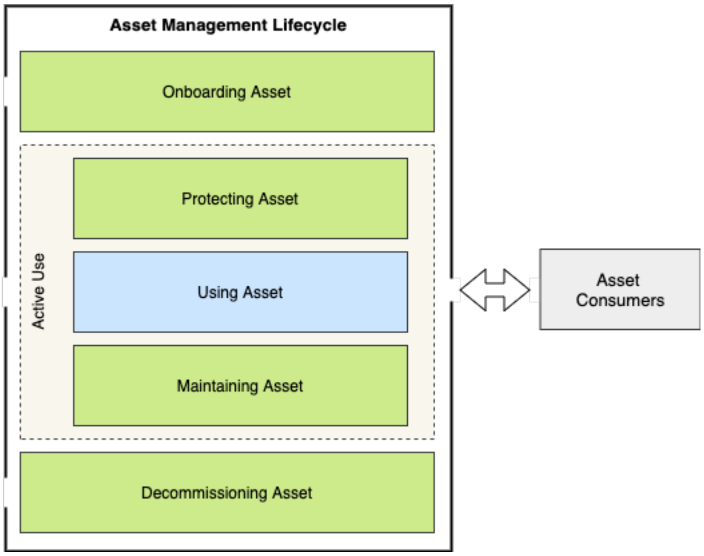

Figure 1: Asset Lifecycle

During its lifecycle, the asset should be tracked from onboarding to decommissioning. The usage agreement for that asset dictates how it should be protected, used, maintained, and even archived or expired from usage.

DESCRIPTIVE INFORMATION

The existence of an asset alone is not enough for good management; your organization will need to add, update, and manage more descriptive information about those assets. Who is responsible for the asset, what is their role? What is the asset related to? What policies apply to this asset? What rules or decisions or thresholds and metrics are used to gauge success with this asset?

The practice of cataloging assets such as data, processes, analytical models, digital locations, and intellectual property enables you to not only create an inventory of the assets, but through their relationships obtain an insight into how they are connected. These connections can be on the business or semantic level, on the operational level, or a conceptual model.

To describe assets on the business or semantic level we usually turn to a tool such as a glossary.

GLOSSARY

A glossary is a collection of common data definitions that are usually focused on defining the meaning of data. A glossary should have a defined owner and organized content to ensure it is accessible to consumers. An organization can have more than one glossary with different definitions.

The components of a glossary are the following:

- terms
- categories
- classifications
- relationships

These components enable your organization to build a vocabulary or common language to communicate about or relate to information or data.

TERMS

Terms are the building blocks of semantics for the glossary. They describe a single concept or idea. These concepts are usually some form of hierarchy and thus most glossaries have hierarchies of terms. This enables the users to define and refine the concept as needed for a specific usage. An example of this is *personal id,* which could be a *government id* or more narrowly defined as a *license id*. By narrowing a concept you can define meaning and apply organization to information.

Semantic assignment is the act of applying a term to information in order to define the meaning or understanding of the information. Assigning meaning enables the business user to understand the IT infrastructure. Understanding also enables the glossary user to find relevant information and understand the meaning of information or data related to a term.

Glossary Categories

Glossary categories are the folders of the glossary that are used to organize terms. There can be a hierarchy of these categories like folders to enable the organization of terms. A term can be assigned to more than one category.

Classifications

Glossary contents can be described further using classifications. A classification adds descriptive information to further describe how something is used. A glossary can have a classification to describe how its content should be used. A category can have classifications such as the following:

- *Subject Area* is the category that describes an important topic area for the organization. In recommended usage, subject areas have owners and are managed carefully.

A term classification assignment can be used to describe how a term is used or how it relates to other terms.

Ontologies & Taxonomies

In most glossaries the organization of terms is usually in a taxonomy or ontology.

“Taxonomies classify. Ontologies specify “(Cagle 2019). This brief but correct statement gives a hint into how and where you should apply each concept. These tools allow you to build the glossaries that are correct for your organization.

TERM RELATIONSHIPS

It is also possible to link two glossary terms together with a relationship. The relationship may describe a semantic relationship or a structural one.

Terms have relationships to each other that help define how they are related or linked together. For example: An *apple* is a *fruit*. Here we apply the "is a" or "isA" relationship to help narrow or more accurately describe the fruit. There are many different relationships for terms that help such as “isA”, “hasA”, synonym, antonym, “seeAlso”, preferred, replacement, and so on, that these relationships provide a rich system for organizing semantics as terms.

GLOSSARY CLASSIFICATIONS

The glossary itself can be classified to denote the organization or structure of content within. Examples are as follows:

- *Taxonomy* means that the same term is not present in more than one of its categories. This is used in glossaries that are designed to provide an organizing structure for other types of information, such as a document library.

- *Canonical Vocabulary* means the glossary only includes terms that have a unique name. Therefore, there is only one definition for any concept.

Category Classifications

A category can be classified to elaborate on the type of content within that category. Here is an example:

- *Subject Area* means the category describes an important topic area for the organization. Typically, subject areas have owners and are managed carefully.

Term Classifications

Term classifications are used to define how a term is used. The following examples are classifications of terms:

- *Activity Description* indicates that the term is a verb, or an activity. Most term definitions are nouns, they describe concepts or things. However, it is useful to be able to define the meanings of particular activities in the glossary. The “Activity Description” classification highlights when a term describes such an activity.
- *Abstract Concept* indicates that the term is an general or not specific area or concept.
- *Data Value* indicates that the term is a valid value for a data item.
- *Context Definition* indicates that the term is a context. Contexts define where a specific definition is used.
- *Spine Object* indicates that the term represents a type of object (such as a person, place, thing).
- *Spine Attribute* indicates that the term represents a type of attribute or data field.
- *Object Identifier* indicates that a term is typically is a type of attribute or data field that is an identifier for an object.

COMMON INFORMATION MODELS

Common Information Models (CIM) allow you to standardize the way you represent information, making it easier to design, deploy, and evolve even the most complex systems.

A recent open example of a CIM is the Cloud Information Model, which defines a standard by which to model, translate, and exchange information across different systems.

COMMON DATA DEFINITIONS

Common data definitions are used to allow for a shared understanding of how to reference and represent data.

Data Classes

A data class provides features to organize your data into logical types. Here is where you would define or relate information for data type detection such as the following:

- rules
- parsing/detection expressions
- regular expressions for type validation

The data class also allows for denoting the preferred implementation types by technology. An example of a data class is social security number (SSN) in the United States. SSN is typically represented in the form XXX-XX-XXXX where X are digits. But this value could have different representations in comma separated files as strings, SAS datasets as numbers with a format applied, or database columns in VARCHAR(9) datatype settings. Data classes indicate what is expected in each representation, but also documents how it should be used. This documentation of usage allows for easy validation and rule creation.

Schemas

A schema is used to document the structure of data, whether that data is in motion in a streaming environment or at rest in a storage structure. The schema type is used to define the elements and allow reuse. The elements of a schema type are defined with schema attribute objects. These schema attributes define the elements or components of the schema type. An example of this is where we have a schema type of relational table, which has schema attributes to represent the columns of the table.

Assets and Schemas

Schemas as described above are related to assets to describe the structure of the asset's information or data. A schema type can be assigned to multiple assets, implying a shared structure but with different data content.

Connections

Assets are accessed through connections which store the information needed to connect to the asset. These connections have connection types that describe the connection method or technology used to make the connection. A connection may have connector type of ODBC or JDBC to connect to a database server, for example.

Discovery

The information about an asset or its metadata should be discovered or brought into a catalog to allow for search, comparison, and information retrieval as needed. This metadata is best discovered through the use of some form of discovery engine. A discovery engine is a process that runs a set of actions to describe the content of an asset. In the case of data, you would want to discover at least the following:

- metadata: who, what, when information of the asset. Example: who created it, what is its size, etc.
- types: What type or data class is this asset?
- ranges: What range of data or information is present?
- metrics and quality: What is the state (missing values, incorrect values, min, mean, etc.) of the data? What is the quality score of the data?

This information is applied either directly to the asset metadata, related data classes, or a related data annotation. Data annotation is where we store the discovered metrics and quality information and associate it with the asset.

GOVERNANCE LANDSCAPE

The landscape of your governance is defined by how your information and practices fit together using the pieces outlined. You can approach this in several ways. One approach is to start with the asset and discover relationships from the asset to connected information. Another approach is to start with the business ideas, define a model, and apply the model to the appropriate assets. The bigger picture here is to connect the people within your organization to the information they need to respond to inquiries efficiently. Here is an overview summary of how the pieces connect:

- Information or data is connected to assets in your catalog. Connection -> asset

- Asset information is defined by a schema. Asset -> schema
- Schema type can be related to a data class for typing. Schema -> data class
- Asset metadata is discovered. Discovery -> asset -> data annotations.
- Schema attribute/data class can have semantic assignment. Term -> Data Class, term -> schema attribute
- Terms can have topology/ontology and/or data modeling. Term -> term.
- Terms are in glossaries. Term -> glossary.

Now that we have covered the pieces or tools to use in your governance practice, we should have some coverage on the why and where governance is important to an organization.

GOVERNANCE AND ORGANIZATION RELATIONSHIPS

Governance is about ethical and effective operation of an organization. The maintenance, usage, and development of assets reflects on the organization itself and how it appears externally. Vision of the organization comes from how it is seen from the outside. The viewers of an organization could be any of the following:

- consumers
- customers
- business Partners

These parties consume goods or services from an organization and provide feedback and/or reward to the organization.

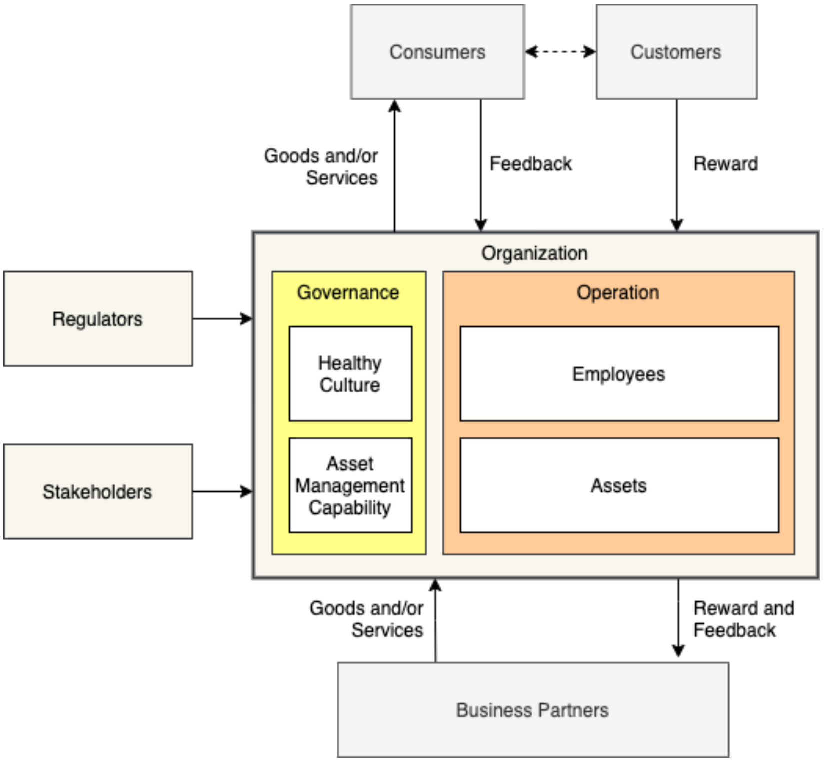

Figure 2: Organizational Influencers

The other set of organizational watchers are ensuring the operations and actions are within expected parameters. Figure 2: Organizational Influencers, also shows regulators and stakeholders who ensure operations are within the expectations, rules, and regulations set forth for that organization. This in-short this defines the compliance standards that must be upheld for social or ethical health along with the value provided to the stakeholders. Stakeholders in this usage could be investors, parent organizations, founders, or committees who want to see their return on investment in assets and/or employees.

MATURITY MODELS

The governance maturity model concept as described by The Journey continues: From Data Lake to Data Driven Organization, describes the levels of maturity in your governance practice. It shows five levels of maturity that build on one another until the organization is able to support a wide range of advanced data use along with appropriate self-service access to data for a majority of people in the organization. In general, organizations are not always at one level of maturity. They adjust their investment to focus on the types of data that will bring the most value.

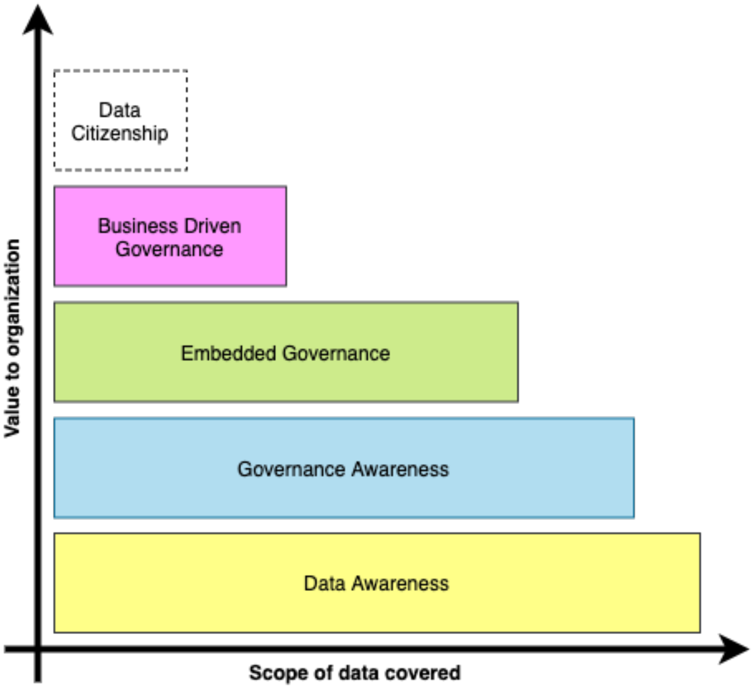

Figure 3: Maturity

The Figure 3: Maturity describes this coverage of maturity and what is more practically implemented. The maturity levels are:

- data awareness - Where is the organization's data and what does it contain?
- governance awareness - How should data be governed?
- embedded governance - How can governance be automated?
- business driven governance - How can the business leaders take ownership of data and governance?
- data citizenship - How can every employee, system and server get the data they need, every day?

These models allow you to realize your stage in growth of your governance practice and where your organization can go from or to from each stage.

CONCLUSION

This document gives a very brief overview of the components or pieces of governance practice and a simple explanation of the moving parts of a system that can be used for your organizations asset governance. The hope is to express the importance and need for a connected system of information that enables you to relate and manage assets within your organization. This document describes the catalogs deployed in most organizations.

If you want to have a good practice for governing assets within of governance on assets within your organization, then it is vital to put into place good policies, procedures, measures, and rules in order to manage those assets. The good management of assets alone is not enough in the connected and information rich world we engage in. We must strive to have good practices for people and processes to follow to ensure compliance. This facilitates good reporting of compliance and safety to regulators and stakeholders. A good company or organization builds confidence with their consumers and/or customers by applying good asset management. These policies and practices need to be feed by good information sources and tools.

These assets are typically described and/or managed with a system such as a catalog. The catalog enables your people to collect and manage data about your information or the metadata of your system(s). This catalog enables you to add information, relate assets, assign responsibilities, and find the information necessary to uphold and implement the policies you designate in your governance practices. The practice of cataloging of assets such as data, processes, analytical models, digital locations, and intellectual property gives you the ability to not only create an inventory of the assets, but through their relationships obtain an insight into how they are connected. This connection or relationship of assets can give you knowledge about how they are created, consumed, produced, processed, classified, or influenced within your systems.

The main justification or reasoning behind governance and security of information or data is the inherit need to share that information or data. The sharing of information should be controlled with your governance practice. The knowledge of your information assets should be cataloged to allow for interaction with systems, consumers, producers and other organizations. This knowledge of your assets leads to more insight into how relevant a collection of information is, it gives you data awareness.

For further information please visit the following:

- ODPi/Egeria guidance on governance for more information on this topic.
- ODPi/Egeria website for more up-to-date information and knowledge sharing.
- ODPi/Egeria project for tools and technology around the practices described here.
- References and Recommended Reading for more detailed information and guidance.

REFERENCES

“Anatomy of a glossary.” May 22, 2019. https://opengovernance.odpi.org/common-data-definitions/anatomy-of-a-glossary.html. Accessed on February 1, 2020.

Cagle, K. “Taxonomies vs. Ontologies.”. March 24, 2019. https://www.forbes.com/sites/cognitiveworld/2019/03/24/taxonomies-vs-ontologies/ Accessed on February 1, 2020.

Cloud Information Model. https://cloudinformationmodel.org/. Accessed on February 1, 2020.

DAMA International. 2017. *DAMA-DMBOK: Data Management Body of Knowledge.* 2d ed. Basking Ridge, NJ: Technics Publications.

"Governing Systems." May 22 2019. https://github.com/odpi/data-governance/tree/master/docs/governing-systems. Accessed on February 1, 2020.

"Open metadata for common data definitions." May 22, 2019. https://opengovernance.odpi.org/common-data-definitions/open-metadata-for-common-definitions.html. Accessed on February 1, 2020.

"ODPi/Egeria Guidance on Governance." May 23, 2019. https://opengovernance.odpi.org. Accessed on February 1, 2020.

"ODPi/Egeria Project." https://egeria.odpi.org Accessed February 20, 2020.

Sebastian-Coleman, Laura. 2018. *Navigating the Labyrinth: An Executive Guide to Data Management*. Basking Ridge, NJ: Technics Publications.

"The Journey continues: From Data Lake to Data Driven Organization". February 19, 2018. http://www.redbooks.ibm.com/abstracts/redp5486.html?Open. Accessed Feburary 1, 2020.

"What is an asset?" https://egeria.odpi.org/open-metadata-implementation/access-services/docs/concepts/assets/. Accessed on February 1, 2020.

ACKNOWLEDGMENTS

I would like to thank Mandy Chessell (IBM) and Ferd Scheepers (ING) and their teams for all the effort and dedication put into the ODPi/Egeria projects and the ODPi/Guidance on Governance. This was not possible without the ODPi/Egeria teams from ING, IBM, and SAS for laying the groundwork for this information and thought guidance.

RECOMMENDED READING

- *The Journey Continues: From Data Lake to Data-Driven Organization (February 19, 2018)*
- *Designing and Operating a Data Reservoir (May 26, 2015)*
- *Governing and Managing Big Data for Analytics and Decision Makers (August 26, 2014)*
- *Common Information Models for an Open, Analytical, and Agile World. (April 8, 2015)*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chris Replogle
SAS Institute Inc.
Chris.Replogle@sas.com