SPONSORED BY
Sandhill Consultants

DATA GOVERNANCE ENABLES DATA MANAGEMENT

database
TRENDS AND APPLICATIONS

HARNESSING THE POWER OF MODERN DATA CATALOGS

Best Practices Series

database
TRENDS AND APPLICATIONS

ACHIEVING DATA-DRIVEN INSIGHT WITH MODERN DATA CATALOGS

Best Practices Series

All companies today seek to become data-driven enterprises. However, such efforts can be hobbled by outdated, incompatible, and discrete data environments. Many organizations' data assets are hidden away in silos or proprietary applications, which can take great amounts of time and resources to locate. This is made more complicated as the amount of data flowing through, into, and out of enterprises keeps growing exponentially. The challenge is to identify, document, and transport the data sources and assets that are meaningful to the success of the business and make the data available to users and specialists at the touch of a button.

Access to data from any platform can be cumbersome. A recent survey found that traditional data lake platforms, for example, require too much data transformation upfront. Time spent moving, migrating, pipelining, or transforming data increased to 7.1 hours per week for respondents who have a data lake. In addition, 30% of respondents indicated that their end-user consumption/visualization tools aren't directly connected to the lake, resulting in data duplication and data movement challenges that limit insights and time to value.

Increasing the difficulty of data access are opaque or rigid corporate cultures that make data sharing difficult. With information flowing through enterprises at an unprecedented rate from an abundance of sources—customer interactions, employee and partner communications, social media platforms, technical documentation, and more—the process of making it accessible to decision makers can be daunting. Information is disconnected and scattered within and outside of organizations—among databases, enterprise resource planning systems, and customer relationship management systems, to name a few. These tend to be managed by different units and teams across enterprises.

Data catalogs enable self-service analytics and data lake modernization, as well as support data governance, privacy, and security initiatives.

In addition to these challenges, there is the unrelenting push into digital transformation, which can be accompanied by layers of systems and channels on top of what may be a dysfunctional and disparate data infrastructure. At the average enterprise, data assets are spread across different departments, applications, systems, and geographies—including onsite data centers and public clouds. This heterogeneity mirrors the increasingly complex and diverse nature of the data landscape. It also intensifies the challenge of finding, inventorying, and analyzing data assets.

GAINING DATA-DRIVEN INSIGHT

To address these obstacles to data-driven insight, more enterprises are turning to data catalogs as a key

component of their modern data strategies. Data catalogs are services comprised of searchable metadata which represents available data across the enterprise. These services are playing a vital role in data governance efforts and provide rapid visibility and faster time to value. The key to enabling access to information as it is needed is to implement data catalogs as part of a data infrastructure.

Interest in data catalogs is high, a recent survey of data managers by Eckerson Group found. Almost all respondents (94%) have a data catalog in use, or plan to deploy one. However, at this stage, only 10% report they are fully deployed. Another 41% report having a data catalog partially deployed. Only 6% have no plans to implement a data catalog due to a lack of budget or research to make a commitment.

Data catalogs enable self-service analytics and data lake modernization, as well as support data governance, privacy, and security initiatives. Many aspects of building a data catalog can be automated, which is a major benefit offered by commercial modern data catalog platforms.

The biggest reason for adopting data catalogs, the Eckerson survey revealed, was to "support data discovery" (72%) and "improve users' ability to find and access data" (66%), which are essentially equivalent. These were followed by the ability to "improve data curation and governance" (68%) and "support our self-service initiative" (60%).

Data catalogs are becoming essential tools for data professionals across all categories of staff, including data engineers and data scientists, as well as line-of-business users. Access to a data catalog can provide quick answers to the location of essential information assets, such as transaction histories or geolocation data, as well as provide visibility into data management issues such as compliance, security, or retention policies.

BEST PRACTICES

The following are best practices for making data catalogs work for the enterprise:

- **Start small:** Zero in on a specific business area and build a catalog that points to relevant data sources in demand by data analysts and users within the domain. As catalog adoption scales upward, this model can be applied to additional domains.
- **Focus on self-service:** One of the most compelling use cases for data catalogs is their ability to support self-service initiatives and associated analytics.

Many aspects of building a data catalog can be automated, which is a major benefit offered by commercial modern data catalog platforms.

Within today's blizzard of data, finding the right information is similar to looking for a needle in a haystack. Once that needle of data is found, it has to be put in its proper business context. A data catalog should incorporate metadata on the location of the data within the data store or environment.

- **Align with governance, compliance, and legal mandates:** Data governance is no longer just an internal corporate matter—government agencies across the globe are concerned with how data, particularly about individuals, is being managed. Organizations must now show places from which they are extracting data and how it is being handled. Users and compliance managers alike need to understand its lineage—its origins and place in the information chain. In addition, it's important to be able to view changes made to data as it moves from one portion of the enterprise to another and who has had access to the data.
- **Provide for easy search and discovery:** There are many data catalog tools and platforms on the market, which can make things confusing if there are a multitude of these within a single enterprise serving many data environments. The goal is to make searching for the right data as quick and painless as possible. The data highlighted in the catalog may be business metadata or technical metadata serving business users or data specialists involved in building the infrastructure.
- **Keep expanding your variety of data sources:** The business data landscape keeps changing, sometimes from day to day. It is important to be able to incorporate new sources as they become available and others become outdated. This data may be from local storage, cloud-based services, or from edge environments. It may be structured data from SQL-compliant databases, or it may be unstructured data, including content such as graphics or audio files. The sky's the limit.

With tremendous volumes of data moving through enterprises, AI and machine learning can be leveraged as a resource to help users rapidly identify and access pertinent data sources. The key is to eliminate or minimize as many time-consuming tasks as possible in collecting, cataloging, and securing the data.

A data catalog is an essential service for today's ever-expanding data environments. In the process of better data utilization, business leaders and users will gain a deeper appreciation for the data capabilities available to them. ■

—Joe McKendrick

Data Governance Enables Data Management

Data cataloging and data governance have different objectives, but the goal of both is to empower data creators and users to understand, use, and trust their data. Data catalogs promote collaboration across IT and different departments to establish enterprise-wide agreement on data which is critical to creating an integrated data catalog. There are clear connections between building a data catalog and data governance. A data catalog is a core component of data governance. Data cataloging falls within the domain of Data Management. Data management covers distinct functions, enabling a business to extract value from data it collects. Data governance identifies the data owners responsible for ensuring data quality, regulatory compliance, and appropriate data usage. Data governance also identifies data users who are required to follow established governance policies, processes and standards. In a way, Data Governance enables the management of data in a orderly well managed way.

GOVERNING DATA

Data Governance is not a software product that you can purchase. Generally, when you begin to build your data catalog you first need to have a few things in place like data policies, processes, and people to support the technology. Most of the time the data is readily available from the existing data stores, but the governance is missing. The biggest investment in creating a sustainable data catalog is the effort needed to collect all the governance artifacts to populate the knowledgebase that ensures the data remains trustworthy and protected.

How do you minimize the time it takes to get your data catalog operational? Is there a jump-start? It has been stated many times by governance professionals that data governance must be custom tailored to your organization. There is no one size fits all solution to Data Governance. All of this is definitely true but there are some fundamental components that you will find in almost all data governance initiatives. Having a foundation of solid content in which to build upon can shorten the cycle from purchase to productivity. The following is a short description of some content to govern your data catalog initiative.

GOVERNANCE POLICY

Data governance polices are at the heart of data governance. Policies document intentions about data ensuring that an organization's data and information assets are managed consistently and used properly. Policies involving data may be numerous and dependent upon the industry being governed but there are some fundamental policies that should be part of any data governance initiative.

The first and most obvious policy is one that states that data must be governed. That is, strategic and effective decisions regarding the organization's data and information assets must be accountable to some governmental body.

The second is to protect the organization's data assets through security measures that assure the proper use of the data when accessed. This policy ensures that employees have appropriate access to organizational data and information while not interfering with the day-to-day activities of the organization's business.

The third fundamental policy involves data usage. This policy is meant to ensure that a company's data is not misused, used ethically and with consideration for individual privacy. While the data access policy documents who can access data, this policy documents what one can do with data based on security levels.

The last policy is related to data quality and the ability to use data across functional business units and information systems. This policy is to

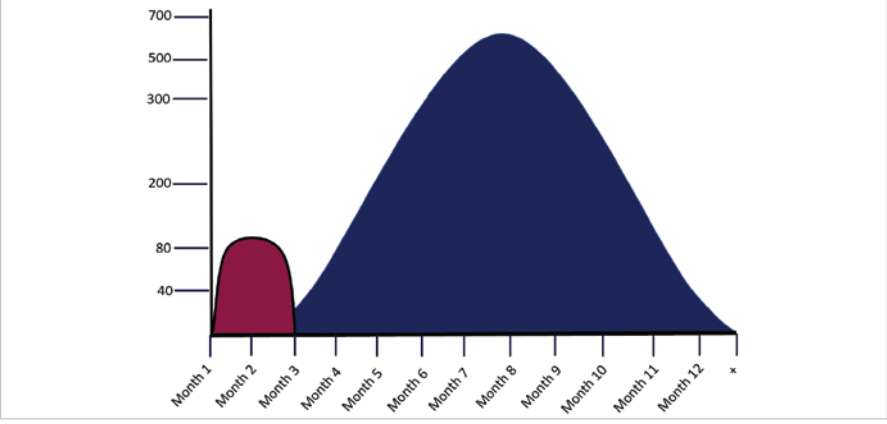

Figure 1: Hours/per Month (Calculation based on Sandhill assessments.) in savings of leveraging a systemized application such as a data governance management application (COMPASS) vs. the traditional approach

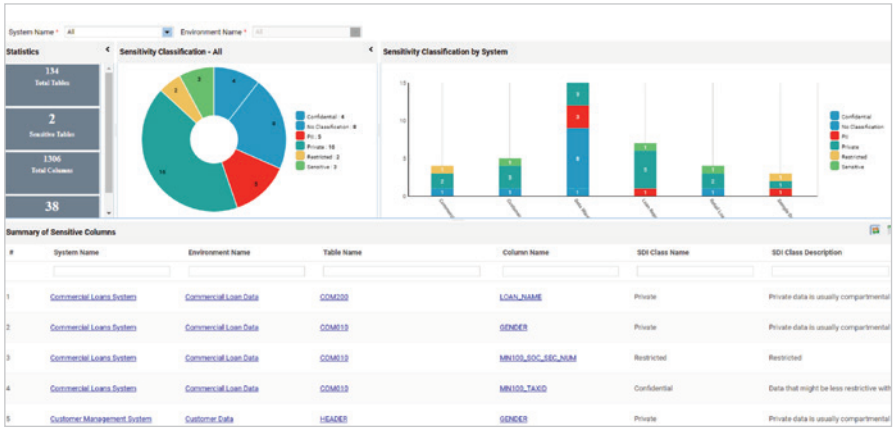

Figure 2: Data Sensitivity Dashboard, erwin Data Intelligence Suite

ensure that the organization's data has good quality and that key data elements can be integrated into the business so that staff, contractors and management can rely on data for information and decision support.

DATA GOVERNANCE PROCESS

Like policies, there are some basic processes that all data governance programs should have. Polices will document what data governance activities must be done, the processes describe how to perform the activities. The two basic processes describe how decisions are made and how conflicts are resolved. Without these processes there is no actual governance going on. Inevitably, there will be change within the data governance program because we are moving from a state of ungoverned data to a state of governed data. It makes sense to understand what we do when data does not conform to what is expected before we encounter it.

DATA GOVERNANCE STANDARDS

How do we know if data conforms or does not conform? This is where standards come into play. Standards tell us what is expected. Often organizations will conflate policy with standards but the two are distinctly different. There are two basic data standards. The first is what is the data name and the second is what does the data element mean? When speaking about data names we are concerned with the business name and the technical name. Names must be consistent and conform to some standard. Names are how you find things in data catalog. A very popular naming standard references the ISO metadata standard 11179 series. In addition to its name data must have quality and security specifications. This allows the IT organization to assess the suitability of the data for use by the business and promotes integration.

What data means is very important to the business. When it comes to key data elements there cannot be confusion about what is the definition of customer or product or other elements that might appear on a report. While there is an ISO standard for names, there is no official standard for data definitions. There are some universal standards like no tautologies such as, a cheeseburger is a burger with cheese or the customer name is the name of the customer. In addition, there are component parts of a definition which can standardize the construction of a definition such as kind-of, characteristics, or usage. For example, an order is a kind of contract between parties that has a specified set of terms and conditions used for selling or purchasing goods.

DATA GOVERNANCE ORGANIZATION

Governance policies document what must be done, processes document how it is done, standards document conformance, and the governance organization documents who is accountable. Without accountability there is no method to ensure that governance is being performed. We can state that data is an asset to be governed, must meet conformance standards, adhere to a process flow but someone must do the work of governance.
The structure of a data governance organization generally follows a tiered pyramidal approach. The top tier sets the policy in accordance with business objectives and is the final decision maker. The middle tier executes the policies by transforming policy into

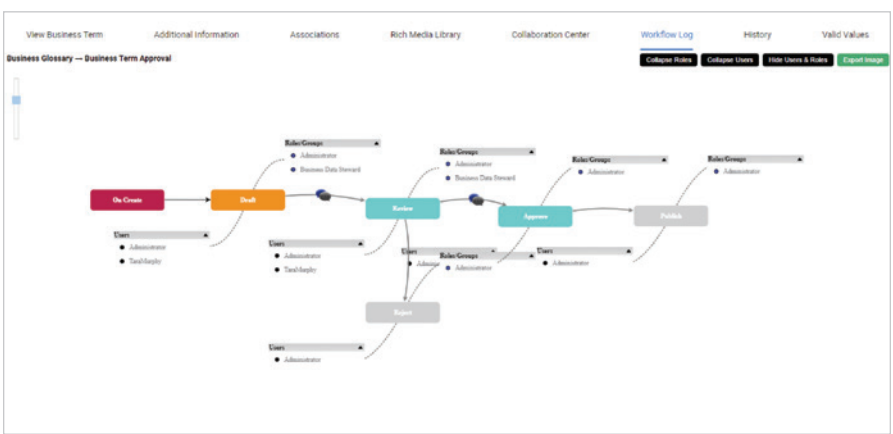

Figure 3: Workflow Manager, erwin Data Intelligence Suite

actionable work with measurable standards. The bottom tier is the actual work of data oversight. Common tier names for a data governance authority include Data Governance Council, Data Governance Committee, and Stewardship among others. The tiers are composed of roles with varying levels of accountability based on the context of the work being performed. Some names for the roles include Data Steward, Data Owner, and Data Custodian. Accountability levels are usually expressed as RACI. This translates to Responsible, Accountable, Consulted, and Informed. The role of Data Steward is responsible for the day-to-day oversight of data. Stewards check to see if the data conforms to quality standards. If change is required, the process is in place to hand-off to an accountable party. It may escalate all the way up to the top tier. Accountability is dependent on the data, the change required, or the issue to be resolved. A RACI chart helps to sort out who is accountable in what context.

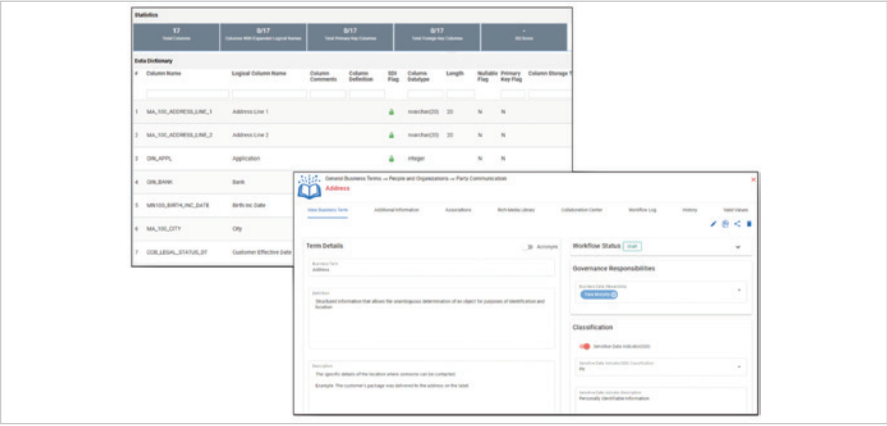

Figure 4: Data Dictionary and Business Glossary, erwin Data Intelligence Suite

A DATA GOVERNANCE MANAGEMENT SYSTEM

Much of the Data Governance components previously described can be operationalized though the use of a Data Governance Management System. The application of a data governance management system for the implementation and deployment of Data Governance accelerates the achievement of the data governance critical success factors by providing the following:

- Knowledge base content and templates in the functional areas of data governance, including policies, processes, standards and organization
- Predefined data governance roles, responsibilities, and accountabilities
- Predefined data governance process workflows
- Roadmap, plan, and metrics for data governance deployment

Having a prebuilt data governance knowledge base prevents the uncertainty of where to start or borrowing from different popular sources that don't fit together. The knowledge base supports the numerous staff resources and stakeholders who are involved in the Data Governance activities. It provides expertise all in one source, with a common vocabulary and definitions, improving data literacy across the organization. It accelerates time to deployment, improves consistency, and decreases chance of error. Prebuilt templates for the Data Governance executive charter, policies, and processes and their metadata are included to provide time-savings, consistency, and the foundations of Data Governance metrics.

Given the accelerator capabilities of a Data Governance Management System with prebuilt content and processes, a general estimation suggests that an organization could save hundreds of hours and resources, compress the time-to-solution by many months, achieving time-to-value in a fraction of the cost of developing Data Governance traditionally.

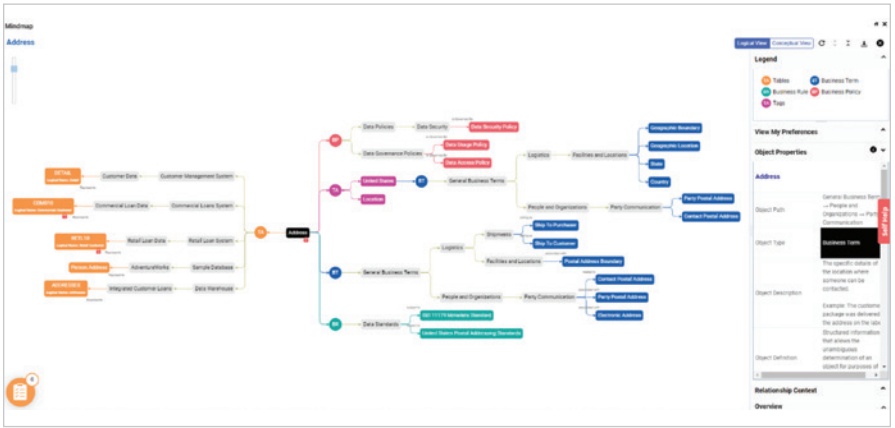

Figure 5: Broad view of lineage from technical to business contexts, erwin Data Intelligence Suite

erwin Data Intelligence Suite

Let's look at how erwin Data Intelligence Suite (DIS) can provide the functionality required to support data governance. Governance provides the oversight needed to ensure a certain level of data quality, findability,

accountability, and meaning that makes data a valuable asset to the business. erwin DIS provides the means to execute data management activities. That is, it does the heavy lifting by providing a centralized listing of an organization's available data elements. Management activities include:

- Metadata Classification
- Change Management
- Search and Discovery
- Lineage

METADATA CLASSIFICATION

Cataloging data requires you to organize it so that data can be structured in a way so that it can be searched, secured, and understood. One of the fundamental policies discussed earlier was the policy ensuring data security. erwin DIS can enable this policy by assigning a security level to certain metadata while visualizing on a dashboard (see figure 2). In addition, tags can be assigned to data to provide a way for subject matter experts and knowledge workers to contribute business knowledge in the form of user-defined associations.

CHANGE MANAGEMENT

Change is an inevitable part of managing a data catalog. As business changes so do the policies, standards, and processes that govern the data. The important thing to understand about change is that change must be governed. If change is not managed the data quality and dependability cannot be assured. Your data catalog will become just another data silo in the ocean of data silos you already have. Data governance describes the process for how change is managed but does not enforce its execution. erwin DIS contains a workflow engine that allows a data governance authority to assign a particular workflow to a data governance role, like Data Steward, so that an auditable trail of decisions can be monitored against the policies and standards (see figure 3).

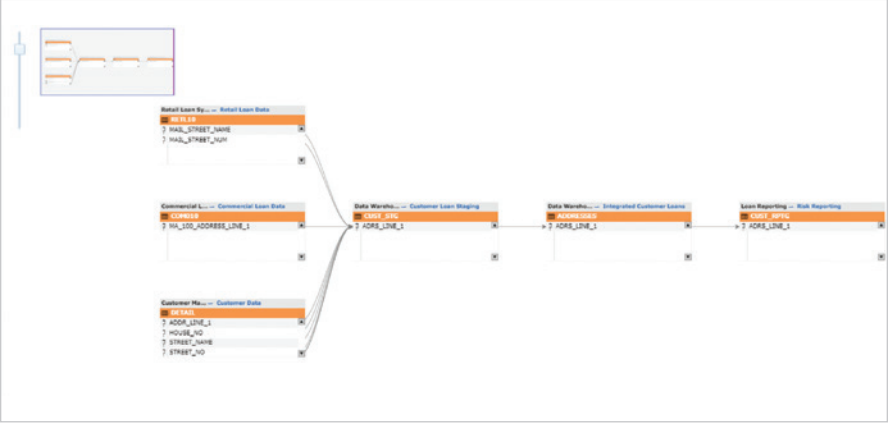

Figure 6: Narrow view of lineage showing technical ETL flow, erwin Data Intelligence Suite

This is one of the primary methods of bringing ungoverned data into a governed state and thus improve quality, consistency, and trust

SEARCH AND DISCOVERY

erwin DIS has flexible searching and filtering options to allow users to quickly find relevant sets of data or browse metadata based on a hierarchy of data assets. Having consistent data standards facilitates the search and discovery process. This is when naming data consistently and providing definitions pays off. When perusing the data dictionary looking for 'address' you may find it abbreviated as ADDR, ADRS, or ADDRESS. If you're only looking for ADDR you will miss ADRS. In addition, automating the harvesting, scanning, and mapping of metadata is made easier as well. Having a single approved abbreviation and definition goes a long way toward trusting in the catalog (see figure 4).

LINEAGE

One of the most powerful features in DIS is the ability to associate related artifacts together and visualize the connections. Lineage can be broad or narrow and within different contexts. Lineage is especially useful in understanding the impact change will have on data assets. From a governance perspective, it is useful to know what policy is related to which standard that affects the privacy classification of data assets within the data catalog. In the diagram below, the left side represent the technical data stores for 'address' in black and the right side represents the business representation (see figure 5). From a detailed technical context, it is useful to understand how data changes from where it is collected to where it gets distributed (see figure 6).

DATA GOVERNANCE EMPOWERS THE DATA CATALOG

The Data Catalog is powerful tool to enable your Data Governance program to contribute to and interact with an inventory of metadata about the data definition, production, and usage of data assets. Data can only be considered an asset if it is actively governed. The true potential of data catalogs cannot be realized without first having a plan about the who, what, where, when and how of the data to be managed.

For more information on Sandhills approach to delivering Data Governance Management Systems (COMPASS) or erwin Data Intelligence Suite, please contact info@sandhillconsulants.com. ■

SANDHILL CONSULTANTS
www.sandhillconsultants.com