Data Governance Preparedness: Using Data Catalogs to Ensure Your Organization Has the Relevant Insights to Weather Any Crisis

Prepared by:

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

Sponsored by:

Introduction

The COVID-19 pandemic caught global organizations of all sizes and industries by surprise. Many have suffered dramatically, some closing their operations permanently, while others have been able to adapt but not without challenges.

In times of crisis, sound decision-making is paramount to short-term survival as well as long-term resiliency and it is reliant on timely and accurate data. Determining how to respond to the coronavirus itself is a prime example of data's important role in shaping not only public health policies but also business processes. Transform or die is not hyperbole when it comes to the health of individuals or companies, so data literacy and intelligence is more critical than ever before.

There are obvious examples of business transformations necessitated by COVID-19. For example, in the healthcare industry, medical practices, hospitals, and health departments had to implement coronavirus testing and solidify their processes for reporting to local and state agencies. Pharmaceutical companies were expected to rapidly scale up their research and development and clinical trials all while adhering to strict protocols for carrying out their work, protecting patient privacy, and aggregating and reporting all the relevant data to various government agencies. And in the insurance industry, well-known companies are voluntarily offering discounts and/or refunds to drivers who have been quarantining at home, off the roads, so there have been fewer accidents or other auto claims to process.

All of these business decisions have been made and implemented with *data-driven insights*.

Now as the rest of the world tries to transition from a state of emergency response to disaster recovery, organizations must attempt to understand how they need to operate in a post-COVID world. From a strategic perspective, they need to know which processes and systems are key to business continuity so they can emerge from a pandemic, a data breach or any sort of major disruption that threatens their workforce, operational effectiveness or regulatory compliance. These three points should be accepted and addressed as part of a successful data governance and intelligence strategy:

1. Situations can erupt unexpectedly, which require rapid adjustments to ensure business continuity.
2. Changes within the business environment, either planned or in response to an unforeseen disruption, require awareness, access and protection of critical data assets.
3. When the need for distributed operations and remote collaboration increases to accomplish the organization's tasks, awareness of and access to relevant data assets will speed the ability to pivot quickly.

Business transformation has to be based on accurate data assets within the right context, so organizations have a reliable source of truth on which to base their decisions so they can weather a future emergency and streamline the recovery process as well as respond to market forces. An agile, intuitive, robust data governance capability with business user friendly visualizations paves the way for evaluating an organization in its current state and then evolving it to serve new objectives. To produce such data intelligence, data practitioners can coordinate automated processes to survey the data

landscape and collect, accumulate, document, collate and publish critical information about the data assets available for use across the enterprise. And organizations can minimize business continuity risks by migrating away from on-premise environments to hybrid cloud platforms that provide democratized data access, inherent redundancy and high availability.

A data catalog is a key technology enabling enterprise-wide data governance and intelligence. But unfortunately, the increasing pervasiveness of vendors that identify their tools as "data catalogs" has only led to confusion. The goal of this paper is to explore the types of capabilities a modern data catalog provides, differentiating needs based on specific use cases and offering clarity when considering data catalog requirements.

Perceptions of Data Catalog Tools

Depending on the context and use cases, there are different perceptions about what a "data catalog" is and how a particular type of tool supports organizational data governance. Reviewing the different use cases will determine the capabilities required in a data catalog tool to support data governance and the subsequent intelligence it delivers. Some examples include:

- **Technical metadata management.** This type of data catalog is used to capture structural information about the schemas associated with data assets to facilitate data extraction, transformation and interchange. This type of tool is used by developers needing knowledge about the data to be extracted from various sources and transformed for loading into a target environment such as a data mart of data warehouse.
- **Data lineage.** A data lineage data catalog documents the data creation process by reverse-engineering data movement processes (such as data integration, ETL or ELT procedures) and business intelligence, reporting and analytics end points as a foundation for inferring data lineage and for impact discovery and analysis. Data stewards use the resulting inventory of data pipelines for instituting data controls, tracking emerging data quality issues, and analyzing how modifications to incoming data sources might potentially impact downstream data consumers.
- **Machine learning data asset inventory**. This style of data catalog lists the data assets deemed important to the organization and contains information about where the data assets are stored and managed plus aspects of data ownership and stewardship. This type of data catalog enables downstream data analysis by surfacing the data assets that are most appropriate for certain types of reports, dashboards and analytics. It allows data consumers to look for data assets that best meet specific needs in relation to specific search terms, coupled with analysis of usage patterns for the collection of available data sets. Individual data consumers sharing their data use experiences increases data awareness by leveraging collected knowledge, allowing the data catalog to subsequently recommend data sets that are best suited to different use cases.
- **Data portal.** A data portal allows users to peruse available data sets, review their data layouts, and in some cases browse a selection of records. These types of data catalogs are popular for open data publication and used by national, state and local governments as part of mandated

data transparency activities. However, any organization can benefit from simplifying data consumer visibility into the enterprise data landscape.

- **Data governance.** A data catalog intended to support data stewardship typically is used as a repository and control framework for data quality policies and rules and integration of data validation controls. This combination of capabilities allows data stewards to operationalize data policy management and is used frequently for managing data quality, data auditing and ensuring compliance with defined data policies.
- **Data security and protection. A** data catalog used for managing data protection policies enables data owners to define and manage data protection policies. Data owners can characterize and classify the content of their data assets, define the roles and groups, and specify data protection directives. An implementation framework can deploy these rules as access controls preventing unauthorized access to sensitive information. These types of data catalogs allow data security teams to operationalize data access control and log attempts of unauthorized access.

While each of these data catalogs is used for a slightly different purpose, they all overlap somewhat in two ways. First, they are all fundamentally dependent on metadata, although each data catalog type uses a subset of metadata in its own way. Second, the information used to support any of these activities can be leveraged to support the other activities. To best understand how metadata aligns with data catalogs, it is worth reviewing the different types of metadata relevant to the data asset lifecycle.

Metadata Classification

Confusion about the nature and value of metadata has, in some ways, prevented organizations from truly recognizing its importance. One challenge is that different people have different perceptions of what constitutes “metadata,” so the organization ends up with incomplete visibility into the data landscape. This broader classification scheme for the different types of metadata provides a more complete picture of the information that comprises organizational data intelligence:

- **Structure metadata**. Most data practitioners are familiar with structure metadata, which includes information about the organization of the data and is most often associated with structured and semi-structured data assets, including database tables, CSV (comma-separated values) files, XML and JSON objects, and spreadsheets and similar data sets. Structure metadata documents what the data looks like, including the names of the data elements mapped to columns, t definitions of data element names, the types of values represented by each data element (such as “integer,” “numeric” or “string”), the sizes/lengths of each data element, and the file layout (such as whether values are organized in a fixed field size or whether they are segregated by field separators, and if s what field separator is used). Structure metadata also includes tags, keys that uniquely identify items in the data set, foreign keys, as well as value domains from which data elements draw their values.
- **Supplier metadata.** As the enterprise data landscape expands, it becomes more difficult to track original data sources. Supplier metadata is information associated with the sources and providers of organizational data assets. It documents a data asset’s origination point along with

any directives and constraints that accompany the data asset prior to its use. Supplier metadata identifies the data owner, the origination point of the data asset (i.e., was it produced internally or was it acquired from an external source), as well as license details regarding the number of consumers and whether there are any data consumption requirements. Supplier metadata also might encompass demographic information about the data asset, such as the number of records (if the data asset is structured), size in bytes, the date the data asset was produced, and possibly information about the original sources from which the data asset was created.

- **Processing metadata.** Related to supplier data, processing metadata describes the production processes of the data in the data asset. It documents the data lineage, which details the processing stages through which the data asset is created, including the extraction process (if the data set is extracted from other data sources), the set of transformations applied, the formulas for derived data elements (such as calculated aggregate values), and the processing flow manifested in terms of data pipelines.
- **Search metadata.** When data consumers want to find the right information for their business processes, a means of classification and tagging helps narrow the scope and simplify the exploration and discovery process, requiring metadata describing each data asset's content accompanied by corresponding content classifications (such as "private personal data" or "data about products"). Search metadata details information associated with the content and classification of the data asset and incorporates a business glossary listing the business terms and their definitions. Importantly, search metadata also includes meta tags and classification taxonomies and ontologies that can be used to build a semantic index searchable by a variety of terms. This type of metadata also may incorporate historical usage data that tracks the types of queries data consumers perform and how they select and subsequently use selected data sets.
- **Actor metadata.** It is critical to understand which parties are accessing corporate data assets, showing the need for actor metadata that is associated with the individuals or systems that touch, manipulate or consume the information contained within the data asset. Actor metadata includes information about data consumers the groups to which they belong, and the types of roles they play. In addition, actor metadata lists the data owners and data stewards tasked with overseeing the quality and usability of the data asset.
- **Governance metadata**. Managing oversight suggests a need for information about the obligations and governance of the data asset. Governance metadata incorporates assertions and policies for data validity and data quality, as well as the policies used to implement data protections, manage access and use, and observe obligations assigned to the data set. For example, supplier metadata such as data licensing might limit the number of simultaneous data consumers, and data controls implemented using governance metadata ensure licensing constraints are obeyed.

Aligning Metadata with Data Catalog Type

By reviewing the different types of metadata, you can see how different data catalog tools rely on the collection and use of a combination of metadata categories:

- **Technical metadata management catalog**. A technical metadata management data catalog captures and provides structural information about source and target data for integration development and ETL. This data catalog mostly relies on structural metadata such as data element names, data types, and data element sizes, supplier metadata such as data asset demographic information, processing metadata including data transformations and data derivations, and aspects of search metadata such as a business glossary and data element definitions.
- **Data lineage tool**. A data lineage tool combines supplier metadata such as the data owners along with the details of the original sources from which the data asset is manufactured. It also captures the data production details with the data transformations, data derivations and the structure of the data processing pipelines from the processing metadata.
- **Machine learning data catalog**. A machine learning data asset inventory blends the practical aspects of the production of the data asset from both the structure metadata (data element names, lengths, types), the processing metadata (data transformation, derivations, and pipeline process maps), and search metadata including the semantic details and historical usage to produce a searchable data catalog.
- **Data portal**. The objective of a data portal is transparency, and a data portal typically scans and then previews accessible data assets. To enable this, data portals combine structure metadata, supplier metadata and search metadata to provide a listing of available data assets, data element metadata, information about the different data sources, and data asset demographics such as number of records or size in bytes. It also provides a means for browsing a subset of data instances within the data asset.
- **Data governance tool**. Data governance tools ensure data usability by monitoring data quality and alerting data stewards when issues emerge. These tools have evolved from metadata repository products to incorporate the definition of data quality policies and support operational data stewardship processes and procedures.
- **Data security and protection catalog**. These types of data catalogs draw on actor metadata to collect information about the different users, groups and roles, different classifications pulled from the search metadata, and data protection directives from governance metadata to enable the definition and implementation of runtime data protection and security policies.

No single solution is limited in the type of metadata consumed and utilized, and no single tool's capabilities satisfy the breadth of need for a data catalog solution.

Reframing the Data Catalog *Solution* for Enterprise Data Governance and Intelligence

We can learn some key lessons about enterprise data governance and intelligence from the sudden retreat into a distributed virtual working environment triggered by the coronavirus pandemic. In a post-COVID world that hungers for data to support both ongoing operations and agile analytics, data consumers are enabled through data governance infused with intelligence that relies on a holistic

collection of data catalog capabilities that cut horizontally across the different data catalog types this paper describes.

Data catalogs provide data intelligence enabling the organization to react rapidly to crises, particularly being able to maintain authorized access to shared enterprise data assets. The data catalog can help maintain data awareness among a physically distributed community of data producers and consumers by level-setting knowledge among the data communities about the content of shared data assets. At the same time, data catalogs provide confidence in the continued availability and accessibility of data while also ensuring that sensitive data is protected from unauthorized access.

Yet the breadth of the data catalog tools market poses some interesting challenges in establishing the right data catalog *solution*. Not all tools fit all use cases and environments, and in some cases a combination of capabilities will be needed.

Identifying the right data catalog solution requires attention to the organization's most critical operational use cases and requirements, such as:

- The degree to which enterprise-wide **business glossaries and data definitions** are expected to be published and shared
- Using **metadata standards** and defined procedures for collecting, documenting and sharing the different classes of metadata
- Inferring data models and lineage through **reverse-engineering**
- Attentive **data curation** that establishes standardized processes for data asset configuration and preparation
- Simplifying **intelligent searching** so data consumers quickly can find what they need and enabling **data previewing** for those seeking data assets to answer ongoing and emerging business questions
- Engineering, implementing and monitoring **data pipelines** and the processing stages through which data streams for end-user reporting and analytics
- **Operational data governance** and assessing existing data governance and stewardship roles and their responsibilities
- **Data validation and quality assurance** for data trust
- **Collaboration** among data producers and different data consumers
- Data content **classification** and how it relates to **data organization** and **data protection**

Steps in Evaluating Data Catalog Solutions

The data governance and intelligence enabled by data catalogs ensures organizational information preparedness. And while not every data catalog tool necessarily fits all enterprise needs, establishing information preparedness to help react to and recover from a crisis demands forethought in instituting a data governance and intelligence framework. Therefore, if you are considering a data catalog solution, you can take the following steps in your evaluation:

1. **Classify data members of the data stakeholder communities:** A data catalog's value directly corresponds to how well it meets the needs of different data communities. Identify who the data producers and data consumers are and classify them according to their specific needs in terms of producing and/or consuming the different kinds of metadata.
2. **Define use cases**: Sit with the data consumers and monitor how they look for data sets, determine which ones are best suited to their needs, and the processes they employ to use data. Defining these use cases will help identify the data intelligence capabilities most desired in a data catalog solution.
3. **Enumerate and harmonize expectations**: Each data community has expectations for producing, publishing, seeking and using data. Engage the different data producers and consumers and discuss the defined use cases to solicit their expectations for data intelligence and how they relate to the data catalog solution features.
4. **Prioritize capabilities**: Collecting and harmonizing expectations will surface repeated requests for desired characteristics, and that will inform organizational priorities.
5. **Establish evaluation criteria**: As desired capabilities emerge, consider how a vendor's product will be assessed in terms of supporting those capabilities. Establish measures and metrics for assessment and evaluation.
6. **Engage vendors and down-select**: Reach out to the different types of tool vendors. Don't just look for a one-size-fits-all product. Instead, look for the combination of technologies that are suited for your organization's use cases and expectations. Choose a set of products for evaluation.
7. **Proof of concept**: Propose and execute one or more proof of concept projects that will not only enable your team to review tool capabilities, it will also raise awareness of the value of a data catalog for data intelligence
8. **Assess suitability**: Use your evaluation criteria to determine which overall solution best meets the needs of your data communities.

The COVID-19 crisis certainly has elevated the importance of data literacy and intelligence and made the right data governance platform, including a modern data catalog, essential to organizations so they can make both short- and long-term decisions regarding their business processes and also their entire business models.

Business transformations have to be based on accurate data assets within the right context, so organizations have a reliable source of truth on which to base their decisions. If you take the steps outlined above, your organization will have the ability to support not only business continuity in a crisis but also demonstrate agility in harnessing its data assets for the intelligence it needs to be relevant and competitive day in and day out.

About the Author

David Loshin, president of Knowledge Integrity, Inc. (www.knowledge-integrity.com), is a recognized thought leader, TDWI affiliate analyst, and expert consultant in the areas of data management and business intelligence. David is a prolific author on topics related to business intelligence best practices. He has written numerous books and papers on data management, including Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph and The Practitioner's Guide to Data Quality Improvement, with additional content provided at www.dataqualitybook.com. David is a frequently invited speaker at conferences, online seminars, and sponsored websites and channels including TechTarget. His best-selling book, Master Data Management, has been endorsed by many data management industry leaders. David is also the Program Director for the Master of Information Management program at the University of Maryland College of Information Studies. He can be reached at loshin@knowledge-integrity.com.

About erwin by Quest

erwin takes a metadata-driven approach to delivering an "enterprise data governance experience," enabling organizations to plan and document how they will discover and understand their data in context, track its physical existence and lineage, and maximize its security, quality and value. The erwin EDGE also helps enterprises operationalize all these steps by integrating business process, enterprise architecture and data modeling with data intelligence software. As the heart of the platform, the erwin Data Intelligence Suite combines data catalog, data literacy and automation capabilities for greater awareness of and access to available data assets, guidance on their use, and guardrails to ensure data policies and best practices are followed. The result is an automated, real-time, high-quality data pipeline from which accurate insights can be derived so both IT and business stakeholders can collaborate for risk management, innovation and business transformation initiatives. For more information, please visit us at www.erwin.com.