Transforming How
Texas Government
Serves Texans

Implementing Data Catalogs

Data Management Advisory Committee

Office of the Chief Data Officer
Texas Department of Information Resources

July 2023

Table of Contents

Introduction 1
Key Features and Benefits of Data Catalogs 1
Steps for Implementing a Data Catalog 3
Stage One: Data Catalog Needs Assessment and Buy-in 3
Stage Two: Data Catalog Preparation and Selection 4
Stage Three: Data Catalog Implementation 5
Best Practices 6
Conclusion 7
References 8
Contributors 8

Introduction

Texas state agencies and institutions of higher education depend on data to facilitate informed choices and streamlined operations. Robust data management programs and systems have become increasingly essential to help leverage this valuable asset while minimizing risks associated with data loss or mismanagement. Data catalogs are a crucial component in managing data assets, allowing Texas state agencies and institutions of higher education to utilize data strategically and effectively.

A data catalog is a wide-ranging inventory of an organization's data assets. Data catalogs simplify data discovery and access, enhance data governance, and promote transparency into the data assets managed by the organization. Adopting a data catalog can significantly improve collaboration, inform decision-making processes, and ensure that the organization has the resources available to achieve its missions and objectives.

This paper provides Data Management Officers, and others responsible for the management and governance of data, information needed to effectively select and implement a data catalog. The paper describes some of the major features and benefits of a catalog, suggests steps for implementing a catalog tool that best meets the needs of the organization, and provides some best practices for consideration. Given the variety of catalog offerings on the market, and the distinct needs of each state agency and institution of higher education, Data Management Officers may identify and follow alternative implementation steps and best practices when selecting and implementing a data catalog.

Key Features and Benefits of Data Catalogs

Data catalogs organize information about data assets for strategic use. At a minimum, data catalogs provide a means for data users to find and understand the data managed by the organization. On the other end of the spectrum, some data catalog tools have advanced capabilities that include extensive data preparation, analysis, and visualization functions, as well as security and governance features.

Regardless of what tool is used to inventory data or the tool's capabilities, data catalogs provide information to allow users to find data available within the organization, choose the appropriate dataset, and view the metadata to understand the data's origins, meaning, access privileges, and approved purposes. Data catalogs can also provide indications of data quality and frequency of usage, to help users understand which data assets might be the most appropriate to use and share. This readily available information not only facilitates efficient and appropriate use but also protects sensitive, confidential, or protected data assets from inappropriate use and sharing.

Some key features and benefits of data catalogs include:

- **Transparency**: Public agencies benefit from enhanced data accessibility and the reduction of data silos, leading to increased accountability and public trust.

- **Improved data management and governance:** Public agencies must have a robust data governance framework to handle huge volumes of data with skill and efficiency, which results in better governance and management of data. Data catalogs support a structured approach to data management, helping improve data accuracy and consistency, and regulatory compliance.
- **Metadata management:** Data catalogs provide a format to document the technical, business, and operational metadata associated with data assets throughout the data lifecycle, beginning with data's creation or acquisition, and into its processing or ingestion into systems. Curation activities may involve creating and updating metadata, which then guides users on the appropriate use of data in the sharing and using stage. Metadata management also supports legal and regulatory compliance by facilitating proper handling, storage, and access to data, as well as its archival and disposal.
- **Enhanced data discovery and access:** Many catalogs offer advanced search capabilities such as smart filters, keyword searches, and natural language queries to aid in discovering data lineages, and data models and hierarchies. By providing a centralized platform that stores, manages, and shares information about data assets, users can efficiently locate, access, and use datasets relevant to their needs and appropriate for their role.
- **Data lineage:** Users can use data catalogs to track data's origin, its relationship to other data, its transformations over time, as well as data ownership and stewardship. This improves data quality, fosters trust in the data, identifies those responsible for the data, and supports change management initiatives by providing valuable insights into data dependencies, impact analyses, and problem-solving efforts.
- **Integration:** Many catalogs offer seamless interoperability with other data management tools (e.g., extract, transform, load tools; data integration across sources; data quality; master data management) which enhances the user experience and contributes to improved data quality. Integrating these tools with data catalogs reduces management efforts and increases efficiency.
- **Data security:** To meet the growing need for data protection, data catalogs include robust governance and security features such as data retention management, compliance adherence, data classification, and role-based access controls. These measures assist in safeguarding sensitive, confidential, and regulated information from unauthorized access and improper use.
- **Regulatory compliance:** Data catalogs help organizations comply with relevant regulations, bolster information security, and adhere to data privacy laws and regulations, thereby minimizing potential legal complications, reputational loss, and penalties. Managing, classifying, and sustaining compliance with applicable laws and regulations are essential functions provided by data catalogs.
- **Cost savings:** There are significant cost savings to be gained from using data catalogs. They can allow the discovery of duplicate data, provide opportunities to eliminate

duplicate data and improve storage efficiency, ensure better data quality, and facilitate prompt and efficient data access. Records retention schedules integrated into the catalog can help identify historical records that may need to be archived or disposed of, which can save costs related to data storage. Additionally, data catalogs can be leveraged as part of an organization's disaster recovery and business continuity plans.

- **Automated catalog population:** Data catalog tools typically include technologies like artificial intelligence and machine learning algorithms to automate catalog population. Automating tasks such as identifying data assets, importing data classifications, metadata extraction, and data quality assessments can alleviate manual work, reduce human error, and save time.

Steps for Implementing a Data Catalog

Implementing a data catalog successfully involves a three-stage approach: assessing data catalog needs and organizational readiness, preparing for implementation, and carrying out the implementation.

Stage One: Data Catalog Needs Assessment and Buy-in

Before pursuing a data catalog development project, public agencies should assess their data catalog needs and level of organizational readiness. This evaluation should determine whether the organization has the necessary resources and support from leadership and stakeholders, the maturity level of the organization's data governance framework, and whether the necessary technical infrastructure is in place to support the functionality of a data catalog. Conducting these assessments up-front uncovers potential roadblocks early on while ensuring alignment of the developed catalog with existing governance frameworks.

A successful data catalog plan depends on well-defined goals that align with the organization's overarching mission and strategic objectives. This will help guarantee leadership's support for the data catalog initiative and ensure the commitment of resources.

Another important initial step involves engaging key stakeholders representing a diverse range of data user groups, including both data producers and data consumers, keeping in mind users may be internal or external to the organization. Through this collaboration process, data users can provide business requirements and use cases that highlight the significance of a data catalog, identify key systems or manual data stores that should be included, and identify the key features required in a data catalog tool.

Assessing the current maturity level of the organization's existing data governance framework is another way to determine how prepared your organization is to adopt a data catalog, and what gaps might need closing first to help enable its success. Questions could include:

- To what extent has a data governance structure been established with resources to oversee the data environment effectively and resolve any potential governance issues?

- Are data governance policies, procedures and standards in place and are they complied with?
- Are there technology tools to assist in the automation of data governance processes?
- Is your organization ready to institutionalize structures and processes to comply with the data governance charter and any new legislation?
- Is data valued as an asset?
- Do you have a dedicated experienced team in place to understand, design, and execute a data strategy and roadmap?
- Are processes in place to develop a data strategy and execute against the strategy?
- Are there technologies and tools in place to support the development of the data strategy and roadmap?
- Do you have change management processes in place for documenting and communicating data issues and root cause solutions, and enforcing Service Level Agreements?

Analyzing the organization's technology infrastructure and how data users need to use the data is another important step in the planning stage. Questions to keep in mind include:

- What kind of data does the organization have and want to include in a catalog?
- Do both structured and unstructured data exist and need to be included?
- Does the catalog need to include reports, visualizations, and dashboards that may exist in the organization?
- Do data users need to know what datasets and data elements are most commonly used, or most appropriate for certain use cases?
- Do you need to catalog data assets at the system, object, or data element level?
- Which data assets need information on their relationships or connections to increase interoperability?
- Do data users need a catalog that can connect to data assets whether they are on-premises, in the cloud, or in a hybrid environment, or to data housed in multiple environments (e.g., development, testing, production)?

Stage Two: Data Catalog Preparation and Selection

In stage two, the organization prepares for data catalog selection and implementation by examining the business requirements and use cases gathered in stage one to determine needs and priorities, such as data governance, analytics, cloud transformation, privacy and security, risk mitigation, and regulatory compliance, as well as any required key features. For example, data catalog design considerations may focus on data quality, data lineage, metadata management, and searchability. User-friendliness and accessibility for a diverse range of users should also be considered.

Because data catalogs are not one-size-fits-all, it is essential to research and evaluate the variety of offerings based on cost, functionality, scalability, vendor support, and alignment with organizational goals. When selecting a data catalog, an organization should include a description of areas identified for improvement in stage one, how a catalog can contribute to the overall mission and goals, and how success will be measured.

In preparation for stage three, the organization should begin planning implementation and training processes, including an evaluation of data users' skillsets and any gaps that may need to be addressed to promote effective use of the data catalog.

Stage Three: Data Catalog Implementation

In stage three, the organization should establish and document an implementation plan, and establish policies and procedures for governing the data catalog once it has been implemented.

The catalog should be implemented and expanded gradually, adding new data assets as users become more familiar and comfortable using the catalog. This gradual expansion can be broken down into the following stages:

- Initial rollout: Begin by cataloging a limited number of high-priority data assets, enabling users to familiarize themselves with the catalog and provide feedback for improvement.
- Incremental expansion: As users become more proficient in utilizing the catalog, progressively add more datasets and incorporate additional features based on user requirements.
- Ongoing development: Continually refine the data catalog by monitoring usage patterns, addressing user feedback, maintaining and updating asset information as necessary and appropriate, and improving the system's overall performance and capabilities.

Analyze and document the different roles and responsibilities of the staff interacting with the data catalog both during and after implementation. Users are individuals who access and utilize the catalog for various data-related tasks, such as data discovery, understanding, analytics, and collaboration. Administrators, on the other hand, are responsible for managing the catalog, including its configuration, expansion, maintenance, and ensuring data governance policies are applied.

In this stage, the organization should also develop and implement a training program to promote effective use of the catalog. Training should encompass various aspects respective to different user responsibilities, such as:

- Understanding the purpose and objectives of the data catalog.
- Understanding the different roles and responsibilities associated with developing, maintaining, and using the data catalog.
- Familiarizing the user with the features, functionalities, and user interface.

- Learning to search, access, and collaborate on data assets.
- Mastering the application of data governance policies and metadata management.
- Applying best practices for maintaining the data catalog and ensuring data quality.

Showcasing the benefits of the catalog and sharing its outcomes and successful experiences with the organization's leadership and stakeholders, as well as tracking progress towards informed decision making through use of the catalog, will help increase user adoption. This user-centric approach guarantees that the catalog remains a valuable and accessible tool for all stakeholders. Further, by analyzing this information on a regular basis, organizations become better equipped with insights required for enhancing their data catalogs accordingly.

Best Practices

Adhering to best practices supports the effective implementation of a data catalog, improved data management, increased accessibility, and better data-driven decision making.

- Defining clear objectives, key performance indicators, and gathering and comprehending stakeholders' data requirements are crucial for a successful data catalog strategy. To effectively tailor the data catalog, it is important to gain insights into their diverse needs. Engaging stakeholders in the data catalog development and execution process is vital for aligning the catalog with their objectives and expectations.
- Encourage a data-driven culture within the organization by promoting the importance of data catalog usage and emphasizing its benefits for decision making and collaboration.
- Establish and enforce data governance policies, including data ownership, data quality, and data security. Communicate these policies to all users of the data catalog and promote adherence to them.
- Allocate time and resources to gather information on source systems and manual data stores, involving cross-functional teams in the process and encouraging them to contribute their department's data sources to the data cataloging effort.
- Plan to catalog all data but build incrementally by selecting a cost-effective and versatile data cataloging tool that is compatible with various data types, scalable, and user-friendly. Start with smaller, high-impact data assets, and add additional data assets as users become familiar with the tool and benefits are realized. By cultivating users' skills and understanding of the best practices with smaller data assets, you can effectively build their capacity and prepare them to tackle larger, more complex data assets.
- Implement data validation, cleansing, and enrichment processes before adding data to the catalog to enhance accuracy and reliability. Perform regular audits of the catalog and address any data quality issues that arise.

- Prioritize metadata management because it is essential for data discoverability and understanding. Collecting accurate and consistent information about source systems and the data within those systems increases the effectiveness and usability of the data catalog. Evaluate and select appropriate data cataloging tools by considering their capabilities in handling large amounts of metadata, supporting metadata extraction from source systems, and balancing cost-effectiveness with user-friendly design.

- Determine the process of gathering and ingesting both technical and business metadata to make sure connection requirements are compatible with business requirements. Pay particular attention to data assets containing confidential or sensitive information with access controls that may interfere with automated uploading of metadata and require manual input of metadata.

- Prioritize data privacy and security in the management of data catalogs by establishing a framework that protects sensitive information, upholds individual privacy, and fosters openness and accessibility for the broader community.

- Determine integration needs between different data management tools. Some cataloging tools are better than others at integrating with other data management tools. For example, it may not be easy to combine tools from different vendors, and in-house developers may need to create custom code to enable integration.

- Assemble a proficient data catalog administration team that possesses both technical and program or business area expertise to effectively implement and maintain a data catalog, champion its use, and address challenges that may arise.

- Implement a training program for those who will be administering and using the catalog to develop their understanding of its features, functions, and best practices.

- Establish a feedback loop with data catalog users to understand their needs, challenges, and areas for improvement. Perform regular assessments of the catalog's effectiveness and make any necessary adjustments.

Conclusion

Data has become an integral component in Texas state agencies and institutions of higher education. To fully unleash the potential of data, agencies should adopt a data catalog. Data catalogs create a centralized platform for controlling, exploring, and understanding data, thereby fostering effective data management strategies with transparency and accountability to drive the organization’s mission and goals.

Leveraging data catalogs offers several benefits for organizations, including better access to information for sound decision making, compliance with laws and regulations, and lower costs through greater efficiencies in processing and analyzing data. Data Management Officers can determine the best data catalog for the organization by identifying the organization’s specific

needs and priorities, assessing the maturity of the organization's data governance framework, and assessing the organization's technical infrastructure to support the functionality of a data catalog to best meet data users' and the organization's needs.

Finally, implementing gradually, providing training to users, and following the best practices described here will facilitate the development of a successful data catalog.

References

https://www.dataversity.net/data-governance-and-data-catalogs-where-do-they-intersect/
https://www.dataversity.net/ten-recommendations-for-building-great-data-catalogs/
https://research.aimultiple.com/data-catalog/
https://www.topbots.com/choosing-a-data-catalog/

Contributors

Ricky Beverlin, Texas Department of Information Resources
Monica Smoot, Texas Department of Information Resources
Leah Porras, Texas Health and Human Services Commission
Edward Kelly, Texas Secretary of State
Carol Tucker, University of Houston – Downtown