Eckerson Group

The Ultimate Guide to Data Catalogs

Key Things to Consider When Selecting a Data Catalog

By Dave Wells

March 2018

Co-sponsored by

Informatica™

About the Authors

Dave Wells is an advisory consultant, educator, and research analyst dedicated to building meaningful connections along the path from data to business impact. He works at the intersection of information and business, driving value through analytics, business intelligence, and innovation. With nearly five decades of combined experience in information management and business management, Dave has a unique perspective about the connections of business, information, data, and technology.

Knowledge sharing and skills building are Dave's passions, carried out through consulting, speaking, teaching, research, and writing. He is a continuous learner—fascinated with understanding how we think—and a student and practitioner of systems thinking, critical thinking, design thinking, divergent thinking, and innovation.

About Eckerson Group

Eckerson Group is a research and consulting firm that helps business and analytics leaders use data and technology to drive better insights and actions. Through its reports and advisory services, the firm helps companies maximize their investment in data and analytics. Its researchers and consultants each have more than 25 years of experience in the field and are uniquely qualified to help business and technical leaders succeed with business intelligence, analytics, data management, data governance, performance management, and data science.

Research
Consulting
Assessments

Executive Summary

Data management is difficult in today's world. Big data, cloud hosting, self-service, and heightened regulation contribute to data management complexity. In response, organizations are turning to data catalogs, which are proving to be state-of-the-art data management tools. From modest beginnings as data inventory and search tools, data catalogs have grown to support business analysts, data scientists, data stewards, data curators, and data engineers.

Data catalogs also have strategic value. Chief data officers (CDOs) and chief analytics officers (CAOs) view the catalog as strategic not just for data inventory, but also for data asset management, data governance and self-service analytics. Compliance officers see cataloging of sensitive data as particularly critical with the rapid approach of General Data Protection Regulation (GDPR). Culturally, cataloging drives a shift from centralized, IT-centric data management to community-focused curation with extensive business collaboration. In short, the data catalog has become an essential component of modern data management.

Choosing a data catalog that meets all these needs, addresses everyone's interests, and fits your environment and culture is a big job. This report describes 20 criteria to help evaluate and select a data catalog. Choosing the right catalog is the first step toward joining the modern age of data management.

A Data Management Imperative

The difficulties of data management have intensified at a steady pace over the past several years. The management complexities of big data, cloud hosting, self-service analytics, and tightening regulations can't be ignored. Effective data management has become a top priority for most organizations, but getting there is challenging. Data catalogs fill essential roles in overcoming these challenges.

The Roles of a Data Catalog

Data catalogs were introduced to help data analysts find and understand data. Before data catalogs, most data analysts worked blind, without visibility into existing data sets or their contents, or the quality and usefulness of each. As a result, they spent much of their time finding data, understanding data, and recreating data sets that already existed. Data catalogs were designed to address these issues.

The data catalog has become an essential component of modern data management.

From modest beginnings as a means to manage data inventory and expose data sets to analysts, the data catalog has grown in functionality, popularity, and importance. Modern data catalogs still meet the needs of data analysts, but have expanded their reach. They are now central to data stewardship, data curation, and data governance.

For example, data catalogs have become strategically important. Chief data officers (CDOs) and Chief analytics officers (CAOs) view the catalog as strategic not just for data inventory, but also for managing data assets and improving analytic quality and productivity.

Compliance officers see cataloging of sensitive data as particularly critical with the General Data Protection Regulation (GDPR) rapidly approaching and introducing new governance requirements to strengthen and unify data protection for individuals. Data strategists and architects understand the important role of cataloging to enable the Enterprise Data Marketplace, a services-based approach to data provisioning built on a digital marketplace model. Culturally, cataloging drives a shift from centralized, IT-centric data management to community-focused curation with extensive business collaboration. Clearly, the data catalog has become an essential component of modern data management.

The Critical Role of Metadata

It seems that everyone wants data management but few care about metadata. Just as you need data about finances for effective financial management, you need data about data (metadata) for effective data management. You can't manage data without metadata. The data catalog has become the new gold standard for metadata and a cornerstone of data curation.

The data catalog has become the new gold standard for metadata.

Metadata is the core of a data catalog. Every catalog collects data about the data inventory and also about processes, people, and platforms related to data. Metadata tools of the past collected business, process, and technical metadata, and data catalogs continue that practice. The metadata management game changers with data cataloging are the following:

- **Crowdsourced metadata.** Much of catalog metadata is collected automatically by applying algorithms and machine learning. But sometimes the most valuable metadata is the knowledge and experiences of individuals and groups. Collecting that knowledge as user ratings, reviews, tips, and techniques enriches the metadata collection and converts tribal knowledge into a shared and enduring data management resource.
- **Data about people.** Data management and data analysis are ultimately human activities. Knowing which people have data roles and relationships and the nature of those roles is valuable. Data catalogs capture metadata to identify data users, data creators, data stewards, and data subject matter experts.
- **Automated metadata discovery.** Organizations with massive data holdings—literally tens of thousands of databases—simply don't know about all of the data they have. It is impossible to catalog a petabyte data estate without automated discovery.

The real value of metadata is found in the answers it can provide. People who depend on data have questions about trustworthiness, latency, lineage, sensitivity, preparation, and much more. Sometimes they want to find others who know or have worked with the data to get human perspective. And they need to know about access, privacy and security constraints, cost, etc. Robust metadata ranging from data set names and properties to usage, access, licensing, and subject experts is the key to answering the many questions that data users and data managers will ask.

Data Catalog Tools

The selection of data catalog tools has grown rapidly in recent years. Several data cataloging tools are available today with new tools emerging and catalog functions being added to existing tools regularly. Scope of data varies widely among data catalogs with some focused on enterprise data from internal systems, and others on external data captured in a data lake, and still others on published reports and dashboards.

Data catalog tools exist today in several forms as described in the table.

Catalog Type	Catalog Characteristics
Standalone	• Catalog of data sets and operations • Supports data set search and evaluation • Seamless user experience requires high level of interoperability
Integrated with Data Preparation	• Catalog of data sets and operations in a tool that includes extensive data preparation features and functions • Seamless user experience for finding, evaluating, and preparing data • Requires high level of interoperability with analysis tools
Integrated with Data Analysis	• Catalog of data sets in a tool that includes extensive data analysis and visualization features and functions • May catalog operations and support basic data preparation • Seamless user experience to find and analyze data • Requires high level of interoperability with data preparation tools when advanced data preparation capability is needed
Fully Integrated Solution	• Catalog of data sets and operations in a tool that includes extensive features and functions for data preparation, analysis, visualization, governance, and security • Seamless user experience throughout the analytics lifecycle • Interoperability becomes important in organizations where multiple data preparation and/or analysis tools are used

The value of a seamless user experience throughout the analytics lifecycle is evident, so the trend in catalog evolution is toward convergence. Most tools will mature to become fully integrated solutions supporting all three capabilities—cataloging, preparation, and analysis. Convergence, however, does not eliminate the need for interoperability, as self-service analysts often want to make their own choices of preparation and analysis tools.

The tools landscape is continuously evolving, but the things that a cataloging tool must do are clear. Every catalog must provide functions for initial build of the catalog, for growth and ongoing maintenance, and for regular use by people who seek and work with data.

Building the Data Catalog

The initial build of a data catalog typically scans massive volumes of data to collect large amounts of metadata. The scope of data for the catalog may include any or all of data lakes, data warehouses, data marts, operational databases, and other data assets determined to be valuable and shareable. Collecting the metadata manually is an imposing and potentially impossible task. The data catalog automates much of the effort using algorithms and machine learning to accomplish the following:

- Find and scan data sets.
- Extract metadata to support data set discovery.
- Expose data conflicts.
- Infer semantic meaning and business terms.
- Tag data to support searching.
- Tag privacy, security, and compliance of sensitive data.

Some data catalogs also support bulk import of business glossaries from legacy data management or EIM systems to help accelerate implementation.

Manual effort is required to complete the initial build. Automation is powerful but completion requires gathering tribal knowledge about data sets that are useful to and valued by data analysts. Curation work to enrich the tags and metadata that support data set discovery, evaluation, and understanding is also an important part of completing the catalog. The catalog should include support for the following:

- Crowdsourcing of metadata for data set discovery.
- Curator-added metadata for semantic meaning.
- Crowdsourced subject matter expert (SME) knowledge to tag and describe data.
- Curator-added business terms for tagging and searching.
- Curator identification of curators, stewards, and SMEs.
- Curator-added annotations with tips, techniques, and cautions for working with data.
- Crowdsourced usage metadata.
- Curator-added usage metadata.

This long and imposing list shouldn't discourage you from getting started with a data catalog. The manual work can be performed as part of the initial build or as an ongoing activity that enriches catalog content through use and active curation. Much value can be derived from a data catalog built with maximum automation and limited manual effort.

Growing and Maintaining the Catalog

Catalog growth is typically a process of expanding the scope of data that is cataloged. Incrementally building a data catalog is a practical approach where the first build might catalog a warehouse and certified data in a data lake. Following increments may expand scope to include non-certified data, operational data, and other sources. When growing the catalog incrementally, each new data source can be approached using the same processes and practices as for the initial build, with automated processes at the forefront.

Maintaining the catalog is primarily a curation responsibility to keep metadata fresh, current, and relevant. Learning through the experience of using data sets, cataloging newly created data preparation operations and workflows, crowdsourcing of tips and techniques, removing obsolete data sets, and updating when changes occur are central to catalog maintenance. While some updates may be automated, much of the maintenance effort relies on manual curation activities. Data curator contributions to the catalog are important when adding annotations that provide guidance and cautions for working with a data set. The curator role is especially important in designating trusted data sources.

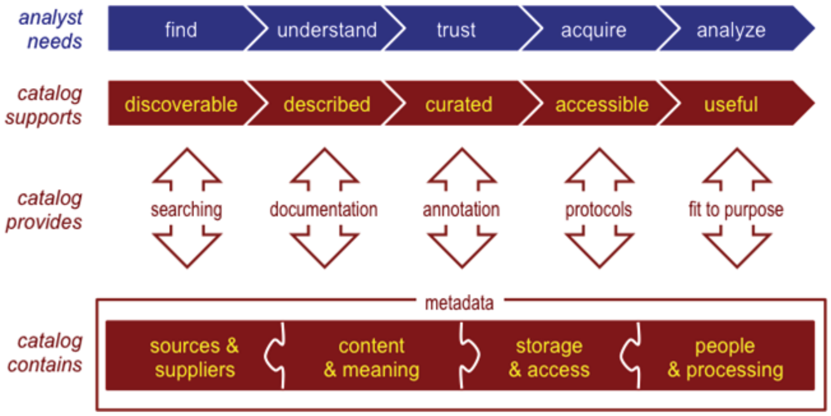
Using the Data Catalog

The ultimate users of a data catalog are people who report, visualize, and analyze data. Users range from non-technical business people to highly skilled data analysts. Usability is an important consideration. It is essential that a data catalog support non-technical business users who work with data. A catalog that supports the needs of line-of-business analysts will easily satisfy the requirements of more technical users.

Analysts work with data. (See figure 1.) To work with the right data, and to work with data in the right ways, they must be able to find the data, understand it, judge its trustworthiness, and acquire (download, transfer, or access) it. Only then can they analyze data effectively. The data catalog supports analyst needs by making data discoverable (finding), described and curated (understanding and trusting), and accessible and useful (acquisition and analysis). Catalog support is provided through functions and processes for searching, documenting, and annotating as well as describing the protocols to acquire data and the purposes for which it is suited.

Catalog functions are built on a metadata foundation that collects data about sources and suppliers of data, content and meaning of data, storage of and access to data, and the people who use data and how they use and process it.

Figure 1. Using the Data Catalog



While oriented to self-service data analysts, the user base also includes highly technical people. Data engineers and technical staff in IT organizations will frequently use the catalog to determine if needed data exists and avoid creating redundant data sets. Compliance officers, risk officers, and data governance professionals may use the catalog to determine where data resides throughout an organization and perform audits to identify exposure and risk.

Evaluating Data Catalog Tools

Data catalog stakeholders span a continuum from business and data analysts to C-level executives, and catalog impacts range from day-to-day tactical activities to long-term strategic position.

Usability is essential to widespread catalog adoption.

Choosing a catalog that meets all of the needs, addresses all of the interests, and fits your environment and culture is a big job. Usability is a paramount consideration with the variety of users and the broad spectrum of data and technical skills. Intuitive user interface and ease of use are essential to widespread catalog adoption. The twenty criteria listed here are designed to help you work systematically through the evaluation process and find the catalog best suited for your organization.

1. Cataloging Data sets. A data catalog should support automated discovery of data sets, both for initial catalog build and ongoing discovery of new data sets. Use of machine learning for metadata collection, semantic inference, and automated tagging is important to get maximum value from automation and to minimize the manual effort of data cataloging. Consider the scope of data that can be cataloged automatically by crawling file systems and data stores including data lakes, data warehouses, operational data, SaaS databases, and copies of data formatted for specific data analysis tools. Also look for the ability to catalog reports, dashboards, and other published data assets. Though much of the data can be cataloged automatically, a data catalog should also offer manual cataloging functions for data curators.

2. Cataloging Data Operations. A data catalog should include cataloging of data preparation operations and associate them with data sets. Operations include processes to improve, enrich, format, and blend data. Data blending operations must be associated with multiple data sets, and many other operations may apply to multiple data sets. Expect the catalog to support many-to-many operation-to-data set associations. Look for ability to catalog data preparation workflows to prescribe sequences for a set of operations. Also consider ability to designate mandatory operations such as masking or obfuscation of personally identifying information (PII).

3. Searching. Searching for data sets is a fundamental requirement of catalog users. Robust search capabilities include search by facets, keywords, and business terms. Natural language search capabilities are especially valuable for non-technical users. Ranking of search results by relevance and by frequency of use are particularly useful and beneficial features. When search functions intersect with security authorizations, look for capabilities to hide or mask (gray out) data sets the user is not authorized to access. You'll want to hide data sets when unauthorized users should not know of their existence, and mask them when existence can be shown but authorization is needed for access.

4. Recommendations and Relationships. When searching for data sets, a recommendations engine is an especially valuable feature. Leveraging usage history metadata and machine learning to develop recommendations based on past user experiences accelerates the search process, improves quality of match between search results and user needs, and makes strong connections between data sets and data preparation operations and workflows. Automated detection and display of relationships among and overlaps between data sets is a powerful feature that supports advanced capabilities for data discovery, data curation, and data blending recommendations.

5. Data set Evaluation. Finding data sets is only the beginning for the data analyst. Choosing the right data sets depends on ability to evaluate their suitability for an analysis use case without first needing to download or acquire the data. Important evaluation features include capabilities to preview a data set, view data profiles, see user ratings, read user reviews and curator annotations, and view data quality information.

6. Data Access. On completion of data set evaluation, desired data sets should be accessible directly from the catalog, providing a seamless user experience from search through data acquisition. Consider the variety of data set types to which the catalog can connect. RDBMS, flat files, tagged files (JSON, XML, etc.), document stores, graph databases, geospatial data, and text documents (Word, PDF, etc.) are common source types. Data access functions should include access protections for security, privacy, and compliance of sensitive data.

7. Usage Metadata. Collection of usage metadata enables other important features including data set evaluation and intelligent recommendations. Look for ability to collect information about each data set including: Who has used the data set? For what use cases has it been used? How frequently is it used? What is the user experience (reviews and ratings)? With what other data sets is it typically used or combined?

8. Data Valuation. As the catalog becomes a data strategy component of interest to CDOs and CAOs, data valuation is a consideration. How will the catalog help you quantify value of data assets? Knowledge about frequency of use and analytic use cases provides a starting point. Does the catalog include specific data valuation features? Don't expect the data catalog to calculate a dollar value for each data set, but it should provide usage information that contributes to value estimation.

9. Metadata Catalog. Consider the richness of metadata that is collected. What data is collected about data sets? What data is collected about processes, and does it support full data lineage traceability? What metadata supports searching? Does it include data about curators, data stewards, SMEs, and data SMEs? How comprehensive is usage metadata? Is data captured about suppliers, service providers, and other external entities?

10. Security. Data security is the first of four essential data governance capabilities. Check for ability of the catalog to work with your existing security infrastructure and processes for user authentication and authorization. User security should at minimum distinguish between administrative users such as curators and analytic users. Also consider the levels at which security constraints can be imposed. Can you secure at a data set level? At record or row level? At column or field level? Can access be constrained based on the values of specific fields?

11. Lineage. Data lineage is a core data governance consideration. Ability to trace data from original source, through analysis and reporting processes, to final analysis and reporting is a key component of trusted data. It is also valuable for change management, impact analysis, troubleshooting, and problem solving.

12. Compliance. Regulations focused on data protection are increasingly common, and a significant governance responsibility. GDPR is an immediate concern, but a multitude of other data-related regulations exist—many of them industry-specific, such as HIPAA and Dodd-Frank. Look closely at how the catalog handles PII, protects data privacy, and supports compliance with regulations. A common and important compliance feature is ability to automatically obfuscate or mask sensitive fields, such as exposing only the last four digits of social security numbers.

13. Quality. Data quality is a fourth major governance concern—one that has become more complex with the adoption of big data and data lakes. The catalog won't cleanse data or improve data quality, but it does have an important role in data quality management. Smart algorithms may expose data conflicts and identify data quality deficiencies. Displaying automated and human judgments of data quality helps analysts evaluate and select data sets and decide how to work with less than perfect data. User reviews and curated annotations may offer recommendations for tuning analytic models to compensate for noise in data. Communication and collaboration are key quality management tools for which the data catalog can prove valuable.

14. Data Curation. Data curators interact frequently with the data catalog and fill a critical role in making it useful and valuable. Evaluate the richness of curation capabilities including ability to add data sets, hide or remove data sets, add annotations, create metadata, add search terms and tags, identify stewards and SMEs, tag security- and compliance-sensitive data, share tips and techniques, and encourage crowdsourcing of metadata.

15. Socialization. As cataloging drives cultural shifts to collaborative data management and to community curation, socialization becomes an important element. Evaluate social capabilities such as crowdsourcing of metadata, collaboration features, posting of user ratings and reviews, and capture of user feedback. Then look beyond the capabilities to consider usability and motivation. Social features not used provide little value. What does the catalog offer that makes it quick, easy, and desirable to participate in social aspects of data cataloging? Also consider what you can add organizationally and culturally—gamification, recognition, incentives, etc.—to encourage participation.

16. Integration and Interoperability. The catalog can't operate in isolation. It needs to work seamlessly throughout the analytics lifecycle from problem framing to data visualization, and to be seamless regardless of data preparation and analysis tool choices. How well will the catalog work with your data preparation tools? How well will it work with your data analysis and visualization tools? Will it integrate gracefully with your security and access controls?

17. Deployment. You'll certainly need to consider how the catalog fits into your current and future technical infrastructure. Does it offer options for on-premises, cloud, and hybrid deployments? Can it support both server-based and Web-based implementation? How well will it support your mobile and geographically dispersed users?

18. Services. The nuances and details of catalog implementation can sometimes be challenging, and consulting services may prove valuable, especially when working with non-traditional data types. Data catalog users are likely to need some introductory training, and data curators may require more depth of training. Be sure to ask what kinds of training and consulting services are available. To meet basic training needs look for free or low cost services such as video or built-in tutorials. Also look for user groups or online forums as sources of knowledge and problem solving.

19. Pricing. Budget and cost are always considerations when acquiring new technology. Ask about the vendors' pricing models. Will you pay by user seats? By volume of data? By number of data sets? Or by other criteria? What should you expect as initial costs and ongoing costs? How can you estimate TCO?

20. Vendor Roadmap. What plans does the vendor have for future features and functions? Will they expand integration with data preparation and data visualization tools? Will they increase interoperability with various preparation and analysis tools? Do they plan to offer connectors to various challenging data sources? How many data sources can they currently connect to? Are they adding advanced collaboration and socialization features or moving toward enterprise data marketplace capabilities?

Join the Age of Modern Data Management

Data management has changed radically in recent years. It is no longer a technical discipline practiced entirely within IT organizations. Business has a significant and growing role in managing data. We're moving rapidly to an era where communication, collaboration, and crowdsourcing are the mainstays of data management. The data catalog is the glue that holds it all together. Managing data in today's world without a data catalog is ill advised and impractical.

Managing data without a data catalog is ill advised and impractical.

Choose your catalog technology carefully. It will become a core component of your data management infrastructure, and it will be highly visible to all who have a stake in data management. Select the right tool, and then build the catalog. Set scope for the first cataloged data sets. It isn't practical to catalog everything at once, so choose a starting place and plan to grow the catalog incrementally. Automate as much as you can. Use the machine learning capabilities to populate the catalog with limited human effort. A small team and a good tool are more effective than a large team and a labor-intensive cataloging process.

Need help with your business analytics or data management and governance strategy?
Want to learn about the latest business analytics and big data tools and trends?
Check out **Eckerson Group** research and consulting services.

About the Research Sponsor

Informatica™

Digital transformation changes expectations: better service, faster delivery, with less cost. Businesses must transform to stay relevant and data holds the answers.

As the world's leader in Enterprise Cloud Data Management, we're prepared to help you intelligently lead—in any sector, category or niche. Informatica provides you with the foresight to become more agile, realize new growth opportunities or create new inventions. With 100% focus on everything data, we offer the versatility needed to succeed.

We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption.

To learn more about Informatica Enterprise Data Catalog, please visit:
https://www.informatica.com/products/big-data/enterprise-data-catalog.html