Introduction

In many companies, data is still a highly underutilized asset. It is very likely that data collected in one business unit for a specific purpose can create value in many other usage scenarios. However, the knowledge about which data exists and its ownership is not always obvious. As a result, data is hardly used beyond its original context, and many opportunities to create value from data remain unused.

One approach to solve this problem is to implement a data catalog. A data catalog is a company-wide inventory of existing data. The purpose of a data catalog is to help its users to discover, understand and trust potentially relevant data assets that they do not know well or did not know before. It contains information that helps to understand the technical characteristics and the business context of all data assets of a company. With a data catalog in place, it is easy to find and understand data assets even if the context of data collection is entirely different from the context of data consumption.

The advantages of having a data catalog are apparent. However, implementing a data catalog is not always straightforward. Companies that want to set up a data catalog have to tackle some challenges including the development of requirements, the selection of tools and the design of processes and supporting structures.

In this point of view, we describe a capability maturity model for data catalogs that provides orientation during the implementation of a data catalog. This perspective is based on cross-industry project experience, interviews with international experts, and literature sources. We first define a framework of five capabilities that can be used to develop requirements for data catalog implementations. We further provide a maturity model that enables companies to assess their status quo and plan their target level of maturity. Finally, we give recommendations for data catalog implementations based on former project experience.

What is a data asset?

A data asset is data, which can be used to generate value for the company. You need to understand the data and know how to interpret it to realize its potential. The existence of data about data, or metadata, is a necessary condition to turn data into an asset. With the help of meaningful metadata in a central catalog, data can be found, understood and used even beyond the context of its collection.
Capability Model

Our data catalog capability model defines five capabilities that companies require to increase the level of data utilization.

The capabilities are company-level skills embedded in people, technologies, and processes. Companies can use the capability model as a starting point to define requirements, select tools, and design or adapt processes related to the data catalog.

**Figure 1: Capability map for metadata systems**

- **Discovery**
  - Provides a clear and comprehensive overview of all data assets
  - Displays sample data entries and summary statistics
  - Has search, query and recommendation functions

- **Collaboration**
  - Has an intuitive user interface
  - Supports task-sharing and communication
  - Rewards user contributions e.g. the creation and revision of descriptions

- **Trust**
  - Information on data quality and coverage
  - Information on data lineage
  - Information on the accountable person e.g. data steward or data owner

- **Provision**
  - Supports manual and automated capturing of metadata
  - Can perform data profiling and tagging
  - Detects and suggests data lineage relationships

- **Data Governance**
  - Supports cross-functional workflows, e.g. for approvals
  - Supports compliance with legal and company requirements
  - Provides access rights management
Discovery
Discovery describes the process of browsing the catalog in search for data assets. Supporting data discovery is the most important capability of a data catalog. A clear and comprehensive overview of all available data assets is the key feature to increase utilization and monetization of data assets. The availability of search, recommendation and query functionality further facilitates the discovery process. Users should be able to identify relevant data assets quickly and to have a look at sample data entries as well as summary statistics.

Provision
Provision is the process of adding new entries to the data catalog and enriching them with metadata, e.g., who owns the data? When was it collected and where is it stored? Why was it gathered, what was the unit of measurement and how were values calculated? A data catalog should support both ways of metadata generation: manual and automated. Some types of metadata can and should be captured automatically such as the information where the data is stored. The descriptions of the business context and purpose of data collection have to be entered manually and should be supported well.

Trust
Trust stands for the capability of the catalog to enable its users to assess the reliability and quality of the data. Data scientists developing advanced use cases need confidence in the data they use.

Collaboration
Collaboration summarizes the capability of the data catalog to support its users in task sharing and to reward the creation and revision of descriptions. An obvious starting point is an intuitive, attractive and powerful user interface. The catalog should also facilitate cross-functional workflows. For example, if an analyst needs further information on a data asset, but the explanation of the business context is missing, the users should be able to request the data steward to add. Such a request feature helps to allocate efforts and prioritize the manual process of describing the business context of data.

Governance
Governance captures the capability to support and satisfy compliance with legal and company standards. An example is the management of access rights. A data catalog should list all data assets and display example entries, but it should also comply with data privacy regulation. Hence, it must not show example entries of personal data unless a user has requested and received the corresponding access rights.

“A collaborative approach for business taxonomy is necessary, since a centralized development of a comprehensive taxonomy is unrealistic.”

Nathan Jones
Director Deloitte Analytics Switzerland
Maturity Model

Companies differ significantly in their progress of inventorizing data assets. Our maturity model helps to assess the actual state and define the target state of a data inventory along the five capabilities described above.

Companies that do not use a specific data catalog tool are in the first or second maturity level. Companies that already use a dedicated catalog-tool are in the maturity levels three to five.

After the introduction of a specialized tool, the maturity levels differ in the degree of automation, the effectiveness of incentive structures, and the efficiency of workflows.

---

**Figure 2: The maturity model for metadata systems**
Maturity level 1: Initial
In the first and lowest maturity level, no effort to centrally inventorized data assets has been made. In this initial level, much of the valuable information is not available centrally to data scientists. Even if an analyst knows about the existence of a specific data asset, the only way to understanding its business context and lineage is to contact the source system owner or data engineers that were involved with the ingestion of the data asset.

Maturity level 2: Managed
In the second maturity level, a listing and documentation of data assets exists. Responsible persons record and maintain elementary metadata using standardized templates. Most of the work is done manually in a spreadsheet or a wiki. Users can get information on data assets from a list and are able to use basic search functionality. Catalog entries become obsolete relatively quickly, e.g., when changes to the data assets are not communicated or because it takes a long time until the manual adjustments are made.

Maturity level 3: Tool-based
In the third maturity level, a data-cataloging tool is used and managed by a dedicated team. The tool helps to automate the capture of metadata, e.g., it automatically detects and tags account numbers or human names. Applied taxonomy and predefined tags facilitate the manual provision of metadata in the business context of data assets. An advanced search function makes it easier to find potentially relevant data.

Maturity level 4: Optimized
In the fourth maturity level, sample data entries and summary statistics are available in the catalog for most of the data assets. The catalog also contains information on data lineage which is captured automatically. The lineage information helps users to gain trust as they are able to quickly determine the origin of the data as well as all prior processing steps. Nudges or gamification features facilitate all manual steps of captured metadata while workflows support all governance processes.

Maturity level 5: Automated
The fifth maturity level data catalog recommends data assets to users. For all data assets in the catalog, automatically generated quality metrics are available. A machine learning algorithm suggests tags and categorizations for those data assets that were not yet described by users. All manually maintained catalog entries are of high quality as a financial bonus rewards valuable contributions. The process of getting access to potentially critical data is mostly automated. Human intervention is only necessary in exceptional cases.
Recommendations

It is too early to talk about best practices for data catalogs. Leading technology companies and financial institutions are still in the process of introducing or further developing their data catalogs. Nevertheless, based on our initial experience, we can give the following recommendations that should be taken into account when launching a data catalog.

**Strive for the broadest possible reach**
A central promise of the data catalog is that data is used, combined or analyzed in other ways than anticipated during data collection. Therefore, you should not omit any sources or systems when taking inventory of the data. Hence no source system or type of data should be excluded and the effort should strive for the broadest possible reach of data assets. The catalog can then be used to support prioritization and qualification of the data assets.

**Align the initiative with the data strategy and adapt data governance structures**
A well designed and implemented data catalog is just one element of successful data utilization. The overall data strategy should provide guidance on why and how the company in general aims to generate advantages with data and define how the data catalog can be implemented. Furthermore, a data catalog initiative should also include a review and adjustment of a company’s data governance structures.

**Consider not only established but also new vendors when selecting a tool**
The number of providers of database catalog tools is continually growing, as is the variety of features offered. In addition to established vendors such as Informatica and Collibra, there are young companies including Alation and Waterline that provide exciting data cataloging tools. Companies can achieve the best results when they take the time to precisely define their specific requirements and then consider a broader field of vendors which includes both established and new ones.

**Test carefully if smart data tagging and profiling features live up to their promises**
Many tools offer machine-learning based data tagging or data profiling. Examine whether these features work well with your company’s data. If a large number of data assets should be listed in the catalog, test thoroughly whether computationally-intensive functions scale well.
Conclusion

Implementing a data catalog is vital to increase the utilization of data in the company. A fully operationalized data catalog facilitates data discovery and understanding across business units and helps users to generate previously unrealized benefits with the data.

The five capabilities: discovery, trust, provision, collaboration, and governance, provide a framework to structure and guide discussions in the early phases of a data catalog implementation. All capabilities emphasize the importance of the people that use the catalog. Although a data catalog at a high maturity level is a highly automated IT system, it remains a tool used and filled by people. That is why usability, incentives, and rewards are as crucial as the technical features.

Deloitte has helped several clients to increase data utilization with a data catalog, particularly by defining the company-specific requirements, selecting and adapting tools, and adjusting data governance structures.
Think Tank

Olaf Peter Schleichert  
Partner  
Lead Deloitte Analytics Germany  
Head of Deloitte Analytics Institute  
orschleichert@deloitte.de

Oliver Bieh-Zimmert  
obieh-zimmert@deloitte.de

Stefan Kraus  
stekraus@deloitte.de

Michael Engel  
mengel@deloitte.de

Mehr Informationen  
http://www.analytics-institute.de