





Functional Requirements for Enterprise Data Catalogs: A Systematic Literature Review

Dimitri Petrik^{1,2} , Anne Untermann², and Henning Baars² 

¹ Graduate School of Excellence Advanced Manufacturing Engineering (GSaME), Nobelstr. 12, 70569 Stuttgart, Germany

dimitri.petrik@gsame.uni-stuttgart.de

² University of Stuttgart, Keplerstr. 17, 70174 Stuttgart, Germany

{anne.untermann,henning.baars}@bwi.uni-stuttgart.de

Abstract. Organizations must gain insights into often fragmented and isolated data assets and overcome data silos to profitably leverage data as a strategic resource. Data catalogs are an increasingly popular approach to achieving these objectives. Despite the perceived importance of data catalogs in practice, relatively little research exists on how to design corporate data catalogs. It is also obvious that the existing market solutions have to be customized to the specific organizational needs. This paper presents a list of functional requirements for enterprise data catalogs extracted from a systematic literature review. The requirements can be used to frame and guide more specific research on data catalogs as well as for system selection and customization in practice.

Keywords: Data catalog · metadata · metadata management · requirements

1 Introduction

Recent technological developments in cloud provisioning, analytics technologies, and the Internet of Things foster data collection and analytics which in turn create novel opportunities for organizations to gain a competitive advantage [1]. The automotive industry, for instance, is impacted by analytics-based innovations in manufacturing, product design (i.e., connected and autonomous cars), collaborative services, and – based on that – novel business models [2, 3]. In other industries, too, organizations are increasingly trying to monetize their data together with the own employees’ knowledge and are trying to bundle them to knowledge-intensive services [4]. In doing so, refined data acts as a key strategic resource for organizations that supports identifying optimization opportunities and sustainable efficiency gains in business processes [5]. To leverage these opportunities, organizations require integration and harmonization of data within and beyond the organizational boundaries [6].

Consequently, organizations need an overview of distributed data assets to acquire a sufficient understanding of the data inventory already available to fully exploit the potential of refined data [6]. Typically, the available data is fragmented. It is stored in a multitude of disparate IT systems by numerous departments as well as external actors,

resulting in isolated data silos. Data silos are also a significant hurdle to overcome as suppliers, customers, and the manufacturing organizations themselves are trying to form data ecosystems with big data analytics that lead to even more complex data landscapes. Increasing complexity and, at the same time, decreasing transparency about existing data inventories hamper the discoverability of meaningful datasets and obscure important information about the interrelationships of data, as well as collaboration possibilities of actors, remain hidden. The search processes for relevant data have become long and costly [7]. This, in turn, firstly impedes the provision of knowledge services. Secondly, it prevents relevant initiatives e.g., for self-service analytics and data democratization, in which employees of operational departments are directly involved in value creation and empowered to perform analytics and share data assets without dedicated data experts [8, 9].

To overcome these challenges, organizations require robust data management concepts [10]. Data catalogs are established solutions to tackle those [9]. A data catalog is an enterprise system for metadata management and data curation [11]. It functions as a knowledge and collaboration hub, supports organizations in building sovereign data infrastructures in continuously expanding networks [11], and supports data analysts and other data consumers during the search for data sets, storage locations, intended uses, and other essential information, thus ensuring a better understanding of the existing data landscapes [12].

Multiple commercial (e.g., IBM, AWS or Oracle) and open-source (e.g., Apache Atlas) tools for cataloging are available [11, 14]. It needs to be considered that these are designable and customizable systems that usually cannot be applied off-the-shelf and their tailoring and organizational and technical implementation are non-trivial tasks. Despite the criticality of data catalogs for software-intensive business, issues of their design remain largely under-researched [8]. An initial analysis of the current scientific research literature reveals a lack of design-oriented research and results regarding the subject of enterprise data catalogs. Existing literature reviews indicate that the current research literature has so far mainly concentrated on domain-specific “open data” topic e.g. in the realms of government data, research data, or geospatial data, and is therefore not directly applicable to enterprise scenarios [15]. This state reveals a research gap in the design of enterprise data catalogs, especially in the industrial and inter-organizational data ecosystem contexts. Therefore, we ask: *What are the relevant requirements to design enterprise data catalogs?*

Reflecting on the state of research on data catalogs in the enterprise context, confirms the need for further scientific research on the design and implementation of enterprise data catalogs. For this reason, this paper particularly aims to identify and extract functional requirements for enterprise data catalogs from a systematic analysis of the scientific body of knowledge.

2 Data Catalogs and Metadata Management

Enterprise data catalogs are recognized as enterprise information systems to collect, create and maintain contextual information (i.e., metadata) from heterogeneous source systems [15]. They are context-specific digital data directories in which metadata, i.e.,

data about data, for all existing data objects can be stored centrally and managed securely in order to catalog them in a way that adds value [5]. In an enterprise architecture, data catalogs complement other existing systems for working with data. Functional models often see data catalogs as complementary to data lakes and they are sought to ensure that the data lakes remain manageable and do not become data swamps [10, 16]. They are usually stand-alone software systems (as evidenced by the existing software product landscape [11]) that work hand-in-hand with other data-related subsystems of an enterprise data architecture. For instance, while data quality tools specialize in identifying data problems and fixing them (e.g., through format alignment, standardization, cleansing, and profiling) [17, 18], data catalogs can make the qualified data assets accessible to different roles [11]. In the cross-organizational context of data ecosystems, data catalogs function, for example, complementary to data marketplaces, which provide data brokerage services [10], integrated in interoperable data platforms [11, 19]. To conclude, data catalogs are an integral part of data-driven solutions and thus of software-intensive business, supporting business intelligence and analytics within enterprises or a data ecosystem.

In the existing academic research literature, enterprise data catalogs are associated with data democratization. “Data democratization” implies that non-IT employees are given access to existing data sets and are empowered to use them for data-driven purposes [8]. Accordingly, by providing a conceptual structure as well as various data access functions, data catalogs should facilitate **findability, accessibility, interoperability, and reusability** (FAIR principles) of data assets for the different casual and technical (i.e., analytics experts) users to support the democratization of data. In the literature, this is considered one of the core benefits of their deployment. For this purpose, data catalogs can provide appropriate search mechanisms so that users can discover data sets for their specific use cases [8]. A pertinent design of a data catalog should therefore ensure that the different users can find out which data objects are registered and provide consistent descriptions of the data assets and their locations [8, 20]. Therefore, data catalogs simultaneously function as abstractions of various documentation levels and thereby should facilitate a centralized data access point within and across organizational borders (in a setting with a data catalog that supports a data ecosystem) [11]. Once a user has identified appropriate data sets, they should be made accessible directly through the data catalog. Since data catalog implementation aims to make data from different domains and previous data silos available and usable, ensuring the comprehensive quality of data sets scattered in heterogeneous source systems [21], an **assessment of the quality** of the registered objects plays an eminent role, as this is the only way to generate actual added value for the data consumer. The main component of a data catalog to make data searches possible is the so-called **data inventory**, which models and describes the available data supply [8]. Data might be manually captured by users or automatically collected through interactions with the respective source systems; particularly when pre-built metadata models foster a standardized data capture [8, 22]. Another essential aspect of the data inventory is the detailed documentation of the data sequence (also known as **data lineage**). Data lineage describes the ability to trace data records back to their original source, i.e., data provenance [5, 15, 22, 23]. Because data catalogs are intended to replace manual searches, they should be able to consolidate and **automate**

the corresponding processes which are otherwise often time-consuming and inefficient [8, 23, 24].

Since enterprise data catalogs support metadata management, this section also presents the related work on metadata. Metadata includes information about data sets and can be generated either manually by the data creator or automatically by a system. Metadata can include information about the data creator, record contents and contexts, or timestamps of data creation [25]. In data management, metadata is significant in facilitating access, management and sharing of structured and unstructured data [26]. The National Information Standards Organization (NISO) supports this statement and adds that consistently maintained and structured metadata are used, on the one hand, to help users find appropriate data sets in heterogeneous data structures of information systems and, on the other hand, to capture and subsequently share essential information about these data, thereby promoting data understanding and transparency [27]. Three metadata types can be distinguished [27]:

- Descriptive metadata (1) provides information about the content of data sets and makes it easier for data consumers to identify and understand appropriate data objects for their specific use or research purpose. Exemplary metadata elements are titles, descriptions, or keywords.
- Administrative metadata (2) is a collective term for data related to managing or creating data sets and can be divided into three segments: 1. Technical metadata, such as information about the physical structure of the data set, such as file format, software used, or encoding; 2. Legal metadata, such as information about access rights, copyright restrictions, or intellectual property rights; 3. Data provenance metadata, such as information about the lineage, last modifications, and reasons for the creation of the data set. The information provided thus assists users in interpreting the identified datasets.
- Structural metadata (3) represents the relationship and interaction between the sub-elements of the data set, such as the hierarchy levels or foreign-key-relationships.

Other metadata classifications may also be useful for the discovery of data sets. For example, metadata can be divided into business metadata (i.e., information about the business context and policies), operational metadata (i.e., the information generated automatically during data processing, such as the information about data quality), and technical metadata (i.e., information about the data structure such as the data format or scheme) [28, 29]. This classification can be beneficial because business metadata promotes data understanding by technical or non-technical-savvy staff and enhances interdisciplinary exploration and interpretation of data sets, while operational metadata enables the derivation of insights related to quality development, security, and compliance, and technical metadata is used to document data composition and types [23]. The different existing metadata typologies are often interrelated and, therefore, not always generated and documented separately [29]. Finally, it is helpful to reconstruct the lifecycle of data elements through consistent metadata to enable the search of data objects within complex information systems. Thus, metadata promises to provide real economic value when, for example, it is at least partially automated, and previously collected information is reused to avoid redundant or obsolete metadata and streamline the curating process [30]. When metadata is generated in a way that is readable by both

machines and humans, it promotes interoperability and integration of metadata on the one hand, and allows data sets to be described, discovered, and contextualized [25, 27, 30]. To achieve this, enterprise data catalogs represent the information systems to realize metadata documentation and provisioning [24].

3 Methodology

As a literature review aims to synthesize the existing state of knowledge on a selected phenomenon, we consider it to be a suitable research methodology for extracting functional requirements for enterprise data catalogs as a form of codified design knowledge. We follow established guidelines for a systematic concept-centric literature review on a database level [31]. For the definition of the sample of relevant literature sources, we started with an unsystematic literature search on Google Scholar EBSCOhost and ScienceDirect (with the generic search terms “Data Catalog”) which helped us pinpointing more specific search criteria. From the results we refined the following keywords: ‘data catalog’, ‘metadata catalog’, ‘enterprise’, ‘data repository’, and ‘data register’. The publication period was set to 2006–2023 as data catalogs in their current form represent a relatively new concept. Another relevant selection filter was the accessibility of the publications as well as a focus on conference and journal contributions (academic journals, conference papers, or proceedings): We tried to avoid that incomplete texts, non-accessible papers, or non-peer-reviewed articles. In total, we formulated two search terms that we applied separately across the five databases Web of Science, SpringerLink, ACM Digital Library, IEEEExplore, and AISeL:

1. “data catalog*” OR “metadata catalog*”
2. “data catalog*” AND enterprise

This generated a total of 750 hits with the first search term and 11 with the second. After applying the aforementioned filter criteria, the sample for the first search string was 408 papers, and for the second search term 10 papers. After excluding the duplicates, the sample went down to 391 papers. In the next step, the titles and abstracts were manually analyzed to determine whether they fit the research question and indeed have “data catalogs” as their research subject. Articles dealing with data catalogs in the domains of medicine, politics, astronomy or geography were excluded, as they do not deal with corporate and industrial contexts of use of data catalogs. Nevertheless, a few articles from these research areas were retained if they contained information that could be transferred to the entrepreneurial context. Since the titles and abstracts were often not meaningful, we performed diagonal reading to minimize subjectivity. Here, the introductions, the conclusion of the articles, and the figure and table titles used were examined with respect to the inclusion and exclusion criteria. A total of 45 articles remained. After reading the full texts, a backward search resulted in six additional articles. After the full-text screening, additional papers were removed from the sample that for instance only described projects with happened to include data catalogs. The authors discussed each paper of the initial sample, seeking a consensus within the research team to increase the objectivity of the exclusion. In doing so, the final sample was reduced to 21 relevant articles.

Due to the limited amount of scientific literature on data catalogs in the enterprise context, we broadened our search and explicitly included grey literature, esp. White papers and research reports. After all, white papers and practice reports are considered recognized explanations of practice, which can prepare qualitative expertise and recommendations regarding a specific topic in a consolidated manner. Thus, adhering to standard guidelines for including grey literature in systematic literature reviews [32], we have broadened our sample by including only grey literature with high credibility and high outlet control. Our selection criteria exclude marketing documents from tool providers, focusing solely on reports from reputable research institutes or established management consultancies that are known for leveraging software- and data-driven projects. In addition to assessing the authority of the sources, our inclusion of grey literature was also guided by the perceived objectivity of their statements. In this way, three additional publications could be added. Due to length constraints, the literature sample compiled is detailed in an external appendix, accessible via the following URL: <http://bit.ly/49Jbbp5> (Fig. 1 illustrates the sample creation process).

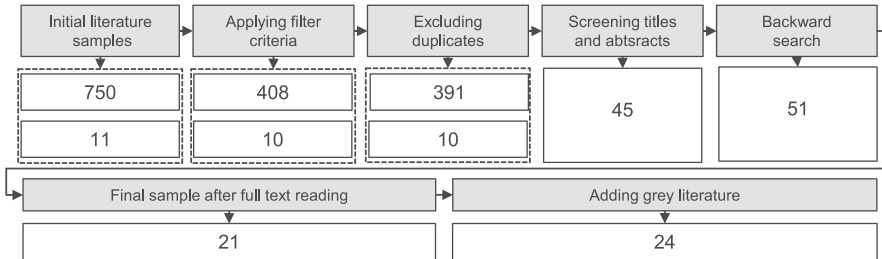


Fig. 1. Illustration of the literature search process and sample creation

During the content analysis of the remaining papers [33], we inductively formed categories for the derivation of functional requirements, guided by the expertise within the research team. According to the inductive technique, the abstraction level is successively increased to develop theory-based main categories from a large number of groupings from the available texts. Each researcher independently reviewed the articles in the created sample, applying coding techniques and labeling the functionalities. These codes were then collectively discussed by the research team to foster a shared understanding and to collaboratively formulate the requirements. In this process, a total of 13 functional requirements were derived.

4 Requirements for Enterprise Data Catalogs

The derived requirements have been grouped into the following six categories, each represented by a unique identifier: metadata management (Requirements R1-4); data inventory (Requirements R5-6), data governance (Requirements R7-9), interoperability (Requirement R10), interface (Requirement R11), collaboration (Requirement R12), intelligent automation (Requirement R13). The requirements were grouped based on

their functional similarity during discussions within the researcher team. Figure 2 integrates the requirements in a functional view on an enterprise data catalog, embedded either in a data lake or in a data platform, based on [11]:

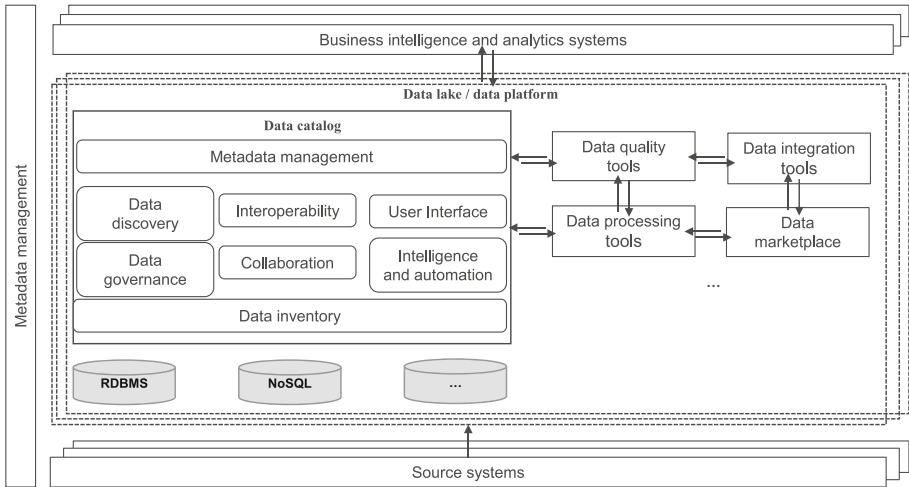


Fig. 2. Functional view on enterprise data catalogs

Data catalogs function as central indexed searchable sources for finding data [8, 24]. To ensure successful and seamless data set searches, robust search functionalities should be integrated into data catalogs that enable users to find data objects for a specific analytics purpose [22, 34]. In particular, the search for keywords, business terms, or metadata should be offered. In addition, using functions that utilize a natural language simplifies the search for data consumers of a non-technical domain [22, 25, 35]. This includes, for example, full-text or semantic search (which is also used in Google searches) to deal with the content of search queries. Designations or titles of data sets, data domains, or business units are first classified and then indexed, resulting in the display of data relating to the content entered [23, 36].

In addition, the role-specific requirements of individual users should be included to avoid missing necessary functionalities or integrating superfluous functions that hinder the search [22]. This results in the following requirement:

R1: *Enterprise data catalogs should be equipped with robust search functionality to enable employees to identify needed data sets by entering, for example, keywords, metadata, or full text, considering role-specific search requirements.*

Furthermore, data catalogs should allow the user to enrich the recorded data objects with complementary information to improve the findability of the data sets and to facilitate the search by giving additional clues about how data objects are related. Finally, high information content promotes the user understanding of data sets and makes data knowledge more consumable. Accordingly, it should ideally be possible to associate data with labels, *identifiers*, and to link them to a searchable source, which provides additional insights into the content and the characteristics of the data [8, 13, 20]. Essential for

indexing is an adequate description of the information about the individual data objects, whereby priority should be given, for example, to the descriptions' completeness, simplicity, and relevance. Based on this information, users can decide whether the data sets are suitable for the respective analytics projects, so the description should be created rather carefully [37]. Tagging functions also improve discoverability significantly [5, 13, 38]: The data is labeled, and it is determined on which level the previously defined metadata variables or attributes are assigned to the respective data. [5, 15]. Depending on the context of the use of the data catalog, data can be tagged at four levels: dataset level (original source dataset), record level (for all data entries in the dataset), entity level (for each data entry), and column level (individual columns in the dataset) [5]. This results in the following requirement:

R2: *Enterprise data catalogs should allow the linking of registered data objects by data providers to adequate identifiers and appropriate indexes to ensure data discovery and facilitate the evaluation of data sets by system users, particularly if the data catalog consolidates data objects from different usage contexts.*

Besides, data catalogs must support metadata documentation while supporting the applicable metadata standards (if applicable). To enable reusability of data objects by aligning enterprise and system-oriented views of data, a complete documentation of metadata should be based on a conceptual (i.e., the context of the creation and the application of the data), a logical (i.e., entities and their relationships to each other as well as associated business objects and attributes) and a physical level (i.e., information related systems, interfaces, data structures and attributes etc.) [8, 21]. Constructive here would be the enrichment of the data with contextual information that can (1) describe the operational context in terms of the domain or subject area in which the data operate, on the one hand, and (2) characterize the technical context through technical details regarding the data source or data set, on the other [5, 15, 36].

R3: *Enterprise data catalogs should promote a unified understanding of data sets for all user groups by documenting metadata on multiple levels, distinguishing between the conceptual, logical, and physical documentation levels, in order to support heterogeneous user groups in retrieving data.*

Following common metadata standards is also recommended when designing data catalogs. These can be public domain-independent metadata standards or ontologies [8, 15]. Standards promote homogeneous access across heterogeneous descriptions and support data interoperability at the user level [25]. In this way, the utility of data objects is improved, and data consumers and producers are linked by building a common consensus [15, 37]. This influences the interoperability of catalog systems and promotes compliance with FAIR Principles [15]. Concerning the system infrastructure of data catalogs, various metadata standards have already been established, which can be applied in combination depending on the context of use. According to [8], these include the Dublin Core Schema (DC), the Data Catalog Vocabulary (DCAT), the ISO 11179-3 Metadata Registry Metamodel and Basic Attributes (MDR), and the Common Warehouse Metamodel (CWM). Consequently, the requirement is as follows:

R4: *Enterprise data catalogs should support metadata standards to provide users with adequate search results and seamless access to heterogeneous data sets.*

Implementing a business glossary offers advantages for the value and acceptance of the data catalog among users. Clear business terms help to understand the context of the use of the data objects and the data itself by employees of the departments [8, 15, 21, 24]. Business glossaries are central repositories containing key business terms agreed upon by cross-functional subject matter experts [15]. On the one hand, company-wide terms, objects, and attributes can be explained, and on the other hand, domain or business unit-specific terms can be defined [21, 23]. To further optimize the interpretation of the data and their usage environments, the created metadata here can also further be enriched by additional context variables [15]. As a result of a better understanding, the data sets can subsequently be used or adapted for other analysis projects, which is an essential prerequisite for the reusability of the data sets.

R5: *Enterprise data catalogs should be equipped with a complementary business glossary to describe the data objects from an operational perspective to create a uniform understanding regarding specific terms for all user groups and to prevent misinterpretations, given the fact that the user groups come from different domains or companies and have different expertise.*

As integrated platforms that link the various data-oriented user groups (e.g., data owners and data analysts) and enable informal information exchange, it also makes sense to provide efficient data management functions in a centralized manner. These include registration functionalities such as “data connectors” that enable the automatic collection of metadata from source systems or “data imports” that independently import the descriptions of data sets from data tables, which can significantly reduce time-consuming tasks [23]. Furthermore, there are functions for data organization and management (curation of data) that enable, for example, annotations or tags, the creation of metadata, or the labeling of security- and compliance-relevant data [34]. Adding tags or compliance-related information can also influence catalog user collaboration by transparently sharing knowledge and expertise and improving search results. This results in the following requirement:

R6: *Enterprise data catalogs should be equipped with a comprehensive range of data management functions, such as data object registration and curation functions, to facilitate the integration into, the administration of and navigation among the meta data sets.*

Data catalogs are commonly seen as necessary for the implementation of a data governance. This in turn implies that the definition of an enterprise-wide data governance is closely intertwined with the data catalog design. On the one hand, a data governance fosters (or even enforces) compliance with internal and external data management regulations and data protection guidelines and, on the other hand, can support the definition of technical standards to ensure interoperability and thereby maximize data value [22, 23, 39]. In conclusion, data catalogs should fulfill prerequisites that contribute to the implementation of the defined data governance [22]. In this field, the documentation of ownership is an essential prerequisite for assessing responsibilities. This has two benefits. Firstly, contact persons can be identified and contacted directly in case of error occurrences or violations of the defined guidelines. Secondly, contact persons promote collaboration between data consumers and data providers [5, 22]. In addition, knowledge regarding ownership provides information on the relationships between data sets,

allowing important insights to be derived for potential synergies [39]. Thus, ownership representation creates transparency and establishes collaboration opportunities between data consumers and providers. This way, contact persons can be accessed directly in case of questions or problems. In addition, a role model acts as an important prerequisite for system-wide collaboration, as tasks can be distributed and responsible users identified. The following requirement is derived from this:

R7: *Enterprise data catalogs should support clear and consistent data governance structures, including unambiguous role models, ownership, and policies regarding data quality and data provenance that act as an organizational framework to ensure the responsible use and management of data sets.*

Access control mechanisms are central for protecting sensitive data from misuse and complying with regulations [15, 34]. This is true for all data bases but data catalogs in particular which is why their design should include data access functionality. This can include automated workflows for approval processes and user authentication mechanisms [8, 15, 25, 40]. Such functionality ensures that the visibility of catalog content needs to be unlocked by access requests and the assignment of appropriate access keys [5, 41]. As a more recent development, Artificial Intelligence (AI) can be used to identify sensitive or secret data by assigning attributes or to display data sets that are not accessible to the user [15, 23, 24]. Another prerequisite for access control is the definition of user groups and role-specific data authorization levels through which suitable approval processes can be created [21, 23]: Data catalogs should document the approval history and reasons for the access request to analyze the contexts of use of the data and trace potential compliance violations [8].

R8: *Enterprise data catalogs should be equipped with reliable mechanisms for role-specific access controls, secure process flows, and usage policies that regulate data usage, management, and access in terms of security and privacy and that allow only authorized users to access data sets to prevent sensitive data from being misused.*

In addition, data catalogs should ensure the quality and reliability of data and meta-data through various functions. Ideally, the tools encourage the users to define quality standards and measurable data quality metrics in advance and allow to continuously check them later. This way, errors, deviations, and duplicates can be detected early after launching a data catalog [23, 39]. Dashboards can also be a valuable tool for the support of data quality management activities as they can graphically display quality metrics for the selected data sets, visualize developments over time, and signal issues with alerting mechanisms [23, 24]. It should also be possible to add new quality rules or modify existing ones [23]. To ensure the quality of the data in the long term, the users need to continue developing procedures for the maintenance and upkeep of the data sets, including clear responsibilities for each individual process instance. By doing so, it can already be ensured during the context of the design that the catalog system that it can provide coherent and valuable data sets over the entire life cycle of the data catalog [15, 22].

R9: *Enterprise data catalogs should provide adequate control mechanisms in the form of qualitative standards, guidelines, and predefined quantitative data quality metrics that can be continuously reviewed to avoid unreliable or erroneous data objects within the data catalog system.*

Furthermore, there is a need to embed data catalogs in existing infrastructures so that data consumers have standardized access to distributed resource descriptions and information systems [25, 38]. Two building blocks are necessary to ensure sufficient interoperability. Firstly, data catalogs should be equipped with standardized application programming interfaces (APIs) to access the source systems [8, 21, 35, 39]. Of particular interest are interfaces to other data catalogs (especially in large organizations or data ecosystem settings) and the functionality to connect with leading enterprise systems (i.e., ERP, CRM, SCM, CRP, or MES) as well as with business intelligence tools [11]. Secondly, uniform standards, schemas, terminologies, and formal and comprehensively applicable languages for the description of data sets and metadata should be used [15, 24, 25, 37].

R10: *Enterprise data catalogs should incorporate standardized application programming interfaces to query the data sets, their description, and metadata to facilitate the integration into existing technical infrastructures and source systems and give access to different functional units of an organization.*

Since data catalogs should enable both technical and non-technical expert users to access data, user-friendly graphical user interfaces (GUI) are a common essential requirement. Ideally, those GUIs can be parameterized depending on the respective user role [23]. Additionally, data catalogs can include visualization functionalities that advance an understandable and descriptive representation of data sets, metadata, terminology, and data sequences. Data flow diagrams or knowledge graphs have proven to be a viable tool for this [22, 24]. Existing empirical research on data catalog suggests that data analysts value graphical representations of entire metadata collections and logging of historical queries to save users (especially inexperienced ones) the effort to develop queries [16].

In addition, data exploration and visualization tools can be used to display quality metrics or other KPIs in dashboards. They support users in evaluating and analyzing the data [8]. The visualization should enable the various user groups, especially data analysts, to derive insights from the data sets recorded in the data catalog that can contribute to data-related decision-making and the quality assessment and improvement of the data objects.

R11: *Enterprise data catalogs should foster digital interactions of data consumers through intuitive digital user interfaces that meet the needs of non-technical user groups and are thus customizable and allow visualization of data sets.*

Another goal of data catalogs is to promote the collaboration between different data users by providing functions for the exchange of practice-related knowledge and, if necessary, its transfer to other data projects [23]. The progression of transparency regarding the company's existing data objects is crucial to developing a collaborative environment. A characteristic of this is that data sets become traceable and findable for the various user groups [24]. Comment, tagging or rating functions, as well as workflows or discussion forums are useful for promoting communication and collaboration between users of data catalogs [8, 22, 23]. In addition, chat functions can be helpful in establishing direct contact with data owners or contacts and allow clarifying ambiguities or sharing feedback regarding the quality or usefulness of the data [8, 22]. Functionalities for registration, publication, search, filtering, and localization of data sets are additional pillars for a successful data collaboration [35, 42, 43]. In this context, role-specific

functions can be offered that support the fulfillment of the respective tasks and meet the needs of the different user groups [22]. Possible functionalities would be the provision of data preview to gain initial insights into the contents of data sets, the possibility to follow data sets and receive notifications of changes, or recommendations based on previous search queries or user behavior [8, 22, 34]. However, these functions should be provided modularly to offer users only functions that clearly support the specific user role without overstressing the user.

R12: *Enterprise data catalogs should be modularly equipped with collaboration and communication features that enable synergies between data-driven user groups and promote collective decision-making so that users with different levels of knowledge and experience can make better data-based decisions.*

The analysis of the selected publications clearly shows that a high degree of automation is indispensable to achieve the sustainable performance of the data catalog by implementing the previously presented requirements with sufficient performance. There are various use cases for automation in data catalogs, particularly concerning data-driven analysis projects. For example, processes can be automated by incorporating workflows (e.g., approval processes for changes or access requests), or machine learning or artificial intelligence (AI) algorithms can be used in detecting anomalies and causes of errors, analyzing data, or generating insights and recommendations regarding data sets [8, 24]. Furthermore, data description, context enrichment, and metadata generation can be supported using automated approaches. Here, the implementation of machine-based dataset profiling techniques is recommended, with the option to automatically create data profiles [36]. Regarding the principle of “reusability,” an automated documentation of generated analyses results can further be used to derive lessons learned or leverage analysis data for more advanced projects [8]. A nuanced reconstruction of the lineage of data sets can also be recorded in an automated manner, increasing the transparency of the origin of data objects and promoting trustworthiness in the data [23]. The automation dimension indicates that support functions such as AI are needed to facilitate data registration and curation. Furthermore, this has the added benefit that company-wide data catalogs become scalable without losing consistency or accuracy [22, 23, 44]. However, it should be considered that the analytics methods often need to be tailored to the targeted analysis contexts.

R13: *Enterprise data catalogs should be equipped with intelligent automation functions to reduce time-consuming and manual activities of data discovery, analysis, and use on the part of data consumers and time-consuming and manual activities of data management and maintenance on the part of data providers.*

5 Conclusion

Enterprise data catalogs are a “hot topic” in practice to support metadata management. This study elaborates and categorizes a set of 13 functional requirements systematically derived from scientific literature and three practical studies. The main goal of this article is to present a list of relevant functional requirements for practitioners who make decisions on the implementation and tailoring of enterprise data catalogs, to improve their design and increase their acceptance by potential users. The requirements support IT

decision makers in designing and customizing data catalogs to support the integration of data into software-intensive services [3, 4] for the facilitation of software-intensive business operations.

Considering the structure and the priority of these requirements, they cover on a foundational set of base requirements that are crucial for the overall functionality of a data catalog. These are at least partially met by existing open-source or commercial tools. The set of requirements also covers key technical functionalities for data storage, access, and management. Without these, the more user-oriented ones would not work as well, revealing also a natural hierarchy within the requirements set. The different target groups (end users, system operations, database administrators, developers) and their use cases build the foundation for sorting the requirements situationally.

We argue that while our focus originates from an enterprise context, the adoption of data catalogs is also becoming increasingly relevant for non-commercial organizations such as government institutions and nonprofit organizations. In this context, we consider data catalogues as enablers for inter-organizational networks and data ecosystems. This is exemplified in the existing data space or data cooperative initiatives to enable scenarios, such as circular economy, which highly rely on sharing metadata resources at scale [45, 46]. The derived functional requirements are not limited to a particular domain or scenario, and can therefore be used in data-driven scenarios in different domains, although specific tailoring might be necessary. It is also important to consider how the nature of such ecosystems evolves when data catalogues become machine-readable, enhanced by the natural language processing capabilities of current Large Language Models (LLMs). Such advancements enable the connection, processing, and utilization of data in these catalogues with minimal human intervention.

Furthermore, the requirements also help service providers and data catalog solution providers with the integration and customizing of data catalogs. Hence, we are confident that the derived requirements support the value proposition deployment of software companies that offer enterprise data catalogs as software products. Our requirements can also be linked to the Fraunhofer ISST functional model, extending it with prescriptive statements about the functionalities that data catalogs must provide [22]. The requirements can be used for context-specific benchmarks and act as a checklist for system designs or development projects. In addition, the requirements provide a starting point for future design-oriented research on data catalogs. To the best of our knowledge, existing data catalog tools only cover the set of requirements only in a basic manner, especially those focused on end-users (R11-R13). This highlights a significant gap that needs to be addressed.

However, the requirements are mainly limited to the scientific literature, which at this point in time, has done relatively little research on data catalogs. Thus, these results present a synthesized knowledge of the literature but without integration of project experience knowledge from the field. Since domain-specific restrictions (e.g., related to interoperability, standardization or data governance) are not included, the requirements catalog is not exhaustive. Yet, the presented requirements build a foundation for further empirical research on the design of data catalogs capturing domain constraints.

Nevertheless, the requirements catalog should be validated and extended in further studies, especially through empirical cases or the analysis of existing data catalog systems

in order to capture seemingly “trivial” requirements or requirements that reflect the dynamics of the field [14]. The latter is a particular problem given the breathtaking speed at which new AI solutions are introduced to the market which support IT-processes in particular. Therefore, we expect that those reshape the functionality of data catalogs and alter the elicited requirements significantly in the mid-term future. Given R1, it can be assumed that search functionality can be expected to benefit considerably in the near future by applying so called large language models that provide both a more user-friendly natural language interface and can extract semantic similarities. Accordingly, future studies should explore solution approaches for novel AI functions for data catalogs for the new levels of data catalog automation, their effectiveness, shortcomings, and their acceptance. In addition, future research can also explore best practices and strategies for implementing enterprise data catalogs. Ideally, this is done by utilizing the action design research approach in order to combine practical requirements, innovative solutions, and theoretical rigor.

References

1. Legner, C., et al.: Digitalization: opportunity and challenge for the business and information systems engineering community. *Bus. Inf. Syst. Eng.* **59**(4), 301–308 (2017)
2. Dremel, C., Wulf, J., Herterich, M.M., Waizmann, J.-C., Brenner, W.: How AUDI AG established big data analytics in its digital transformation. *MIS Q. Exec.* **16**(2), 81–100 (2017)
3. Hunke, F., Heinz, D., Satzger, G.: Creating customer value from data: foundations and archetypes of analytics-based services. *Electron. Mark.* **32**, 503–521 (2022)
4. Ksouri-Gerwien, C., Ebel, M., Bittner, K., Poepplbuss, J.: Offering knowledge as a service – a taxonomy of knowledge-intensive business services. In: *Proceedings of the 31st European Conference on Information Systems, Kristiansand (2023)*
5. Shanmugam, S., Seshadri, G.: Aspects of data cataloguing for enterprise data platforms. In: *2nd International Conference on Big Data Security on Cloud*, pp. 134–139. IEEE (2016)
6. Otto, B., Jarke, M.: Designing a multi-sided data platform: findings from the International Data Spaces case. *Electron. Mark.* **29**, 561–580 (2020)
7. Gluchowski, P., Gonschorek, E.: Data Catalog – Transparenz durch Dateninventarisierung. *Rethinking. Finance* **3**, 11–14 (2019)
8. Labadie, C.: *Essays on Data Democratization & Protection in the Data-driven Enterprise*. Doctoral thesis, University of Lausanne (2021)
9. Eichler, R., Gröger, C., Hoos, E., Schwarz, H., Mitschang, B.: Data shopping – how an enterprise data marketplace supports data democratization in companies. In: De Weerd, J., Polyvyanyy, A. (eds.) *International Conference on Advanced Information Systems Engineering (CAISE) Forum. LNBP*, vol. 452, pp. 19–26. Springer, Cham (2022)
10. Eichler, R., Giebler, C., Gröger, C., Schwarz, H., Mitschang, B.: Modeling metadata in data lakes – a generic model. *Data Knowl. Eng.* **136**, 101931 (2021)
11. Jahnke, N., Otto, B.: Data catalogs in the enterprise: applications and integration. *Datenbank-Spektrum* **23**, 89–96 (2023)
12. Spezzati, A., Kheradmand, E., Gupta, K., Peras, M., Zaminpeyma, R.: Note: leveraging artificial intelligence to build a data catalog and support research on the sustainable development goals. In: *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*, pp. 579–584 (2022)
13. Dibowski, H., Schmid, S., Svetashova, Y., Henson, C., Tran, T.: Using semantic technologies to manage a data lake: data catalog, provenance and access control. In: *Proceedings of the*

- 13th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS 2020), Athens, pp. 65–80 (2020)
14. Zaidi, E., De Simoni, G., Edjlali, R., Duncan, A.D.: Data catalogs are the new black in data management and analytics. Gartner, Consultancy Report (2017)
 15. Ehrlinger, L., Schrott, J., Melichar, M., Kirchmayr, N., Wöß, W.: Data catalogs: a systematic literature review and guidelines to implementation. In: Kotsis, G., et al. (eds.) Database and Expert Systems Applications - DEXA 2021 Workshops. CCIS, vol. 1479, pp. 148–158. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87101-7_15
 16. Gunklach, J., Michalczyk, S., Nadj, M., Maedche, A.: Metadata extraction from user queries for self-service data lake exploration. *Datenbank-Spektrum* **23**, 97–105 (2023)
 17. Altendeitering, M., Guggenberger, T.: Designing data quality tools: findings from an action design research project at Boehringer Ingelheim. In: Proceedings of the 29th ECIS, Marrakesh (2021)
 18. Ehrlinger, L., Wöß, W.: A survey of data quality measurement and monitoring tools. *Frontiers Big Data* **5**, 850611 (2022)
 19. de Reuver, M., Ofe, H., Agahari, W., Abbas, A.E., Zuiderwijk, A.: The openness of data platforms: a research agenda. In: Proceedings of the 1st International Workshop on Data Economy, New York (2022)
 20. Choi, M.-Y., Moon, C.-J., Jung, S.-J.: Building methods of intelligent data catalog based on graph database for data sharing platform. *ICIC Int.* **11**(1), 953–959 (2020)
 21. Mamrot, S., Nowak, F., Ryzyszczak, K., Kaczmarek, Ł., Krzywy, J.: Applying central data catalogues to implement and maintain digital public services. a case study on catalogues of public administration in Poland. In: Janssen, M. et al. (eds.) Electronic Government. LNCS, vol. 13391, pp. 31–46, Springer, Cham (2022). https://doi.org/10.1007/978-3-031-15086-9_3
 22. Jahnke, N., Spiekermann, M., Ramuzeh, B.: Data Catalogs. Implementing Capabilities for Data Curation, Data Enablement and Regulatory Compliance. Fraunhofer Report (2022)
 23. Russom, P.: The Data Catalog’s Role in the Digital Enterprise. TDWI Checklist Report (2017)
 24. Labadie, C., Eurich, M., Legner, C., Fadler, M.: FAIR enough? Enhancing the usage of enterprise data with data catalogs. In: Proceedings of the 22nd Conference on Business Informatics (CBI), pp. 201–210. IEEE (2020)
 25. Quimbert, E., Jeffery, K., Martens, C., Martin, P., Zhao, Z.: Data cataloguing. In: Zhao, Z., Hellström, M. (eds.) Towards Interoperable Research Infrastructures for Environmental and Earth Sciences. LNCS, vol. 12003, pp. 140–161. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52829-4_8
 26. Kerhervé, B., Gerbé, O.: Models for metadata or metamodels for data? In: Proceedings of 2nd IEEE Metadata Conference, Silver Spring, pp. 1–12 (1997)
 27. Riley, J.: Understanding Metadata. What is metadata and what is it for? <https://groups.niso.org/higherlogic/ws/public/download/17446/Understanding%20Metadata.pdf>. Accessed 26 Feb 2023
 28. Oram, A.: *Managing the Data Lake*. O’Reilly (2015)
 29. Diamantini, C., Giudice, P.L., Musarella, L., Potena, D., Storti, E., Ursino, D.: A new metadata model to uniformly handle heterogeneous data lake sources. In: Proceedings of the 22nd European Conference on Advances in Databases and Information Systems (ADBIS 2018), pp. 165–177 (2018)
 30. Research Data Alliance Homepage. <https://www.rd-alliance.org/groups/metadata-ig.html>. Accessed 26 Feb 2023
 31. Tranfield, D., Denyer, D., Smart, P.: Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *Br. J. Manag.* **14**(3), 207–222 (2003)

32. Garousi, V., Felderer, M., Mäntylä, M.V.: Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Inf. Softw. Technol.* **106**, 101–121 (2019)
33. Mayring, P.: *Qualitative Inhaltsanalyse: Grundlagen und Techniken*, Beltz (2015)
34. Wells, D.: *The Ultimate Guide to Data Catalogs*. White Paper of the Eckerson Group (2018)
35. Lapi, E., Tcholtchev, N., Bassbouss, L., Marienfeld, F., Schieferdecker, I.: Identification and utilization of components for a linked open data platform. In: *IEEE 36th Annual Computer Software and Applications Conference Workshops*, Izmir (2012)
36. Skopal, T., Klimek, J., Necasky, M.: Improving findability of open data beyond data catalogs. In: *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, pp. 413–417 (2019)
37. Barbosa, E.B., Sena, G.: Scientific data dissemination a data catalogue to assist research organizations. *Ciência da Informação* **37**(1), 19–25 (2008)
38. Stillerman, J., Fredian, T., Greenwald, M., Manduchi, G.: Data catalog project—a browsable, searchable, metaIndata system. *Fusion Eng. Des.* **112**, 995–998 (2016)
39. Joshi, D., Pratik, S., Rao, M.P.: Data Governance in Data Mesh Infrastructures: The Saxo Bank Case Study. In: *Proceedings of the International Conference on Electronic Business*, Nanjing (2021)
40. Lefebvre, H., Legner, C., Fadler, M.: Data democratization: toward a deeper understanding. In: *Proceedings of the 42nd International Conference on Information Systems*, Austin (2021)
41. Czajkowski, K., Kesselman, C., Schuler, R.: ERMrest: a collaborative data catalog with fine grain access control. In: *13th International IEEE Conference on e-Science*, Auckland (2017)
42. Shi, C., Zhang, Y., He, R.: Design and implementation of a P2P resource sharing system based on metadata catalog. In: *Proceedings of the 9th International Symposium on Computational Intelligence and Design*, Hangzhou (2016)
43. Holl, P., Gossling, K.: Midas: towards an interactive data catalog. In: Gadepally, V., et al. (eds.) *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*. LNCS, vol. 11721, pp. 128–138. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33752-0_9
44. Labadie, C., Fadler, M., Eurich, M., Legner, C.: All hands on data: a reference model for enterprise data catalogs. In: *Essays on Data Democratization & Protection in the Data-Driven Enterprise*, pp. 71–108 (2021)
45. Serna-Guerrero, R., Ikonen, S., Kallela, O., Hakanen, E.: Overcoming data gaps for an efficient circular economy: a case study on the battery materials ecosystem. *J. Cleaner Prod.* **374**, 133984 (2022)
46. Jäger-Roschko, M., Petersen, M.: Advancing the circular economy through information sharing: a systematic literature review. *J. Cleaner Prod.* **369**, 133210 (2022)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

