IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 February 2025, accepted 22 April 2025, date of publication 9 May 2025, date of current version 19 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3568542

SURVEY

A Survey on the Functionalities of Data Catalog Tools

JASMIN KROPSHOFER[1], JOHANNES SCHROTT[2], WOLFRAM WÖß[1], AND LISA EHRLINGER[3]

[1]Institute for Application-oriented Knowledge Processing, Johannes Kepler University Linz, 4040 Linz, Austria
[2]Software Competence Center Hagenberg GmbH, 4232 Hagenberg, Austria
[3]Hasso Plattner Institute, University of Potsdam, 14482 Potsdam, Germany

Corresponding author: Jasmin Kropshofer (jasmin.kropshofer@jku.at)

This work was supported in part by the Bundesministerium Innovation, Mobilität und Infrastruktur (BMK), Bundesministerium Wirtschaft, Energie und Tourismus (BMAW), and the State of Upper Austria within the scope of the Software Competence Center Hagenberg (SCCH) Competence Center Integrated Software and Artificial Intelligence Systems (INTEGRATE) Forschungsförderungsgesellschaft (FFG) through the FFG Competence Centers for Excellent Technologies (COMET) Program under Grant 892418.

ABSTRACT Finding all data distributed across numerous systems, understanding its meaning, and assessing its quality are major challenges for many companies and organizations. As a result, both researchers and practitioners have become increasingly interested in data catalogs, as such tools maintain a repository of technical metadata annotated with domain knowledge. Data catalog tools thus significantly improve the findability, accessibility, interoperability, and reusability (FAIR principles) of datasets. Currently, there is no generally accepted definition or interpretation regarding the required functionality of data catalog tools. This has not only led to a wide range of so-called data catalog tools but has also made it difficult for practitioners to gain an overview in order to make a targeted selection of a tool. Therefore, the main contributions of this paper are 1) an analysis and discussion of the most important data cataloging functionalities and 2) a systematic survey that investigates the extent to which existing data catalog tools implement these functionalities. The detailed results of this survey (i.e., the identified features, data source connectors, support of artificial intelligence for each data catalog tool) are additionally provided in a table that can be customized, sorted, and filtered. While the evaluation table is intended primarily to support practitioners, and in particular data stewards and data engineers, we want to promote a common interpretation of data catalogs in the scientific community with the results compiled in this paper.

INDEX TERMS Data catalog tools, metadata management, data source description, systematic review.

I. INTRODUCTION

Managing ''data as an asset'', as promoted by data governance initiatives, is becoming increasingly important in enterprises and organizations [1], [2]. One reason is the pivotal role of data in business analytics or computational models created, for instance, by machine learning (ML) methods. Despite the growing awareness of data governance, organizations are still struggling to leverage the value of their data [3], [4], [5]. Many organizations cannot adequately answer the question of which systems store the desired data in which quality level. An important step towards answering these questions (i.e., location and quality level of data) and thus achieving the goals of data management and analysis (cf. [6]) is comprehensive metadata management (i.e., ''data that defines and describes other data'' [7]). Data catalogs play an important role in metadata management by maintaining an inventory of data assets and facilitating the discovery and understanding of data [5]. Although the popularity, and therefore the diversity, of data catalogs has increased steadily since 2016, a generally accepted interpretation or definition of the term remains elusive [2], [3], [5], [8], [9].

In this paper (based on [2], [8], [9], [10], [11]) , we define that a data catalog tool provides at least the following two main functionalities : (i) collection and storage of technical metadata and (ii) annotation of this technical metadata with domain knowledge (business context). To represent data items technically, data catalogs manage structural metadata

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

(e.g., about relational tables and their attributes) as well as descriptive and administrative metadata (e.g., title, creation date, data types) of data sources [8], [12]. In addition to technical metadata representation, annotation with "business context" is an important functionality to enable accessibility and added value of metadata for business users [2], [8]. Annotation of domain knowledge can be implemented using, for instance, tags [13], a glossary [2], or ontologies [14], [15].

The main contributions of this paper are (i) an analysis and discussion of the most important data catalog functionalities, and (ii) a systematic survey that investigates the extent to which existing data catalog tools implement the main data catalog functionalities. First, we aimed to define minimum requirements for a data catalog in order to achieve consensus on a general interpretation within the scientific community. Second, by evaluating data catalog tools, we aimed to support practitioners in selecting a suitable tool.

Besides non-peer-reviewed data catalog tool surveys aimed at practitioners (e.g., [5], [6], [9], [16], [17]), four peer-reviewed surveys of data catalogs were published, three of which ([2], [8], and [18]) addressed the general concept of data catalogs and did not examine specific tools or their functionalities. The fourth study [3] investigated how data catalogs can be integrated into existing enterprise-wide metadata management landscapes. This aspect is important because the large number of tools referred to as "data catalog" and the many additional functionalities that a data catalog offers (beyond the main functionalities mentioned above) make it difficult to select the right tool for specific requirements [9]. In contrast to [3], which focused on the integration into enterprise-wide metadata management landscapes, our study focused on the ability of data catalog tools to model and store metadata, which is particularly important to enable reuse of metadata within an organization, for instance, for data quality checks. Our survey is therefore complementary to [3]. A detailed comparison of the studies and a discussion of the results is provided in Section IV-C1.

For further investigation, we summarize our evaluation in a table available online [19]. The 75 general-purpose data catalog tools identified by our systematic search (see Appendix for the complete list with links to the websites) were analyzed according to the following six aspects:

- type of domain knowledge representation,
- level of detail (granularity) of the metadata of data sources,
- available data source / application connectors,
- extent to which artificial intelligence (AI) is used,
- available licensing options, and
- deployment type(s) offered (on-premise or software-as-a-service).

The remainder of this paper is structured as follows: Section II describes the methodology used to conduct the survey, including the research questions, the search strategy, and the process of extracting tools from the search results. Section III presents the results obtained by the systematic literature review (SLR), and Section IV provides background on data catalog research and related work. The results are discussed in Section V, and Section VI concludes with an outlook on future work.

II. SURVEY METHODOLOGY

The methodology used for this survey is based on the guidelines for systematic literature reviews (SLRs) in [20]. An SLR analyzes and synthesizes findings from existing primary studies to address specific research questions. Since the focus of this work was to provide an overview of the tools (rather than the papers) available, the literature included was also reviewed for mentions of tools. Figure 2 summarizes the complete process of the SLR developed to retrieve data catalog tools.

Kitchenham and Charters [20] group the tasks of an SLR into three phases, which are accompanied by a review protocol. This protocol defines the procedure and ensures objectivity and repeatability of the survey. While the remainder of this section describes the *planning phase*, the results obtained in the *conduction phase* are discussed in Section III. The third phase, the *reporting phase*, is not discussed, as it deals with the general topic of disseminating an SLR.

The planning phase comprises all tasks that need to be carried out before the survey can be conducted [20]. Section II-A briefly reviews the value and contribution of this survey, while Section II-B defines the research questions. Section II-C describes the search strategy, and Section II-D presents the criteria used to select and analyze the search results.

A. THE MOTIVATION FOR THIS SURVEY

At the time the planning phase was started, there was no peer-reviewed survey of data catalog tools available. In the meantime, Jahnke and Otto [3] published a survey of data catalogs, focusing on how well these tools can be integrated into enterprise-wide metadata management landscapes (see Section IV-C1 for a detailed discussion). To the best of our knowledge, this is the first survey of data catalog tools that (i) investigated the internal metadata management capabilities of data catalogs, and (ii) provided a comprehensive overview of the data catalog tool landscape.

B. RESEARCH QUESTIONS

This survey analyzed the main functionalities (collecting and storing the metadata of data sources and annotating it with domain knowledge, cf. Section I) of existing data catalog tools. Thus, the primary research questions guiding this paper were:

(R1) Which general-purpose data catalog tools are available?

(R2) What is the nature of their support for the main functionalities (and other functionalities)?

The first research question is addressed in Section III-B, which details the exclusion process applied to all tools retrieved in the systematic search. Six different aspects of

state-of-the-art data catalog tools, as listed in Section III-C, were investigated to answer the second research question.

C. SEARCH STRATEGY

For the systematic search, the commonly used online platforms listed in Table 1 were queried. To address both researchers and practitioners, three different platform categories (literature, source code, and general search) were used. The following search expression was used:

$$(\text{“data catalog”} \vee \text{“data catalogue”} \vee \text{“metadata management”}) \wedge (\text{“tool”} \vee \text{“software”} \vee \text{“system”})$$

In addition to the British English and American English spellings of the term “data catalog”, the search expression also included the term “metadata management”, since practitioners often use it as a synonym for data cataloging (cf. [6], [21]). Korte et al. [9] additionally used “data governance” and “data lake management” in their search-based market analysis, but neither term was considered relevant to the research questions of this paper. Data governance clearly includes strategies of involving people in metadata management [22], while we focus on tool functionality, and data lake management also covers the management of data instances, while our focus is on the structure of metadata. Another term sometimes confused with data catalogs is “data dictionary”, which contains technical metadata, such as structural information about data objects and data types [2], [9]. This term was not used in the search expression because it has a narrower scope than “data catalog”, which is additionally associated with business metadata, data responsibility roles, data lineage, and sometimes even data quality assessment [8], [9].

Where possible, the search scope was limited to title, abstract, and keywords. The search was further limited to results with a publication date of 2005 or later, since this was the time when data catalogs were first mentioned in [10]. Figure 1 visualizes the increasing use of “data catalog” or “data catalogue” in titles, abstracts, and keywords since that point in time.

D. PROCESSING THE SEARCH RESULTS

The pipeline we developed (see Figure 2) consisted of (i) processing of the search results to identify data catalog tools (see II-D1), (ii) feeding these tools to a filtering process using exclusion criteria (see II-D2), and (iii) analyzing the remaining tools using the criteria outlined in II-D3.

1) PROCESSING ALL SEARCH RESULTS

Each search result was checked sequentially using a set of criteria, as illustrated in Figure 2. If any exclusion criterion was met, the result was discarded.

Initially, a search result was checked for its availability in English, as the search query was also in English. Second, a distinction regarding the type of result was made: If a search result was a survey, it was evaluated against the criteria that define a primary survey. Kitchenham, Madeyski, and Budgen [24] distinguish between four types of information sources: Formally published literature (white literature), informally published literature (gray literature [25]), self-published information (e.g., social media posts, blog posts, and wikis), and unpublished information (e.g., e-mails, notes). Following the suggestions of [24], only the first two categories (white and gray literature) were considered primary surveys.

To identify as many tools as possible, each result (regardless of publication type) was checked for information about or mentions of data catalog tools. Results that contained or referred to a data catalog tool were selected for further investigation, while those unrelated to data catalog tools were discarded.

2) EXCLUSION OF TOOLS

In the second step, the data catalog tools found were investigated further to select only those relevant to the first research question. Hence, four exclusion criteria (abbreviated as E1-4) were applied to the list of tools in the order described below. If an exclusion criterion was met, the corresponding tool was discarded.

- **(E1)** No support or information available: Since only current tools were of interest, all tools that had been discontinued by their vendor (whose source code repository had been archived or had had no commits since 2018) were discarded. Each tool had to have further information available (e.g., documentation, usage information) beyond the mention by which it was originally identified.
- **(E2)** Main purpose is not data cataloging: Tools with other but often similar purposes, such as data integration or data management tools, were excluded.
- **(E3)** Domain-specificity: Since this work focused on analyzing the main functionalities of data catalog tools, only general-purpose tools were considered and highly domain-specific tools (e.g., a data catalog for cultural heritage) were excluded from further analysis.
- **(E4)** Storage of data: The literature agrees (cf. [11], [12]) that data catalogs have a repository that contains the metadata of data sources, but not the actual data itself. As a result, tools that focused primarily on storing or integrating data were discarded, as this is not the main objective of data catalogs.

Detailed statistics on the tools excluded are provided in Section III-C. All remaining data catalog tools were taken forward for further analysis.

3) ANALYSIS OF DATA CATALOG TOOLS

The remaining tools were analyzed based on six key aspects (abbreviated as A1-6) to provide a comprehensive overview of the features of and emerging trends in data catalog tools. These six aspects were selected to offer insights that are

TABLE 1. Exact search expression per platform.

Platform	Search expression	Notes
ACM Digital Library (https://dl.acm.org/)	Title:((“data catalog” OR “data catalogue” OR “metadata management”) AND (“tool” OR “software” OR “system”)) OR Abstract:((“data catalog” OR “data catalogue” OR “metadata management”) AND (“tool” OR “software” OR “system”))	
GitHub (https://github.com/)	Multiple individual searches were conducted: • “data catalog” “tool” • “data catalogue” “tool” • “metadata management” “tool” • “data catalog” “software” • “data catalogue” “software” • “metadata management” “software” • “data catalog” “system” • “data catalogue” “system” • “metadata management” “system”	According to GitHub’s documentation [23], connecting search terms with “AND” and “OR” is not supported.
GitLab (https://gitlab.com/)	(“data catalog” | “data catalogue” | “metadata management”) + (“tool” | “software” | “system”)	Searching across GitLab requires a free account.
Google (https://google.com/)	((“data catalog” OR “data catalogue” OR “Metadata Management”) AND (“tool” OR “Software” OR “system”))	
Google Scholar (https://scholar.google.com/)	allintitle: ((“data catalog” OR “data catalogue” OR “Metadata Management”) AND (“tool” OR “Software” OR “system”))	
IEEE Xplore (https://ieeexplore.ieee.org/)	((“All Metadata”: “data catalog” OR “All Metadata”: “data catalogue” OR “All Metadata”: “Metadata Management”) AND (“All Metadata”: “tool” OR “All Metadata”: “Software” OR “All Metadata”: “system”))	
ResearchGate (https://www.researchgate.net/)	(“data catalog” OR “data catalogue” OR “metadata management”) AND (“tool” OR “software” OR “system”)	ResearchGate was queried without being logged into a ResearchGate account, as this ensures a finite number of 100 results. For a logged in user, ResearchGate always finds an unlimited number of results, irrespective of the query entered. Filtering the results with respect to their publication date had to be done manually.
ScienceDirect (https://www.sciencedirect.com/)	(“data catalog” OR “data catalogue” OR “metadata management”) AND (“tool” OR “software” OR “system”)	
Springer Link (https://link.springer.com/)	“data catalog tool” OR “data catalogue tool” OR “metadata management tool” OR “data catalog software” OR “data catalogue software” OR “metadata management software” OR “data catalog system” OR “data catalogue system” OR “metadata management system”	Computer Science was selected as “discipline”, and any preview-only content was excluded.

relevant to both academic research and practical use cases. Aspects (A1) and (A2) – the type of domain knowledge representation and the level of detail of metadata, respectively – were based on previous studies of the components of data catalogs [8] and have also been recognized as important aspects in other publications [2], [13]. The features (A3) available connectors, (A5) licensing options, and (A6) deployment types were chosen with a focus on the practical and technical requirements to support selection of a suitable data catalog tool. Finally, the extent to which tools integrate artificial intelligence (A4) reflects an important new trend highlighted in the literature and underscores the growing relevance of AI in modern data systems. This approach ensured that the analysis captured important functional, technical, practical and forward-looking aspects to provide a comprehensive perspective on data catalog tools. A separate and detailed justification for the selection of each analysis criterion is provided below.

(A1) Type of domain knowledge representation:
Data catalogs are aimed not only at information technology (IT) specialists, but also at business users [2], [8]. An effective representation of domain knowledge improves the usability of data catalogs by making it easier for non-technical users to interact with data. This facilitates collaboration between IT and business teams and also ensures that the data is used in a more

meaningful context [8]. In this paper, the annotation of technical metadata with domain knowledge is seen as a main functionality of data catalogs, which is why we analyzed what types of domain knowledge representation the tools used. The literature mentions using additional attributes for tagging the cataloged data assets [13] and creating a business glossary [2] (a repository of agreed business terms [26]) as ways of adding domain knowledge. A further, more expressive alternative is to use ontologies [14], [15].

(A2) Granularity of the metadata about data sources: In addition to representing domain knowledge, collecting and storing metadata is considered a main functionality. Metadata collected by a data catalog can cover various levels of granularity of a data source. This includes, for instance, descriptions of entire datasets, table names, and information about attributes. Metadata at various levels of granularity is essential for specific functionalities that are based on cataloging [13]. Fine-grained metadata enables effective data discovery, fine-grained annotation with domain knowledge, and data lineage functionality and supports data governance and data quality assurance efforts. The data catalog tools were therefore examined for attribute-level data granularity support.

(A3) Availability of connectors: The availability of connectors directly affects the usability and flexibility of a tool. In this paper, a distinction is made between data source connectors and application connectors: The former allow a data catalog tool to connect directly to databases, file systems, or other data repositories to load and/or catalog data (also called ''ingestion''). The latter allow a tool to integrate with external software applications, such as analytics platforms, to perform downstream tasks on the data held by the data catalog or its sources. This survey thus captured the number and types of connectors offered by each data catalog tool (i.e., whether it offers solely data source connectors or also application connectors).

(A4) Artificial intelligence: Artificial intelligence (AI), and in particular machine learning (ML), has gained prominence within the domain of data catalogs and has the potential to enhance data management activities [9], [18]. AI can improve automation, reduce time and human effort required, and opens up new opportunities, such as automated data discovery and business metadata generation. The incorporation of AI into data catalogs is a clearly recognizable trend [9], [18]. Hence, this work examined the extent to which state-of-the-art data catalog tools utilize AI, and in particular ML, identifying the components in which these technologies are most commonly implemented.

(A5) Licensing and extensions: Knowledge of whether all cataloging functionalities are included out of the box is key when choosing a data catalog tool for a particular application. The licensing model is particularly important for practitioners, as it has a direct impact on cost, flexibility, and scalability. Understanding the licensing model helps organizations to align their choice of tool with their budget and long-term requirements. Therefore, the licensing models and possible extensions for adding further functionalities to the cataloging tools were analyzed.

(A6) Types of deployment: Data catalogs differ in their deployment type – on-premise or cloud-based – and thus in the advantages they offer and in the considerations leading to their application. For enterprises it is crucial that they can deploy a tool both technically and organizationally (in compliance with regulations) within their existing infrastructures. By analyzing the deployment options the various data catalog tools offer, this paper aims to enable potential users to make an informed decision based on factors such as scalability and cost effectiveness.

III. RESULTS

This section covers the conduction phase, in which the SLR tasks (cf. Section II) were executed. Section III-A describes the retrieval of data catalog resources from which tools were taken. Here, the term ''resource'' refers to any type of result, including journal articles, conference proceedings, product websites, and product comparison websites. Based on the resources retrieved, Section III-B describes how unrelated tools were excluded and Section III-C how the remaining tools were analyzed. Related surveys extracted from these resources are discussed in Section IV-C3 within the section on state-of-the-art data catalog research.

A. RESOURCES IDENTIFIED

Between April and June 2023, 947 distinct resources were identified by applying the search expressions to the selected online platforms as defined in Table 1.

Table 2 shows statistics of the resources retrieved using the search process shown in Figure 2. Of the 947 resources, 19 (≈2 %) were classified as surveys or lists of data catalog tools, and only five of these, discussed in Section IV-C3, were primary surveys (cf. Section II). Of the 947 publications, 255 (≈27 %) mentioned one or more tool(s). After removing duplicates, 247 individual tools were considered for further investigation.

B. SELECTION OF DATA CATALOG TOOLS

The total number of tools was reduced by sequential application of four exclusion criteria (E1-E4) as defined in Section II-D1 to select only those suitable for further analysis. Figure 3 provides an overview of the number of tools excluded per criterion. In the following subsections, each criterion is discussed in detail.

E1) NO SUPPORT OR INFORMATION AVAILABLE

In total, 85 tools were excluded by the first criterion, 55 of which were no longer actively supported by their

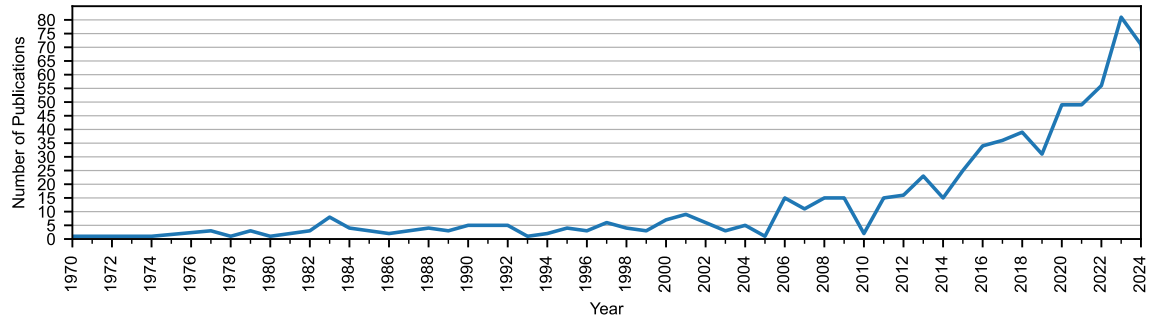


FIGURE 1. Annual number of publications that contain at least one of the terms "data catalog" or "data catalogue" in title, abstract, or keywords according to Scopus (https://www.scopus.com/). Date of query: January 13th, 2025.

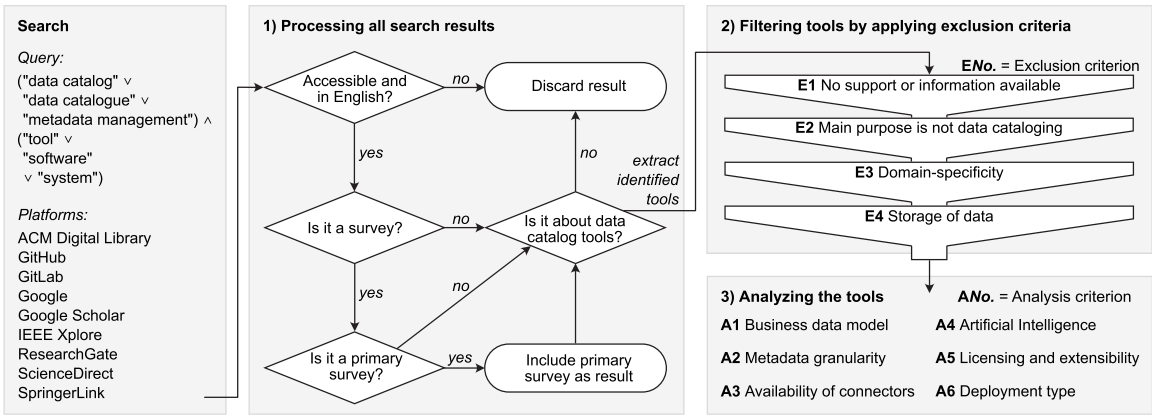


FIGURE 2. The search, exclusion and analysis processes of the survey detailed in Section II.

TABLE 2. Per-platform and total results of processing all resources retrieved.

Platform	No. of resources	No. of surveys	No. of primary surveys	No. of publications on data catalogs
ACM Digital Library	61	0	0	10
GitHub	97	1	0	24
GitLab	1	0	0	0
Google	100	13	0	84
Google Scholar	60	0	0	28
IEEE Xplore	317	2	2	50
ResearchGate	95	1	1	37
ScienceDirect	72	0	0	4
Springer Link	144	2	2	18
Total	947	19	5	255

creators. This included source code repositories that had been archived, tools that had not been updated or changed in the previous seven years (i.e., since 2018), and tools that were explicitly called deprecated by their vendors. An additional 30 tools were excluded because no further information was available beyond the resources found in the search.

E2) MAIN PURPOSE IS NOT DATA CATALOGING

As a second exclusion criterion, the purpose of the individual data catalog tools was examined. 60 of the tools found did not correspond to the interpretation of data catalogs used in this paper (cf. Section I) and were therefore excluded. The distribution of the excluded tools across domains is shown in Figure 4.

E3) DOMAIN-SPECIFICITY

The remaining 102 tools were data catalogs, but only 80 of these were classified as general-purpose (i.e., domain-independent), while 22 were tailored to a specific application domain. The 22 domain-specific data catalog tools were associated with 5 application domains (cf. Figure 5): Cross-domain science and research (9 tools), biological and medical (5 tools), geospatial and environmental (4 tools), cultural heritage (3 tools), and education management (1 tool).

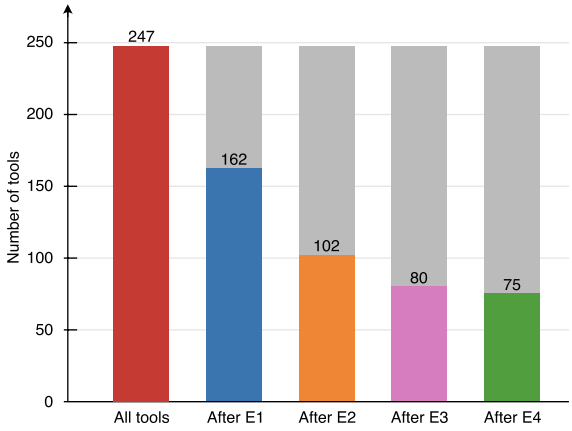


FIGURE 3. Number of tools remaining after applying each exclusion criterion.

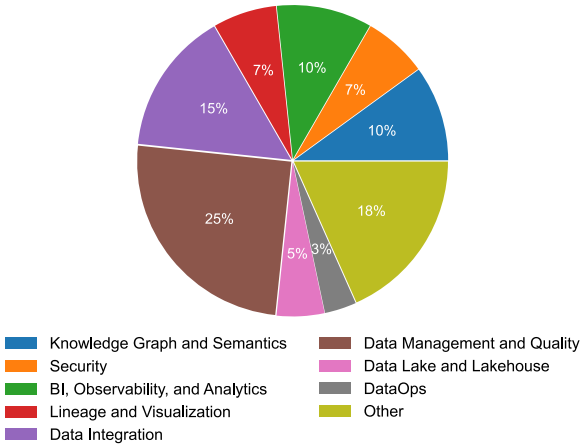


FIGURE 4. Classification of tools not considered data catalogs.

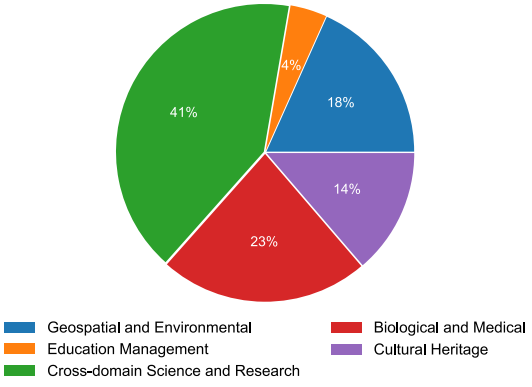


FIGURE 5. Classification of the domains of domain-specific data catalog tools.

E4) STORAGE OF DATA

The fourth exclusion criterion examined whether a data catalog tool stores only metadata and references to data or whether it also stores the data itself. The latter was the case for five of the tools and consequently led to their exclusion. A total of 75 tools (see Appendix) were selected for further analysis, including both standalone data catalog tools and larger ecosystems that included a data catalog.

C. RESULTS OF THE TOOL ANALYSIS

The 75 remaining data catalog tools were analyzed with regard to six dimensions (aspects A1-A6), as described in Section II-D3. The analysis, the results of which are summarized below, was based on information from each tool's corresponding website, including general information about the tool, technical reports, available documentation, and videos demonstrating the functionalities and features. While the complete analysis results are accessible through the supplementary material [19], the list of all 75 tools with links to their websites is provided in the Appendix.

A1) TYPE OF DOMAIN KNOWLEDGE REPRESENTATION

Of the 75 tools, two (≈3 %) do not allow domain knowledge to be represented, as illustrated in Figure 6. Since a single tool can provide multiple ways of representing domain knowledge, the sum of the numbers of tools for each concept in Figure 6 exceeds the number of tools. Tags are the most popular concept, being used by 52 tools (≈69 %), followed by business glossaries, which are used by 48 tools (64 %). Only three tools (4 %) rely on graphs or ontologies, although these offer greater expressiveness (cf. [27]). Seven tools (≈9 %) use other concepts to represent domain knowledge that do not correspond to any other category, for instance, a list of KPIs (Key Performance Indicators) or taxonomies with limited semantic expressiveness.

A2) GRANULARITY OF THE METADATA OF DATA SOURCES

Of the remaining data catalog tools, 69 (92 %) provide metadata about data sources at the level of attributes or columns. The tool CKAN does not provide this functionality out of the box, but it can easily be extended. This analysis criterion is not applicable to Data X-Ray, as the tool is intended for unstructured data sources. For four data catalog tools (≈5 %), it could not be determined whether they support metadata at the attribute level because no corresponding information was available.

A3) AVAILABILITY OF CONNECTORS

Connectors are an essential part of a data catalog, as they enable integration of the tool with external systems. As shown

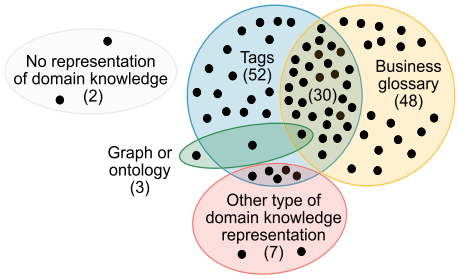


FIGURE 6. Distribution across concepts available for representing domain knowledge.

in Figure 7, eleven of the 75 tools (≈15 %) did not provide enough information to allow concrete conclusions to be drawn, neither on the number of available connectors nor on the type of connectors supported. Three tools (4 %) do not include any connectors. Of the remaining 61 tools (≈81 %), eight (≈10 %) are tied to a specific ecosystem, such as Azure, Snowflake, and Apache Hadoop. 18 tools (24 %) offer up to 49 different connectors, while 15 (20 %) offer between 50 and 100 connectors. Within this category, two tools offer between 50 and 100 connectors and an optional extension at additional cost. 20 (≈27 %) tools offer 100 or more connectors.

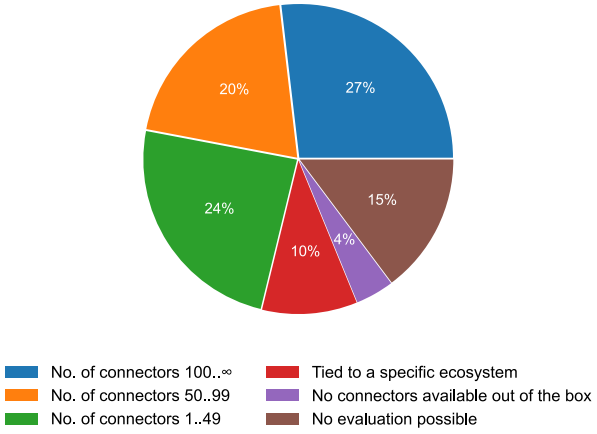


FIGURE 7. Availability of different connectors.

An important goal of a data catalog tool is to maintain a data repository to make it easier for users to find and manage data in an organization. Data catalog tools must provide various data source connectors or application programming interfaces (APIs) to catalog data sources and to ingest metadata into the system [16]. In this paper, data source connectors are therefore considered to be the most important connector type of a data catalog tool. In total, 61 of the 75 tools (≈81 %) claim to provide data source connectors, such as connectors to popular relational databases (e.g., MySQL [28] or PostgreSQL [29]), cloud storage systems (e.g., AWS S3 [30] or Google Cloud Storage [31]), and NoSQL databases (e.g., MongoDB [32] or Cassandra [33]).

In addition to data source connectors, many tools offer connectors to interface with various other applications and platform types, including business intelligence tools, data visualization platforms, and collaboration platforms. The analysis revealed that 42 of the 75 tools (56 %) offer connectors for other application types. Four tools (≈5 %) do not include such connectors per default, but provide options to add connectors either for free or for an additional charge. 16 tools (≈21 %) do not support application type connectors. These are tools that either generally do not provide any connectors (3; 4 %), are tied to a specific ecosystem (8; ≈10 %), or offer more connectors only for data sources (5; 7 %). 13 tools (18 %) do not provide enough information for an evaluation. This includes the 11 tools that already do not provide sufficient information on connector support in general and two additional tools that do not clearly disclose whether they support other application types beyond data source connectors.

Figure 8 illustrates the categorization of tools based on the number of connectors they support, specifically data source connectors and application connectors. Note that the categories "Tied to a specific ecosystem", "No connectors available out of the box", and "No evaluation possible" shown in Figure 7 are not included in Figure 8, since there is no distinction between data source and application connectors for these categories.

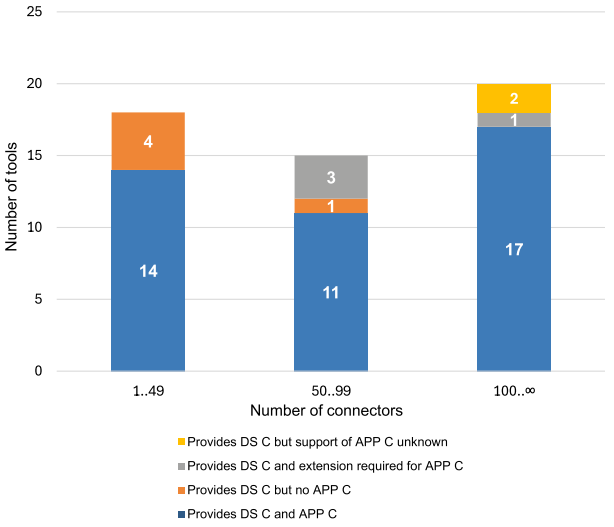


FIGURE 8. Tools categorized by the number of connectors available and their support for data source connectors (DS C) and / or application connectors (APP C); APIs can serve as both DS C and APP C, depending on the implementation of the tool, and are not explicitly distinguished.

A more detailed analysis of the data source connectors and other third-party application connectors available can be performed with the evaluation table in the supplementary material [19].

A4) ARTIFICIAL INTELLIGENCE

About half of the tools analyzed (37; ≈49 %) explicitly claim to offer AI functionalities. Figure 9 clusters these tools according to their functionalities. Since one tool may use AI, and in particular ML, for multiple functionalities, the sum of the functionalities listed below exceeds the number of tools.

The functionality groups most commonly supported by ML or, more generally, AI are generating business metadata (e.g., creating business glossaries, determining business names, generating documentation), supporting the data discovery process (e.g., data scanning, data ingestion), and enhancing the cataloging process (e.g., assigning tags). Less common AI-based functionalities include assisting users with suggestions, analysis of data sources, and data lineage. Only one tool features an ML-powered data governance component. Four tools mention the use of AI, but do not provide further details.

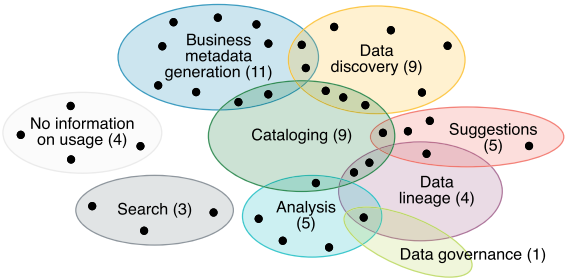


FIGURE 9. Clustering of AI-based functionalities in data catalog tools.

A5) LICENSING AND EXTENSIONS

Since the number of licensing variants (and possible tool extensions) is large, they are grouped into categories for better interpretability of the results. If multiple licensing variants are offered, tools may, as with the other aspects, belong to multiple categories.

- All-in-one product:
  50 of 75 tools fall within the category of feature-complete products that are either available for free, on subscription, or as a one-time purchase.
- Feature-based licensing:
  15 tools offer flexible functionality plans or extensions at additional cost.
- Part of a larger system:
  Two tool vendors state that they offer their tools as part of a larger system. One of these is also available as an ''all-in-one product''.
- No explicit information available:
  For eight tools (≈11 %), no explicit information about licensing and tool extensions was publicly available.

While most tools (63; 84 %) are proprietary, 12 are open source (16 %) and – with the exception of CKAN – ''all-in-one products''. Since CKAN offers a limited number of functionalities, which can be expanded using plug-ins, it was not assigned to any of the above categories, as we do not consider it to be feature-complete out of the box.

A6) TYPES OF DEPLOYMENT

Criterion A6 examined whether a data catalog tool was available for on-premise installation, as software as a service (SaaS) (see [34] for a definition), or both. Some tools that offer a ''hybrid'' deployment type were included in the SaaS category rather than assigned to a separate category because the tool itself is typically provided as SaaS and connects to connectors installed on-premise to ingest data into the SaaS-based catalog. Note that even if a tool is categorized as on-premise, it can run in a private cloud, for instance, on virtual machines or by using containers.

As shown in Figure 10, 20 (≈27 %) data catalog tools are available for on-premise installation only, 25 (≈33 %) only as SaaS, and 22 (≈29 %) as both. For eight (≈11 %) tools, no information regarding deployment was publicly available.

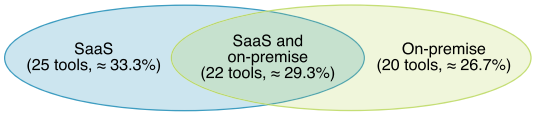


FIGURE 10. Available deployment options.

IV. STATE OF THE ART IN DATA CATALOG RESEARCH

Data catalogs were first mentioned by Franklin et al. [10] in 2005 in the context of the ''dataspace'' concept. At that time, the focus was mainly on the cataloging functionality. Later, other functionalities were also considered to be part of a data catalog, including data pipeline management, data lineage, and data governance functionalities [9]. Despite the growing popularity of data catalogs in the enterprise context [5], [16], the literature on this subject is limited [2], [8]. This section begins with a historical review of cataloging, then examines definitions of data catalogs, and finally presents an overview of related surveys.

A. HISTORICAL BACKGROUND OF CATALOGING

The history of cataloging dates back to ancient times [35], [36], [37], when libraries, such as the library of Alexandria, already had catalogs on clay tables to organize their collections [36]. Especially for large libraries, cataloging is essential, as it enables efficient organization and retrieval of materials and improves findability. Over time, various rules have been established to standardize the cataloging process and have been adapted to accommodate various types of artifact, such as recordings or photographs [35], [37]. By assigning standardized metadata and indexing items, cataloging ensures that resources can be easily found using metadata such as subject, author, and title.

In the course of technological progress, libraries adopted online catalogs [35], [37]. Online library catalogs provide users with the ability to search for materials electronically using keywords, author names, or subject terms. The search results displayed are relevant items along with availability status, locations, and – if applicable – even digital content. The shift to digital cataloging has not only made searching for materials more efficient but has also enabled libraries to manage larger collections.

B. DEFINITIONS AND GENERAL SURVEYS OF DATA CATALOGS

Despite several attempts to define the concept (cf. [5], [9], [10], [11]) there is no generally accepted definition of a ''data catalog''. According to Zaidi et al. [5], data catalogs document metadata and provide context to the data stored across an organization. Data catalogs have been considered as part of larger systems [5], [10]. In contrast, Korte et al. [9] considered a data catalog to be a large system that provides a variety of services, including data governance and data assessment. Korte et al. [9] also contributed a reference model for data catalogs that consists of three interdependent

components: (i) the functional model describes the main functionalities a data catalog should offer, such as data inventory, data governance, and discovery; (ii) the role model describes user groups and use cases; and (iii) the information model (subject of future work) describes how data assets are to be documented [9]. Hilger and Wahl [11] argued that data catalogs should offer a variety of services, including data access rules or features for data discovery. In their view, a data catalog provides a centralized way of managing metadata, which supports better data governance by allowing discovery, understanding, and access to data [11].

Quimbert et al. [12] did not provide a new definition, but focused on a detailed description of the functionality by stating that ''data catalogues exist to collect, create, and maintain metadata'' [12], which allows easier findability and accessibility.

Ehrlinger et al. [8] addressed the lack of research into the conceptual components of data catalogs and identified four main components in their systematic literature review: (i) metadata management, which plays a central role in data cataloging; (ii) business context, which is important for enhancing technical metadata with domain knowledge; (iii) data responsibility roles, such as data stewards; and the (iv) FAIR principles [38], which define guidelines for publishing (meta)data. Note that the FAIR principles are not considered to be technical components of data catalogs, but as a benefit that results from their use [8].

Labadie et al. [2] investigated the extent to which data catalogs support compliance with FAIR principles in organizations. The authors developed a taxonomy that classifies data catalog initiatives in terms of scope, goals, and users. To demonstrate a variety of goals and approaches to the implementation of a data catalog, Labadie et al. [2] presented three case studies: A top-down approach that focuses on data supply; a balanced, step-by-step approach that combines top-down elements with bottom-up elements; and a user-driven bottom-up approach that focuses on the needs of end users. According to [2], the main difficulty in implementing data catalogs is conveying the value of using a data catalog to users. In addition, maintaining detailed documentation and data quality while managing complexity is challenging, but can be optimized by using automation technologies. The study [2] highlighted the differences in the application of the FAIR principles in science versus business: In the latter case, it is more difficult to motivate users to actually use the data, even if the data is technically FAIR [2].

Boufassil et al. [18] presented methodologies, trends, and future prospects for data catalogs. They evaluated various architectural models, including metadata-based, data-based, and hybrid approaches, discussing the advantages and disadvantages of each approach to data catalogs. Contrary to the interpretation of a data catalog used in this paper, the data-based approach they presented involves storing the actual data. The authors emphasized emerging trends, such as integration of AI to enhance data discovery and the shift towards cloud-based architectures for better scalability and flexibility. A use case in a healthcare organization illustrated the benefits of a data catalog. The authors also mentioned disadvantages of integrating a data catalog, such as high cost of implementation and maintenance, complexity of use, security risks due to the sensitive nature of the data, and potential over-reliance on certain tools. Despite these challenges, Boufassil et al. [18] suggested that advances in data catalog development, including interoperability, advanced data analytics, and improved security measures, would continue to improve data management practices.

C. RELATED TOOL SURVEYS

This section first discussed Jahnke and Otto's survey [3] as the most closely related peer-reviewed publication (see Section IV-C1) and then explored non-peer-reviewed data catalog tool surveys in Section IV-C3. Finally, it provides an overview of other related surveys identified in the systematic search.

1) PEER-REVIEWED DATA CATALOG TOOL SURVEYS

To the best of our knowledge, there is only one peer-reviewed survey of data catalog tools [3], which was published after our search process was complete and, therefore, was not included in the search results. In a systematic search, Jahnke and Otto [3] identified 73 data catalog tools, 51 of which they analyzed in more detail; this is comparable to the 75 tools analyzed in this paper. Although there is an overlap of 40 vendors (36 distinct tools) with our survey, they obtained different results because they [3] (i) restricted their list to one tool per vendor (while we included multiple tools, e.g., by Oracle), (ii) also included tools that were marked as deprecated when we performed our search, and (iii) included tools that we classified as larger projects rather than data catalogs (e.g., Gaia-X).

For the tool analysis, information from the vendors' websites was used, including available documentation and video tutorials. One main contribution of [3] is a typology of seven data catalog classes that helps practitioners to identify the right tool for integration into their (existing) enterprise metadata management landscape. This integration can be challenging due to heterogeneity of the data catalog classes [3].

Jahnke and Otto [3] used – similarly to our analysis – five different functionality types to investigate the data catalog tools identified: ''Deployment types'', ''connectors and integrations'', ''federation'', ''data access'', and ''metadata management landscape''. While two of their analysis criteria match those presented here (i.e., ''deployment types'' and ''(A6) deployment type'' and ''connectors and integrations'' and ''(A3) availability of connectors''), ''federation'' and ''metadata management landscape'' were tailored specifically to the focus of [3] on enterprise-wide metadata integration. Compared to [3], we focused on the (internal) metadata management capabilities of data catalogs. Thus, the analysis presented in this paper additionally investigates (A1)

how data catalogs represent domain knowledge internally, (A2) the granularity of the metadata stored, (A4) the adoption of AI functionalities, and (A5) licensing and options for extensions. In addition, we enumerate domains for application-specific data catalogs (cf. Figure 5) and provide an overview of tool classes that fall outside our interpretation of data catalogs, but were discovered through our systematic search (cf. Figure 4).

In conclusion, the two surveys complement each other in terms of tool analysis. While Jahnke and Otto [3] aimed to address the important aspect of enterprise integration of data catalogs, our survey focused on (i) the internal ability of data catalog tools to model and store metadata and (ii) the overall landscape of data catalog tools. The first aspect of (i) metadata management is particularly important to enable the reuse of this metadata for various other tools and tasks within an organization, such as data quality tools that require metadata to have a sufficient level of detail for annotating constraints. An additional important contribution of this study is a tabular presentation of the detailed analysis results. The table provided online can be customized, sorted, and filtered, enabling further analysis, which should help practitioners, in particular, to select suitable tools for their requirements [19].

2) NON-PEER-REVIEWED DATA CATALOG TOOL SURVEYS

A variety of non-peer-reviewed documents are available that provide an overview of data catalog tools. Korte et al. [9] conducted a market analysis of 15 commercial data catalog tools, primarily based on vendor material. Each tool was described and evaluated according to the criteria defined in the reference model the authors proposed [9]. Their evaluation was based on the functionality groups ''data inventory'', ''data collaboration'', ''data assessment'', ''data governance'', ''data discovery'', and ''data analytics''. In addition, Korte et al. [9] included ''automation and machine learning'' and ''data visualization'', which were not addressed by the original reference model, but were added later based on feedback from practitioners. Each tool was evaluated using a five-point scale ranging from ''not addressed'' to ''fully addressed''. Korte et al. [9] concluded that data catalog tools can be divided into two categories: (i) those that focus on data collaboration and data governance, and (ii) those that mainly address the management of data lakes.

In their report ''Magic Quadrant for Metadata Management Solutions'' Gartner Inc. [6] listed strengths and risks of metadata management tools (i.e., data catalog tools according to the interpretation used in this paper). Another report by Gartner Inc. [5] provided a list of tools with data cataloging capabilities but did not evaluate them. De Simoni et al. [6] compiled a representative list of the leading commercial tools with data catalog functionalities, including only those that offered at least the following functions: Metadata repositories, a business glossary, data lineage, impact analysis, rule management, and metadata ingestion and translation from various sources. Their report, however, also lacked a detailed comparison of the tools listed. De Simoni et al. [6] concluded that organizations should carefully evaluate metadata management solutions based on their requirements and should specifically consider the aspects of data governance and collaboration when selecting a tool.

3) OTHER RELATED SURVEYS

The search (cf. Section III-A) yielded five relevant surveys and lists ([39], [40], [41], [42], [43]) – four from white and one from gray literature (cf. Section II). Notably, none of the four white-literature surveys focused explicitly on the topic of data catalogs. However, since these surveys addressed data asset management to some extent, they were considered relevant for our survey. As ''data lake'' is considered a buzzword without a commonly agreed definition (cf. [39]), Hai et al. [39] conducted a peer-reviewed survey of tools that are marketed as data lakes or as having some data lake functionality. The data lake tools were compared in terms of data storage, ingestion, maintenance, and exploration. Metadata management and extraction were also considered important in the data lake context because proper use of metadata is key to using data in the lake effectively [39]. Hai et al. [39] also mentioned the FAIR principles (cf. [38]) as an important aspect.

Macedo et al. [40] examined four different open data publication and management tools for the Comprehensive Knowledge Archive Network (CKAN). Although CKAN is – according to its website – an ''open source data portal software'', it is often classified as a data catalog. In summary, Macedo et al. [40] focused specifically on the domain of smart cities and considered only open source tools compatible with CKAN.

Santos et al. [43] and Oliveira et al. [42] provided two related systematic mapping studies on the topics of data published on the web and data ecosystems. Santos et al. [43] focused on understanding data published and consumed on the web, exploring various aspects, such as data sources, formats, and usage patterns, with the objective of identifying research trends and gaps. Oliveira et al. [42] examined topics such as data governance, integration, sharing, and analytics within the domain of data ecosystems. Both studies provided valuable insights into the challenges faced in understanding and managing data in the digital era, and both mentioned CKAN as a commonly used platform for publishing data on the web. However, in alignment with [40], [42], and [43], we argue that CKAN is neither suitable as a platform for creating a data ecosystem [42] nor sufficient to monitor data consumption [43].

McSweeney's data catalog survey [41] is the only one classified as gray literature. The survey discussed the concepts of data profiling, data catalogs, the FAIR principles, (meta)data management, data integration, and data governance, and showed how these concepts relate to each other. McSweeney [41] described a data catalog as a registry of data

sources that provides links to storage location of the data, contains descriptions of the stored data, and holds metadata in a structured format that can be queried. Data catalogs can thus capture relationships between the data sources. Business glossaries and data dictionaries are mentioned as related concepts, and the Data Catalog Vocabulary (DCAT) as a standard for describing datasets within a data catalog. Further, the importance of the FAIR principles (cf. [38]) was highlighted [41]. McSweeney [41] also mentioned several well-known platforms, such as Zenodo [44], Invenio [45], and Dataverse [46] as data catalogs.

The results of the search process in the present paper show that there is a significant lack of peer-reviewed literature on data catalogs. For instance, McSweeney's survey [41] is neither peer-reviewed nor systematic. The peer-reviewed surveys identified cover only related topics (e.g., data lakes), but do not provide a complete picture of data catalog solutions and their functionalities.

V. DISCUSSION

This section discusses the state of the art in data catalog tools from research and practical perspectives. First, Section V-A discusses terms that the survey revealed to be ambiguous. Subsections V-B and V-C focus respectively on the two target groups of this paper, scientists and practitioners. A discussion of threats to validity in Section V-D concludes this section.

A. DISCUSSION OF CONCEPTS AND TERMS

The systematic search revealed seamless transitions between various concepts, and ambiguously defined terms.

1) METADATA MANAGEMENT

While originally intended to broaden the search results, it became apparent that the term ''metadata management'' not only refers to data catalogs and data management [6], [8], but is also used in other domains. The term features prominently in the contexts of data lakes [47], multimedia content [48], distributed file systems, and sensor management. The latter two are described below.

Metadata management has often been mentioned in the context of ''distributed file systems'' (e.g., [49], [50], [51], [52]). A distributed file system is a network-based storage infrastructure that facilitates efficient and secure file sharing among multiple users across autonomous computers [53]. In this context, metadata management plays a key role in monitoring the organization, access, and manipulation of file system metadata [54]. Metadata in a file system includes information such as file attributes, permissions, and file location [52], [54]. In particular, operations on metadata can constitute a significant proportion – up to 80 % – of total file system operations, which highlights the importance of metadata to the functionality of such a system [54].

Another field in which metadata management is essential is sensor networks and sensor management. Metadata provides context about sensor data, including time stamps, location, and sensor specifications [55]. Dawes et al. [55] highlighted the importance of metadata and metadata management in the context of sensor networks, where it facilitates the identification and resolution of issues. In this field, metadata aids in tasks such as pinpointing faulty sensors for repair, even when specific location information is unavailable, by leveraging details such as serial numbers and database parameters [55].

2) DATA CATALOG

The main challenges in the systematic search were to determine whether (i) a particular resource pertained to data catalogs and (ii) a tool can be classified as a general-purpose data catalog tool.

Some resources identified by the search initially appeared relevant, but upon closer examination turned out to interpret the term ''data catalog'' differently. In many cases, such resources referred to data portals or similar concepts. An example of a tool at the intersection of data repositories and data catalogs is CKAN. Although CKAN provides metadata management and discovery, its primary focus is on data publishing for open data initiatives. However, this survey still categorizes CKAN as a general-purpose data catalog tool, even though governments are using CKAN to promote transparency rather than as a traditional catalog [40]. This classification is justified because CKAN, despite its strong emphasis on open data sharing, fulfills the main functionalities expected from data catalogs and is in line with Jahnke and Otto [3], who classified data portals as a type of data catalog.

As this example outlines, the selection of a publication or tool for further consideration was influenced by our interpretation of a data catalog (cf. Section I). That 75 data catalog tools matched our interpretation emphasizes the variety of tools on the market, which poses a challenge for practitioners seeking to select the most suitable tool for their requirements.

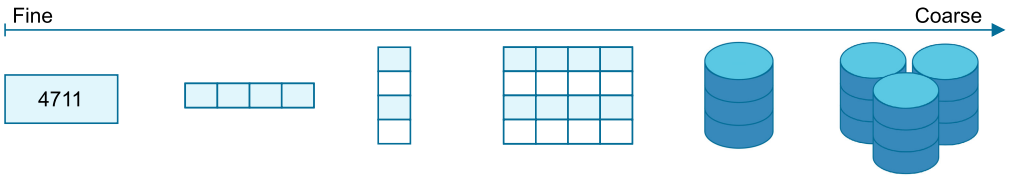
B. TOWARDS A COMMON INTERPRETATION IN THE SCIENTIFIC COMMUNITY

Many documents that discuss data catalog tools focus primarily on commercial solutions [5], [6], [9], are not peer-reviewed, or do not clearly describe the methodology used. This highlights the need for more rigorous research on data catalogs.

1) EXPRESSIVENESS OF DOMAIN KNOWLEDGE REPRESENTATIONS

Although semantic technologies (e.g., ontologies and knowledge graphs) have been available for many years [27], analysis A1 revealed that they are not widely used by data catalog tools – neither for representing (structural) metadata of data assets, nor for representing domain knowledge. According to a comparison of various data models by Feilmayr and Wöß [27], ontologies are the most expressive data model available. Therefore, we argue that the use of ontologies will be key to model complex real-world domains.

TABLE 3. The terminology used in various publications and data models for the levels of data granularity.

Data model / Literature	Levels of data granularity 					
Even and Shankaranarayanan [59] (based on [60])	Data item	Data record	Attribute	Dataset	Database	Database collection
Relational model [61], [62]	Value	Tuple	Attribute	Relation	Database	-
Relational database systems (e.g., PostgreSQL)	Value	Row	Column	Table	Database	-
Spreadsheets (e.g., Microsoft Excel)	Cell	Row	Column	Sheet	File	-
OWL-based ontologies [63]	Data property / object property	Individual	OWL constructor: `DataAllValuesFrom`	Class	Ontology	Ontology

In alignment with [56], which investigated ontologies and knowledge graphs for digital twins, we claim that knowledge graphs are an important technology for modeling semantic relations and their roles.

Most of the tools analyzed use concepts with limited expressiveness to represent domain knowledge (e.g., tags, business glossaries). Only three tools use knowledge graphs or ontologies. We therefore conclude that more research on combining data catalogs with data models of high semantic expressiveness is needed, for instance in the vein of [15], where a generic ontology layer provides data catalogs with enhanced semantics. In this context, also the Data Catalog Vocabulary (DCAT) [57] is noteworthy. The vocabulary is based on the Resource Description Framework (RDF) and the Web Ontology Language (OWL) and can therefore be used when modeling ontologies. DCAT provides terms for representing data catalogs and terms for cataloged datasets and data services (e.g., APIs). To meet the requirements of data portals in Europe and to enable searching across multiple portals, the European Union (EU) offers the specification DCAT-AP [58], which provides guidelines for using the DCAT vocabulary.

2) IMPORTANCE OF FINE-GRAINED DATA

As outlined in analysis criterion A2, particular functionalities require metadata at particular levels of granularity. For instance, detailed data lineage requires column-level metadata, and annotation of tables with context for all rows, such as why a row is part of a table, requires table-level metadata.

The terminology used to describe data levels varies significantly in the literature. Table 3 shows how each level is referred to and used in different contexts. In the remainder of this section, the terminology of relational database systems is used.

The finest level of granularity is the "value" level. Each value belongs to a "column" and a "row", which are both one level coarser than the value level [59]. The "table" level contains multiple columns and multiple rows [59]. The fifth level of granularity, the "schema", is optional and therefore not included in Table 3. Some relational database systems, such as PostgreSQL internally structure tables and other objects at this level [64]. Note that this interpretation of the term "schema" should not be confused with the "schema" that defines the structure of tables (cf. [62]). Depending on whether the schema level is available, a database contains multiple tables and multiple schemas. The coarsest level of granularity, which includes several databases, is referred to as a "database collection" (cf. [59]).

Our analysis revealed that the finest level of data granularity supported by almost all (92 %) of the data catalog tools analyzed is the column level. Interestingly, there is no term in the DCAT vocabulary that allows levels of granularity below the table level to be represented. Based on the observations in this survey, we therefore suspect that the lack of fine-grained granularity levels might be a reason for the low adoption rate of ontologies and knowledge graphs in data catalog tools. In [15], the authors show how to support finer levels of granularity by combining DCAT with the Data Source Description Vocabulary (DSD), which allows to model the internal structure of tables.

3) POTENTIAL OF ARTIFICIAL INTELLIGENCE

The results of applying analysis criterion A4 clearly showed that use of AI in data catalogs is on the rise, since approximately half of the tools examined (37 of 75) incorporate such functionalities in various aspects. As shown in Figure 9, most AI-based functionalities cover tasks that are somewhat tedious for humans. This includes, for instance, discovering data sources and relationships between them, cataloging (classification and tagging of data sources), and generating business metadata or suggestion functions.

While AI-based functionalities relieve humans of tedious tasks, they also come with challenges and ethical concerns. The ethical issues surrounding AI-based functionalities fall under the broader term of ''AI governance'', which refers to the rules and policies that ensure ethical, transparent, and responsible use of AI technologies [65]. Schneider et al. [66] defined AI governance not from a philosophical or societal perspective, but from a business standpoint, combining ''corporate governance'' with ''artificial intelligence''. AI governance for businesses includes the rules, practices, and processes which ensure that an organization's AI technology supports and improves business strategies and objectives [66].

In particular, data privacy and security pose significant ethical challenges. AI systems often process vast amounts of sensitive or personal data, which requires strict measures to prevent unauthorized access and data breaches [65], [66]. The General Data Protection Regulation (GDPR) [67], which was put into effect in May 2018, plays an important role in this context. It sets out strict requirements for the management and protection of personal data of individuals in the European Union (EU). However, the GDPR is not only relevant to organizations operating in the EU, but also has global implications. The requirements apply to all organizations, regardless of their location, that process, transfer, or store personal data of EU individuals. Hence, this also affects non-European companies, which must comply with the requirements of the GDPR if they handle such data [66], [67]. Companies must therefore ensure that information is collected lawfully and used ethically and that the rights of data owners are protected. Failure to comply with the GDPR can result in heavy fines [67].

Without a data catalog, companies face significant challenges in keeping track of all the data stored, as it is often distributed across multiple systems, and thus difficult to locate and manage. This makes it difficult to apply appropriate protective measures and categorizes data according to its sensitivity, and consequently leads to an increased risk of data breaches and non-compliance with regulations such as the GDPR. Additionally, every EU individual has the right to withdraw his or her consent to data processing at any time. In the course of this withdrawal, individuals can also request deletion of their personal data [67]. If a company does not have an overview of where they keep what data, then finding personal data becomes difficult and the risk of overlooking data is high. However, with a suitable data catalog tool, a company can identify exactly where each data element is stored, which makes it easier to handle personal data in accordance with legal requirements. Further, a data catalog tool makes it possible to classify data based on sensitivity and to label datasets as either personal identifiable information (PII) or non-PII. Some of the tools, such as Castor Data Catalog or securiti Data Catalog, already offer such features. Several tool vendors emphasize that data catalogs play an important role in GDPR compliance (e.g., [68], [69], [70]).

In addition to supporting compliance with the GDPR, data catalog tools also open up potential to ensure privacy and ethical requirements of AI applications. First, if a data catalog tool offers options for fine-grained data annotation, it enables organizations to exclude data classified as sensitive (personal) from the AI training process. Second, annotating data of insufficient quality or trustworthiness can also be useful in order to exclude particular data from AI training processes and thus avoid bias. Without a data catalog tool, this task would be very difficult to accomplish. These two aspects are of great importance considering that once data have been used in an AI training process, it cannot be removed from the model.

Transparency of AI decision-making processes to achieve explainable AI remains an ongoing challenge [65], [71]. This is problematic, especially for practitioners, when large language models (LLMs) are implemented in data catalog tools, for instance, to generate metadata or to automatically align mappings between data sources. Currently, most data catalog tool vendors do not provide concrete information on how such AI features are realized. Practitioners need more detailed information for tool selection and for assessing the results of these tools than is currently available. In total, 16 tools claim to provide functionalities based on generative AI, and some of them explicitly mention the use of LLMs, such as ChatGPT. Although application of generative functions may initially sound promising, there are distinct disadvantages. Due to the non-deterministic nature of LLMs, their results cannot be taken at face value and must be questioned [72], and even small errors in the tagging or classification of data sources can have a large impact [73]. This absence of transparency can lead to mistrust and difficulties in auditing AI actions and results. Additionally, explaining compliance actions to regulators – which is required by GDPR Article 5 [67] – may thus become problematic. Recently, the AI Act [74] has come into force, which has created a harmonized regulatory framework to ensure safe, transparent, and ethical development and use of AI in the EU.

In line with current trends and the increasing importance of AI technologies in enterprises, many of the tools analyzed advertise their AI capabilities to varying degrees. Although the potential benefits of AI are significant, ethical aspects and the possible impact of errors caused by such systems must be taken into consideration. Organizations must be aware of the risks involved and implement measures to ensure data privacy, transparency, and reliability of AI

decisions. Therefore, we encourage practitioners to perform a comprehensive risk assessment when selecting a data catalog tool with AI capabilities.

C. RELEVANCE FOR PRACTITIONERS

Since the target audience of this paper includes not only scientists but also practitioners, this section discusses aspects that can support practitioners in selecting the most suitable tool for their needs.

1) FOCUS OF DATA CATALOG TOOLS

The analyses performed as part of this survey revealed that the lack of a common definition of data catalogs is reflected in the various functionalities that are emphasized by the various tools. In addition to the cataloging functionality, some tools prioritize, for instance, data lineage and analytics (e.g., Qlik), whereas others focus on data quality (e.g., rudol). This highlights the multifaceted nature of data catalogs and the range of needs they aim to address within organizations.

The landscape of data catalog solutions is also diverse with respect to the analysis criteria. For instance, some tools offer comprehensive business glossaries and extensive connector support (e.g., OvalEdge, OpenMetadata, Talend Data Catalog), which makes them a robust choice for organizations with a variety of data sources or large data volumes. However, the potential lack of AI-based functionalities and, in particular, advanced ML capabilities for some solutions may limit their effectiveness in more demanding environments (e.g., Ab Initio, DataHub, securiti Data Catalog).

Platforms with flexible licensing or deployment options might be attractive (e.g., sifflet, Accurity Software Suite), but some of these tools (e.g., OpenDataDiscovery) lack AI integration and offer limited connectivity, which can be a disadvantage for data-intensive organizations.

Some tools fulfill simpler data governance needs with features such as tag-based knowledge representation and limited AI capabilities, but may struggle with scalability due to a small number of connectors or limited deployment options (e.g., AWS Glue Data Catalog, magda Data Catalog).

Tools that lack sufficient information to be evaluated against specific analysis criteria (e.g., dataspot, kyrah, Truedat) are less useful to practitioners. Without a clear evaluation of key features such as connector support, AI capabilities, and deployment options, organizations are left with uncertainties that may lead to suboptimal tool selection. Since every tool has advantages and disadvantages, we recommend considering the use cases for which it will be applied when deciding on a particular solution.

2) CONNECTORS TO DATA SOURCES AND APPLICATIONS

Analysis A3 revealed that data catalog tool vendors tend to promote a large number of different data source connectors as an advantage. However, a large number of different connectors may not always be necessary or beneficial. While a wide range of connectors offers high adaptability to various types of data sources, a smaller range can be preferable for simplicity. We claim that basic functionalities, such as gathering metadata, can often be enabled adequately by generic connectors, such as a Java Database Connectivity (JDBC) connector, which is typically capable of connecting to various (relational) database management systems (DBMSs) [75].

While Section III-C details the number and types of connectors offered by data catalog tools, this section provides concrete examples of the most commonly used connectors. The most widely used are application connectors to analytics platforms: 39 of the 42 tools ($\approx$ 93 %) that generally support application connectors, also support some type of analytics platform, such as Tableau [76], Power BI [77], and Looker [78] (e.g., Ataccama One, OvalEdge Data Catalog, and DataGalaxy). In addition, connectors for Apache Superset [79] (e.g., Amundsen Data Catalog), Qlik Sense [80] (e.g., Octopai), and Apache Spark [81] (e.g., Ataccama One) are mentioned. Three tools ($\approx$7 %) do not provide analytics connectors (e.g., Metacat), which may limit their usefulness for organizations that rely on seamless access to analytics applications.

In addition to analytics platforms, several other types of applications (cf. Section III-C) are supported, such as collaboration and project management tools (e.g., Jira [82] and Slack [83]), data quality tools (e.g., Great Expectations [84]), and data streaming platforms (e.g., Apache Kafka [85]). For instance, OvalEdge categorized their connectors into databases, data warehouses, reporting systems, data lakes, applications, ETLs, streaming systems, and other tools that do not fit the aforementioned categories. OvalEdge includes, among others, connectors for Apache Kafka and Jira.

The ability to connect a data catalog tool to both data sources and different types of applications (e.g., analytics platforms, collaboration platforms, data quality tools) can significantly improve its functionality and versatility. While connecting to data sources is essential for cataloging and metadata ingestion, connecting to other applications is valuable but not required. A data catalog limited to data source connectors offers simplicity, which is beneficial for small and medium-sized companies with limited resources. However, additional connectors, especially for analytics applications, enable meaningful use of data by allowing analysts to quickly generate reports and visualizations. Ultimately, effective data governance and analytics strategies depend on the integration of data catalogs within a broader ecosystem.

3) LICENSING

Licensing plays a crucial role in the decision to introduce a data catalog tool to an organization. As revealed by analysis A5, most data catalog tools (i.e., 50) are available as feature-complete products under various licensing conditions (one-time purchase vs. subscription-based), while 15 tools use feature-based licensing. Both variants are commonly used in software licensing.

In both categories, feature-complete and feature-based licensing, there are tool vendors using subscription-based

licensing. Vendors typically use the number of users (''pay per user'', e.g., dataspot), the number of datasets cataloged (''pay per dataset'', e.g., Anjana Data) or a combination of both (e.g., Decube or Select Star) to calculate the exact costs. These calculation options open up scope for discussion: Although they may seem cost-effective at first glance, we argue that this pricing can be problematic, particularly in the context of data cataloging. We believe that a data catalog loses much of its potential if it catalogs only some of the data sources of an organization or allows only a limited number of users to access it. Some tool vendors (e.g., dataspot) have already recognized this problem and exclude users with ''read-only'' permission from the cost calculation.

Although most tools are proprietary (63 of 75), a growing number of general-purpose open source data catalog tools with an active community have emerged (e.g., DataHub or OpenMetadata). A list of all open-source data catalog tools found is available as part of the supplementary material [19]. The rapid development of these tools should be closely monitored, as they can become an alternative to commercial data catalog tools for certain use cases.

4) TYPES OF DEPLOYMENT

In addition to licensing, the deployment types available are also key in selecting a data catalog tool. Analysis A6 distinguished between tools that can be installed on-premise and those available as SaaS. Some tool websites mention having a ''hybrid'' type of deployment, which means that the tool itself is provided as SaaS and connects to an on-premise software component when connecting to data sources. Since the metadata of data sources is consistently loaded into the SaaS-based tool, hybrid is not treated as a separate category.

In agreement with Jahnke and Otto [3], we argue that each type of deployment comes with advantages and disadvantages. On-premise installations entail more maintenance effort for IT departments because tools must be installed, updated and monitored, but have the advantages that the metadata loaded stays within the organization and that connections to highly protected data sources can be realized. In contrast, SaaS-based data catalog tools are maintained and operated by their vendors [34], which reduces the effort for the organization using the tool. When on-premise data sources are loaded into the SaaS-based data catalog tool, information about the structure of data sources and information systems leaves the company, which may be problematic for data protection reasons. Some tools (e.g., Data Cookbook, rudol) also provide data profiling functionalities. This can lead to privacy compliance issues (e.g., with respect to the European GDPR [67]) when data sources containing personal data are cataloged.

5) ROLE OF DUPLICATES IN DATA CATALOGS

There are two types of duplicates: (i) records within a single database (or table) that represent the same entity (e.g., an entity is stored multiple times within a database or table) and (ii) duplicates of entity types caused by distribution (e.g., a company has two company locations in different cities, and the same entity type is stored locally at databases in both locations). This means that we differentiate between duplicates at the instance level and at the schema level. While detection of the first type of duplicates is considered an important research topic with many relevant research contributions (e.g., [86]), we argue that this is not a task primarily performed by a data catalog tool. Removal of duplicates at the instance level is a main task of data quality tools. For an analysis of existing data quality tools and their ability to detect duplicates in a database, see the survey of data quality tools by Ehrlinger and Wöß [87] and the survey by Goasdoué et al. [88] with a more detailed evaluation of duplicate detection.

A data catalog tool should detect duplicates at the schema level. Consider the following example: If a company has a dedicated production database and a payroll database, both containing an employee table. The schema of the employee table in the production database is `Employee(ID, Name, SSN, Dep_ID, Role, Hire_Date)`, while the employee data in the payroll database has the structure `Employee(ID, Name, SSN, Salary, Tax_ID, Bank_Account)`. There will be an overlap of entities in these databases because some employees will be stored in both tables, and a data catalog tool should automatically detect these duplicates at the schema level.

Of the 75 tools analyzed, only a few support automatic duplicate detection, but none provide comprehensive details on the methods used. Some tool vendors do not even describe the type of duplicate detection that their tool supports. Others, including magda and Ab Initio, explicitly mention support of type (ii) duplicate detection. Some tools, such as AWS Glue, do not have built-in automatic entity type duplicate detection, but offer features that facilitate manual deletion of duplicate rows (type (i) duplicates) once they have been identified, while other tools offer type (i) duplicate detection as part of their built-in data quality functionalities (e.g., Dataedo Data Catalog, DataGalaxy. Some platforms, such as Oracle Cloud Infrastructure Data Catalog, Oracle Enterprise Metadata Management, and Octopai, emphasize the importance of duplicate detection and data deduplication on their websites in several blog posts [89], [90], but do not mention implementation of duplicate detection within their tools. Finally, some tools leverage ML techniques for duplicate detection, mainly applying similarity-based measures, such as fuzzy matching.

6) OPINION OF PRACTITIONERS

At the 17th International Chief Data Officer and Information Quality (CDOIQ) Symposium 2023 in Boston, Massachusetts, one of the authors of this survey (L.E.) conducted a poll with the audience of her talk on ''The Impact of Metadata on Data Quality''. Participants were asked to answer the question ''Which features would you expect from

your data catalog tool?'' based on seven predefined options (multiple selection was possible). The results of the poll are shown in Figure 11.

Of all options, cataloging was rated the most important by 77 of 90 participants, which indicates a strong consensus.

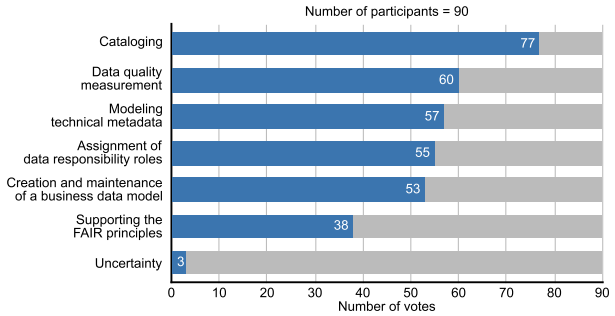


FIGURE 11. Results of the poll regarding data catalog functionality at the CDOIQ 2023.

Measuring data quality received fewer votes than cataloging (60 of 90). This shows that participants recognized the importance of measuring and ensuring the quality of the data within a data catalog. With 57 votes, technical metadata modeling received a similarly high level of support. 55 votes indicated that participants recognized the importance of assigning responsibility roles within a data catalog – an aspect highlighted in [8] but not taken into account in this survey. Labadie et al. [2] found that traditional roles, such as data owner, and new roles, such as data analyst, who exclusively consumes data, are among the most important user roles in the context of data catalogs.

Although slightly fewer participants prioritized the creation of a business data model to represent domain knowledge, with a total of 53 votes the importance of this aspect was recognized. 38 participants rated the FAIR principles [38] as important in data catalogs. With only three votes, ''uncertainty'', which was intended to reflect the trustworthiness of the objects cataloged, appeared to be less of a concern.

The results of the poll do not acknowledge the importance of domain knowledge as a main functionality of data catalogs. This is particularly interesting because cataloging of any type of items always involves adding additional context, such as domain knowledge. Without context, it would be impossible to ever find an item in a catalog again. There may be two reasons for the low number of votes for a business data model: (i) the IT department is often erroneously considered to be the primary target group of data catalogs and (ii) the term ''business data model'' may imply a very high level of expressiveness of the business context, similar to that of ontologies, which – as described previously – are not widely used.

7) GUIDANCE ON SELECTING A DATA CATALOG TOOL

The supplementary material contains the table with the complete results for all data catalog tools analyzed [19]. The table is intended for practitioners and enables them to evaluate and select data catalog tools that offer the desired functionalities. Figure 12 shows an excerpt of the evaluation table, which can be adapted to the respective selection criteria using sorting and filter functions. A practitioner can use our analysis results directly and search for suitable data catalog tools.

An example use case from the automotive industry was reported in [15]: Our industry partner had already deployed a data catalog tool to automatically collect technical metadata from a large number of heterogeneous data sources. In the course of our collaboration, new requirements were identified, such as the need for a highly expressive representation of domain knowledge (business data model in [15]) that provides an abstract view of the underlying technical data sources. Their existing data catalog tool did not meet this requirement, which was why they had to arrive at an informed decision on a more suitable replacement. From our previous experience with industry partners (as also discussed in [8], [91], [92]), we argue that selecting a data catalog tool based on predefined requirements is essential.

Provision of our survey results in the form of a table allows practitioners to filter a comprehensive list of data catalog tools according to predefined requirements. By applying such filters to our analysis criteria, practitioners can customize their search and find the tool best suited to their needs, thus significantly improving their ability to integrate a data catalog solution into their organization. The table contains the analysis results for each tool based on type of domain knowledge representation, availability of fine-grained metadata, number of connectors, and presence of AI-based functionalities. Additionally, it provides information on licensing models and deployment types. This process enables practitioners and data leaders, such as Chief Data Officers, to evaluate the data catalog tool landscape and to minimize the risk of bad investments.

D. THREATS TO VALIDITY

At least two authors thoroughly analyzed each primary survey to address potential validity concerns regarding subjectivity [93]. Unclear assessments regarding the selection of primary surveys or the exclusion of tools according to predefined criteria were examined in detail and discussed within the team before being reassessed.

In order to address limitations caused by a lack of robustness and to obtain the most comprehensive list of data catalog tools, three different platform types were queried in the systematic search: Scientific literature (ACM Digital Library, IEEE Xplore, ResearchGate, ScienceDirect, Springer Link, and Google Scholar), source code (GitHub and GitLab), and Google for a general search.

The search expression was expanded by including alternatives to the key terms ''data catalog'' and ''tool'' to reduce the risk of missing relevant tools. Precision was increased by using a conjunction to combine the term ''data catalog'' (or a synonym) with the term ''tool'' (or a synonym). This ensured

				Analysis 3: Connectors			Analysis 4: AI	
Tool name	Vendor / Developer	Analysis 1: Available type of domain knowledge representation	Analysis 2: Availability of fine-granular metadata on attribute level	Number of avaialbe connectors	Connectors available to data sources (e.g., relational databases)	...	Presence of AI-based functionalities	...
Ab initio Data Catalog	Ab inito	Business glossary available	Yes	50..99	☑		Not explicitly mentioned	
Accurity Software Suite	Accurity by Simplity	Business glossary and tags available	Yes	No evaluation possible	No evaluation possible		Not explicitly mentioned	
Alation Data Catalog	Alation	Business glossary available	Yes	100..∞	☑		Yes	
Amundsen Data Catalog	Amundsen	Tags available	Yes	1..49	☑		Yes	
Anjana Data	Anjana Data	Business glossary available	Yes	50..99	☑		Yes	
Apache Atlas	Apache Atlas	Business glossary and tags available	Yes	1..49	☑		Not explicitly mentioned	
Apgar Data Catalog	Apgar	Business glossary available	Yes	100..∞	☑		Not explicitly mentioned	
Arena	Zaloni	Business glossary and tags available	Yes	100..∞	☑		Yes	
Ataccama ONE	Ataccama	Business glossary available	Yes	100..∞	☑		Yes	

FIGURE 12. An excerpt of the table containing the analysis results for the evaluated tools, focusing on a subset of columns, while the full table contains additional information and columns.

that only results specifically related to data catalog tools were retrieved.

Google returned search results titled "the 10 best data catalog tools" and similar. Since the websites presented these tools as data catalog tools, they were included for further investigation. However, closer examination in the subsequent exclusion process revealed that these websites featured some tools that were not related to data catalogs and often had a (completely) different focus (e.g., data quality or health monitoring). Exclusion criteria E2 and E3 were introduced to mitigate this bias and ensure that the final list of tools included only general-purpose data catalog tools.

VI. CONCLUSION AND OUTLOOK

This paper presented (i) an analysis and discussion of key data cataloging functionalities and (ii) a systematic survey evaluating the extent to which existing data catalog tools implement the main data cataloging functionalities. The survey results aim to assist practitioners in evaluating data catalog tools and to promote a common interpretation of data catalogs within the scientific community.

From a practical point of view, we claim that there remains room for improvement in the functional scope of current data catalog tools, especially with respect to the representation of domain knowledge. Development and implementation of an ontology-based knowledge representation for data catalogs is a promising approach to which we attribute great potential. The expressiveness ontologies provide allows modeling of complex scenarios (required in most enterprise settings), which is not possible with most concepts currently used in data catalog tools to represent domain knowledge, as revealed in Section III-C. Another advantage of semantic web technology is their reuse of and linkage with publicly available vocabularies (e.g., DCAT and DSD as used in [15]), which improves interoperability between systems.

A commonly agreed definition of data catalogs would be of utmost importance to the scientific community. Although this paper provides a solid basis for proposing such a definition, a discussion of the definition itself is beyond the scope of this survey and considered future work. Finally, there are topics that are mentioned in the literature on data catalogs (e.g., compliance of individual tools with the FAIR principles), but have not yet been investigated by a survey.

All links in this paper were last visited in April 2025.

APPENDIX
THE DATA CATALOG TOOLS ANALYZED

All search results per platform, the surveys found and their classification, and the analysis of the tools are available in our supplementary material [19]. The 75 data catalog tools analyzed in this paper are listed below in alphabetical order.

- Ab initio Data Catalog
  Developer: Ab inito
  URL: https://www.abinitio.com/en/data-catalog-quality-governance/
- Accurity Software Suite
  Developer: Accurity by Simplity
  URL: https://www.accurity.ai/
- Alation Data Catalog
  Developer: Alation
  URL: https://www.alation.com/
- Amundsen Data Catalog
  Developer: Amundsen
  URL: https://www.amundsen.io/
- Anjana Data
  Developer: Anjana Data
  URL: https://anjanadata.com/en/
- Apache Atlas
  Developer: Apache Atlas
  URL: https://atlas.apache.org/#/
- Apgar Data Catalog
  Developer: Apgar
  URL: https://apgar-group.com/our-expertises/data-catalog/

- Arena
  Developer: Zaloni
  URL: https://www.zaloni.com/arena-overview/
- Ataccama ONE
  Developer: Ataccama
  URL: https://www.ataccama.com/platform
- Atlan Data Discovery & Catalog
  Developer: Atlan
  URL: https://atlan.com/data-discovery-catalog/
- AWS Glue Data Catalog
  Developer: AWS
  URL: https://docs.aws.amazon.com/glue/latest/dg/what-is-glue.html
- BigID Data Intelligence Platform
  Developer: BigID
  URL: https://bigid.com/data-intelligence-platform/
- Boomi Data Catalog & Preperation
  Developer: Boomi
  URL: https://boomi.com/platform/data-catalog-and-preparation/
- Castor Data Catalog
  Developer: Castor
  URL: https://www.castordoc.com/product/data-catalog
- CKAN
  URL: https://ckan.org/
- Cloudera Navigator
  Developer: Cloudera
  URL: https://www.cloudera.com/products/product-components/cloudera-navigator.html
- Colid
  Developer: Bayer
  URL: https://github.com/Bayer-Group/colid-documentation/
- Collibra Catalog
  Developer: Collibra
  URL: https://www.collibra.com/us/en/products/data-catalog
- D-QUANTUM Data Catalog
  Developer: Synabi
  URL: https://synabi.com/data-catalog/
- Data Cookbook Data Catalog & System Inventory
  Developer: Data Cookbook
  URL: https://www.datacookbook.com/datacatalogsysteminventory
- Data Integrity Suite
  Developer: Precisely
  URL: https://www.precisely.com/de/product/data-integrity/precisely-data-integrity-suite
- Data X-Ray
  Developer: Ohalo
  URL: https://www.ohalo.co/platform
- data.world Data Catalog
  Developer: data.world
  URL: https://data.world/
- Dataedo Data Catalog
  Developer: Dataedo
  URL: https://dataedo.com/product/data-catalog
- DataGalaxy
  Developer: DataGalaxy
  URL: https://www.datagalaxy.com/
- DataHub
  Developer: DataHub Project
  URL: https://datahubproject.io/
- Datameer Cloud
  Developer: Datameer
  URL: https://www.datameer.com/
- dataspot. Data Catalog
  Developer: dataspot.
  URL: https://dataspot.at/
- Decube
  Developer: Decube Data
  URL: https://www.decube.io/
- DvSum
  Developer: DvSum
  URL: https://dvsum.ai/
- Enterprise Data Catalog
  Developer: Informatica
  URL: https://www.informatica.com/de/products/data-catalog/enterprise-data-catalog.html
- erwin Data Catalog
  Developer: Quest
  URL: https://www.erwin.com/de-de/products/erwin-data-catalog/
- erwin Data Intelligence
  Developer: Quest
  URL: https://www.erwin.com/de-de/products/erwin-data-intelligence/
- Google Cloud Data Catalog
  Developer: Google
  URL: https://cloud.google.com/data-catalog/docs/concepts/overview?hl=de
- IBM InfoSphere Information Governance Catalog
  Developer: IBM
  URL: https://www.ibm.com/products/information-governance-catalog
- IBM Knowledge Data Catalog
  Developer: IBM
  URL: https://www.ibm.com/cloud/watson-knowledge-catalog
- K2View Data Product Platform
  Developer: K2View
  URL: https://www.k2view.com/platform/data-product-platform/
- Kylo
  URL: https://kylo.io/
- Kyrah
  Developer: Cogniflare
  URL: https://www.kyrah.io/
- magda Data Catalog
  Developer: magda
  URL: https://magda.io/
- MAGGOT
  URL: https://github.com/inrae/pgd-mmdt

- Metacat
  Developer: Netflix
  URL: https://github.com/Netflix/metacat
- MetaCenter
  Developer: Data Advantage Group
  URL: https://www.dag.com/metacenter
- Metadata Management System
  URL: https://github.com/serginf/MDM
- Metaphor Data Catalog
  Developer: Metaphor
  URL: https://metaphor.io/
- Octopai
  Developer: Octopai
  URL: https://www.octopai.com/
- OneTrust Data Governance
  Developer: OneTrust
  URL: https://www.onetrust.com/solutions/data-governance/
- OpenDataDiscovery Platform
  Developer: Open Data Discovery
  URL: https://opendatadiscovery.org/
- OpenMetadata
  Developer: OpenMetadata
  URL: https://open-metadata.org/
- Oracle Cloud Infrastructure Data Catalog
  Developer: Oracle
  URL: https://www.oracle.com/big-data/data-catalog/
- Oracle Enterprise Metadata Management
  Developer: Oralce
  URL: https://www.oracle.com/middleware/technologies/enterprise-metadata-management.html
- Orion Enterprise Information Intelligence Graph (EIIG)
  Developer: Orion Governance
  URL: https://www.oriongovernance.com/orion-platform/
- OvalEdge Data Catalog
  Developer: OvalEdge
  URL: https://www.ovaledge.com/data-catalog
- Pentaho Data Catalog
  Developer: Hitachi Vantara
  URL: https://www.hitachivantara.com/en-us/products/dataops-software/data-catalog.html
- Promethium
  Developer: Promethium
  URL: https://www.promethium.ai/metadata-data-catalog
- Purview Data Catalog
  Developer: Microsoft
  URL: https://learn.microsoft.com/en-us/azure/purview/how-to-browse-catalog
- Qlik Catalog
  Developer: Qlik
  URL: https://www.qlik.com/us/products/catalog-and-lineage
- Rocket Data Intelligence
  Developer: Rocket Software
  URL: https://www.rocketsoftware.com/products/rocket-data-intelligence
- rudol
  Developer: rudol
  URL: https://rudol.ai/
- SAP Information Steward
  Developer: SAP
  URL: https://www.sap.com/austria/products/technology-platform/data-profiling-steward.html
- Secoda Data Catalog
  Developer: Secoda
  URL: https://www.secoda.co/data-catalog
- securiti Data Catalog
  Developer: securiti
  URL: https://securiti.ai/products/data-catalog/
- Select Star Data Catalog
  Developer: Select Star
  URL: https://www.selectstar.com/
- Sidecar
  Developer: Sidecar
  URL: https://www.sidecar-data.ch/
- Sifflet
  Developer: Sifflet
  URL: https://www.siffletdata.com/
- Sled
  Developer: Sled
  URL: https://www.sled.so/
- Solidatus
  Developer: Solidatus
  URL: https://www.solidatus.com/
- SQL Data Catalog
  Developer: Redgate
  URL: https://www.red-gate.com/products/dba/sql-data-catalog/
- Tableau Catalog
  Developer: Tableau
  URL: https://www.tableau.com/products/add-ons/catalog
- Talend Data Catalog
  Developer: Talend
  URL: https://www.talend.com/products/data-catalog/
- ThinkData Works Catalog
  Developer: TinkData works
  URL: https://www.thinkdataworks.com/#
- Tibco Cloud Metadata
  Developer: Tibco
  URL: https://www.tibco.com/products/tibco-cloud-metadata
- Truedat
  Developer: Truedat
  URL: https://www.truedat.io/
- Y42
  Developer: Datos-Intelligence
  URL: https://www.y42.com/product/catalog-lineage/
- Zeenea Data Catalog
  Developer: Zeenea
  URL: https://zeenea.com/de/daten-katalog/

REFERENCES

[1] T. Korte, M. Fadler, M. Spiekermann, C. Legner, and B. Otto, *Data Governance*. Cham, Switzerland: Springer, 2023.

[2] C. Labadie, C. Legner, M. Eurich, and M. Fadler, "FAIR enough? Enhancing the usage of enterprise data with data catalogs," in *Proc. IEEE 22nd Conf. Bus. Informat. (CBI)*, vol. 1, Antwerp, Belgium. IEEE, Jun. 2020, pp. 201–210.

[3] N. Jahnke and B. Otto, "Data catalogs in the enterprise: Applications and integration," *Datenbank-Spektrum*, vol. 23, no. 2, pp. 89–96, Jul. 2023, doi: 10.1007/s13222-023-00445-2.

[4] R. Bean. (2021). *Why Is It So Hard To Become a Data-Driven Company?*. [Online]. Available: https://hbr.org/2021/02/why-is-it-so-hard-to-become-a-data-driven-company

[5] E. Zaidi, G. De Simoni, R. Edjlali, and D. Alan D. (2017). *Data Catalogs Are the New Black in Data Management and Analytics*. [Online]. Available: https://www.gartner.com/en/documents/3837968

[6] G. De Simoni, M. Beyer, A. Jain, and A. Dayley. (2020). *Magic Quadrant for Metadata Management Solutions*. [Online]. Available: https://www.gartner.com/en/documents/3993025

[7] *Data Quality—Part 8: Information and Data Quality Concepts and Measuring. Standard*, ISO Standard 8000-8:2015. [Online]. Available: https://www.iso.org/standard/60805.html

[8] L. Ehrlinger, J. Schrott, M. Melichar, N. Kirchmayr, and W. Wöß, "Data catalogs: A systematic literature review and guidelines to implementation," in *Proc. Database Expert Syst. Appl. (DEXA Workshops)*. Cham, Switzerland: Springer, Jan. 2021, pp. 148–158.

[9] T. Korte, M. Fadler, M. Spiekermann, C. Legner, and B. Otto, *Data Catalogs—Integrated Platforms for Matching Data Supply and Demand. Reference Model and Market Analysis (Version 1.0).* Stuttgart, Germany: Fraunhofer Verlag, 2019.

[10] M. Franklin, A. Halevy, and D. Maier, "From databases to dataspaces: A new abstraction for information management," *ACM SIGMOD Rec.*, vol. 34, no. 4, pp. 27–33, Dec. 2005, doi: 10.1145/1107499.1107502.

[11] J. Hilger and Z. Wahl, "Data catalogs and governance tools," in *EnMaking Knowledge Management Clickable*. Cham, Switzerland: Springer, 2022, pp. 187–192. [Online]. Available: https://link.springer.com/10.1007/978-3-030-92385-311

[12] E. Quimbert, K. Jeffery, C. Martens, P. Martin, and Z. Zhao, "Data cataloguing," in *Towards Interoperable Research Infrastructures for Environmental and Earth Sciences*, vol. 12003. Cham, Switzerland: Springer, 2020, pp. 140–161.

[13] S. Shanmugam and G. Seshadri, "Aspects of data cataloguing for enterprise data platforms," in *Proc. IEEE IEEE 2nd Int. Conf. Big Data Secur. Cloud (BigDataSecurity) Int. Conf. High Perform. Smart Comput. (HPSC), IEEE Int. Conf. Intell. Data Secur. (IDS)*. New York, NY, USA, Apr. 2016, pp. 134–139.

[14] H. Dibowski, S. Schmid, Y. Svetashova, C. Henson, and T. Tran, "Using semantic technologies to manage a data lake: Data catalog, provenance and access control," in *Proc. 13th Int. Workshop Scalable Semantic Web Knowl. Base Syst. Co-Located 19th Int. Semantic Web Conf. (ISWC)*, Jan. 2020, pp. 65–80. [Online]. Available: https://ceur-ws.org/Vol-2757/SSWS2020paper5.pdf

[15] J. Schrott, S. Weidinger, M. Tiefengrabner, C. Lettner, W. Wöß, and L. Ehrlinger, "GOLDCASE: A generic ontology layer for data catalog semantics," in *Metadata and Semantic Research*. Cham, Switzerland: Springer, 2023, pp. 26–38.

[16] E. Zaidi and G. De Simoni. (2019). *Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders*. [Online]. Available: https://www.gartner.com/en/documents/3957301

[17] R. Gartner, *Metadata*. Cham, Switzerland: Springer, 2016.

[18] A. Boufassil, F. Bouhafer, M. Cherradi, and A. E. Haddadi, "Data catalog: Approaches, trends, and future directions," in *Proc. 17th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Bangkok, Thailand, Nov. 2023, pp. 369–376.

[19] *Supplementary Materials of the Paper a Survey of Data Catalog Tools*. Accessed: Apr. 30, 2025. [Online]. Available: https://zenodo.org/records/13959401

[20] B. Kitchenham and S. Charters. (2007). *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. [Online]. Available: https://www.researchgate.net/publication/302924724GuidelinesforperformingSystematicLiteratureReviewsinSoftwareEngineering

[21] P. Russom. (2017). *The Data Catalog's Role in the Digital Enterprise*. [Online]. Available: https://tdwi.org/research/2017/11/ta-all-informatica-the-data-catalogs-role-in-the-digital-enterprise/

[22] T. C. Redman, *People and Data: Uniting to Transform Your Business*. London, U.K.: Kogan Page, 2023.

[23] *Understanding the Search Syntax—GitHub Docs*. Accessed: Jan. 14, 2024. [Online]. Available: https://docs.github.com/en/search-github/getting-started-with-searching-on-github/understanding-the-search-syntax

[24] B. Kitchenham, L. Madeyski, and D. Budgen, "How should software engineering secondary studies include grey material?" *IEEE Trans. Softw. Eng.*, vol. 49, no. 2, pp. 872–882, Feb. 2023.

[25] J. Schöpfel, "Towards a Prague definition of grey literature," in *Proc. 12th Int. Conf. Grey Literature, Transparency Grey Literature. Grey Tech Approaches High Tech Issues*, Prague, Czech Republic, Dec. 2010, pp. 11–26. [Online]. Available: https://hal.science/sic00581570v1

[26] S. Winningham. (2019). *Knowledge Nugget: Bus. Glossary Vs. Data Dictionaries*. [Online]. Available: https://web.stanford.edu/dept/pres-provost/cgi-bin/dg/wordpress/knowledge-nugget-business-glossary-vs-data-dictionaries/

[27] C. Feilmayr and W. Wöß, "An analysis of ontologies and their success factors for application to business," *Data Knowl. Eng.*, vol. 101, pp. 1–23, Jan. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169023X1500110X

[28] MySQL. *MySQL Database*. Accessed: Apr. 30, 2025. [Online]. Available: https://www.mysql.com/

[29] *PostgreSQL*. Accessed: Apr. 30, 2025. [Online]. Available: https://www.postgresql.org/

[30] Amazon Web Services. *Amazon S3: Cloud Object Storage*. Accessed: Apr. 30, 2025. [Online]. Available: https://aws.amazon.com/s3/

[31] Google Cloud. *Google Cloud: Cloud Computing Services*. Accessed: Apr. 30, 2025. [Online]. Available: https://cloud.google.com/

[32] MongoDB. *MongoDB Database*. Accessed: Apr. 30, 2025. [Online]. Available: https://www.mongodb.com/

[33] Apache Softw. Foundation. *Apache Cassandra: Open Source NoSQL Database*. Accessed: Apr. 30, 2025. [Online]. Available: https://cassandra.apache.org/

[34] A. Zannou, A. Leshob, R. Rab, and P. Hadaya, "A method for selecting a suitable cloud computing deployment strategy," in *Proc. IEEE Int. Conf. E-Bus. Eng. (ICEBE)*, Nov. 2023, pp. 69–76.

[35] B. Dobreski, "Descriptive cataloging: The history and practice of describing library resources," *Cataloging Classification Quart.*, vol. 59, nos. 2–3, pp. 225–241, Apr. 2021, doi: 10.1080/01639374.2020.1864693.

[36] H. Philips, "The great library of Alexandria," *Library Philosophy Pract.*, vol. 2010, p. 22, Aug. 2010. [Online]. Available: https://digitalcommons.unl.edu/libphilprac/417/

[37] B. B. Tillett, "FRBR and cataloging for the future," *Cataloging Classification Quart.*, vol. 39, nos. 3–4, pp. 197–205, Apr. 2005, doi: 10.1300/j104v39n03_12.

[38] M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, Mar. 2016, Art. no. 160018. [Online]. Available: http://www.nature.com/articles/sdata201618

[39] R. Hai, C. Koutras, C. Quix, and M. Jarke, "Data lakes: A survey of functions and systems," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12571–12590, Dec. 2023.

[40] J. Macedo, N. Cacho, and F. Lopes, "A comparative study of tools for smart cities open data publication and management," in *Proc. IEEE 1st Summer School Smart Cities (S3C)*, Natal, Brazil, Aug. 2017, pp. 79–84.

[41] A. McSweeney. (2021). *Data Profiling, Data Catalogs and Metadata Harmonisation*. [Online]. Available: https://www.researchgate.net/publication/351282374DataProfilingDataCatalogsandMetadataHarmonisation

[42] M. I. S. Oliveira, G. D. F. B. Lima, and B. F. Lóscio, "Investigations into data ecosystems: A systematic mapping study," *Knowl. Inf. Syst.*, vol. 61, no. 2, pp. 589–630, Nov. 2019, doi: 10.1007/s10115-018-1323-6.

[43] H. D. A. D. Santos, M. I. S. Oliveira, G. D. F. A. B. Lima, K. M. da Silva, R. I. V. C. S. Muniz, and B. F. Lóscio, "Investigations into data published and consumed on the web: A systematic mapping study," *J. Brazilian Comput. Soc.*, vol. 24, no. 1, p. 14, Dec. 2018, doi: 10.1186/s13173-018-0077-z.

[44] *Zenodo*. Accessed: Apr. 30, 2025. [Online]. Available: https://zenodo.org/

[45] Invenio Softw. *Invenio*. Accessed: Apr. 30, 2025. [Online]. Available: https://inveniosoftware.org/

[46] Dataverse Project. *Dataverse*. Accessed: Apr. 30, 2025. [Online]. Available: https://dataverse.org/

[47] C. Quix, R. Hai, and I. Vatov, "GEMMS: A generic and extensible metadata management system for data lakes," in *Proc. CAiSE Forum 28th Int. Conf. Adv. Inf. Syst. Eng. (CAiSE)*, Jan. 2016, pp. 129–136. [Online]. Available: https://ceur-ws.org/Vol-1612/paper17.pdf

[48] T. H. Zhou, B. M. Heo, L. Wang, Y. K. Lee, D. J. Chai, and K. H. Ryu, "A prototype of multimedia metadata management system for supporting the integration of heterogeneous sources," in *Proc. Adv. Intell. Comput. Theories Appl. Aspects Theor. Methodol. Issues (ICIC)*. Heidelberg, Germany: Springer, Jan. 2008, pp. 1095–1102.

[49] Q. Xu, R. V. Arumugam, K. L. Yong, and S. Mahadevan, "Efficient and scalable metadata management in EB-scale file systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 11, pp. 2840–2850, Nov. 2014.

[50] A. Thomson and D. J. Abadi, "CalvinFS: Consistent WAN replication and scalable metadata management for distributed file systems," in *Proc. 13th USENIX Conf. File Storage Technol.*, Feb. 2015, pp. 1–14. [Online]. Available: https://www.usenix.org/sites/default/files/fast15fullproceedingsinterior.pdf

[51] M.-H. Cha, S.-M. Lee, D.-O. Kim, H.-Y. Kim, and Y.-K. Kim, "High performance metadata management engine for large-scale distributed file systems," in *Proc. 9th Int. Conf. Future Gener. Commun. Netw. (FGCN)*, Jeju, Jeju, South Korea, Nov. 2015, pp. 29–32.

[52] Y. Gao, X. Gao, X. Yang, J. Liu, and G. Chen, "An efficient ring-based metadata management policy for large-scale distributed file systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 30, no. 9, pp. 1962–1974, Sep. 2019.

[53] T. D. Thanh, S. Mohan, E. Choi, S. Kim, and P. Kim, "A taxonomy and survey on distributed systems," in *Proc. 4th Int. Conf. Networked Comput. Adv. Inf. Manage.*, Gyeongju, South Korea, Sep. 2008, pp. 144–149.

[54] H. Dai, Y. Wang, K. B. Kent, L. Zeng, and C. Xu, "The state of the art of metadata managements in large-scale distributed file systems—Scalability, performance and availability," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 12, pp. 3850–3869, Dec. 2022.

[55] N. Dawes, K. A. Kumar, S. Michel, K. Aberer, and M. Lehning, "Sensor metadata management and its application in collaborative environmental research," in *Proc. IEEE 4th Int. Conf. eSci.*, Indianapolis, IN, USA, Dec. 2008, pp. 143–150.

[56] E. Karabulut, S. F. Pileggi, P. Groth, and V. Degeler, "Ontologies in digital twins: A systematic literature review," *Future Gener. Comput. Syst.*, vol. 153, pp. 442–456, Apr. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X23004739

[57] W3C. *DCAT 3 Vocabulary*. Accessed: Apr. 30, 2025. [Online]. Available: https://w3.org/ns/dcat

[58] Eur. Union. *DCAT Application Profile for Data Portals in Europe (DCAT-AP)*. Accessed: Apr. 30, 2025. [Online]. Available: https://op.europa.eu/en/web/eu-vocabularies/dcat-ap

[59] A. Even and G. Shankaranarayanan, "Utility-driven assessment of data quality," *ACM SIGMIS Database: DATABASE Adv. Inf. Syst.*, vol. 38, no. 2, pp. 75–93, May 2007, doi: 10.1145/1240616.1240623.

[60] T. C. Redman, *Data Quality for the Information Age*. Norwood, MA, USA: Artech House, 1996.

[61] E. F. Codd, "A relational model of data for large shared data banks," *Commun. ACM*, vol. 13, no. 6, pp. 377–387, Jun. 1970, doi: 10.1145/362384.362685.

[62] E. F. Codd, "Further normalization of the database relational model," *Data Base Syst.*, vol. 1972, pp. 33–64, Jun. 1972.

[63] F. Baader, I. Horrocks, C. Lutz, and U. Sattler, *An Introduction to Description Logic*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[64] PostgreSQL Global Develop. Group. *PostgreSQL Documentation: 5.10. Schemas, Chapter 5. Data Definition*. [Online]. Available: https://www.postgresql.org/docs/current/ddl-schemas.html

[65] A. Taeihagh, "Governance of artificial intelligence," *Policy Soc.*, vol. 40, no. 2, pp. 137–157, Apr. 2021.

[66] J. Schneider, R. Abraham, C. Meske, and J. V. Brocke, "Artificial intelligence governance for businesses," *Inf. Syst. Manage.*, vol. 40, no. 3, pp. 229–249, Jul. 2023, doi: 10.1080/10580530.2022.2085825.

[67] Eur. Parliament Council Eur. Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27, Apr. 2016 on the Protection of Natural Persons With Regard To the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[68] Ab Initio. (2017). *GDPR Report*. [Online]. Available: https://complexstories.com/wp-content/uploads/2020/06/GDPR-report.pdf

[69] L. de Leyritz. (2023). *How To Comply With GDPR?*. [Online]. Available: https://www.castordoc.com/blog/what-if-you-had-to-comply-with-gdpr

[70] R. L. Sarfin. (2019). *What You Need To Know: GDPR Compliance*. [Online]. Available: https://www.precisely.com/blog/data-quality/what-you-need-to-know-gdpr-compliance

[71] A. Rai, "Explainable AI: From black box to glass box," *J. Acad. Marketing Sci.*, vol. 48, no. 1, pp. 137–141, Jan. 2020, doi: 10.1007/s11747-019-00710-5.

[72] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1–45, Jun. 2024, doi: 10.1145/3641289.

[73] R. Ashmore, R. Calinescu, and C. Paterson, "Assuring the machine learning lifecycle: Desiderata, methods, and challenges," *ACM Comput. Surv.*, vol. 54, no. 5, pp. 111:1–111:39, Jun. 2022, doi: 10.1145/3453444.

[74] Eur. Commission. (2024). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[75] Oracle. *Java SE Technologies: Database*. Accessed: Apr. 30, 2025. [Online]. Available: https://www.oracle.com/java/technologies/javase/javase-tech-database.html

[76] Tableau. *Tableau: Bus. and Analytics Software*. Accessed: Apr. 30, 2025. [Online]. Available: https://www.tableau.com/

[77] Microsoft. *Power BI: Data Visualization*. Accessed: Apr. 30, 2025. [Online]. Available: https://www.microsoft.com/en/power-platform/products/power-bi

[78] Google Cloud. *Looker: Business Intelligence Platform Embedded Analytics*. Accessed: Apr. 30, 2025. [Online]. Available: https://cloud.google.com/looker

[79] Apache Softw. Found. *Apache Superset*. Accessed: Apr. 30, 2025. [Online]. Available: https://superset.apache.org/

[80] Qlik. *Qlik Sense: Modern Analytics*. Accessed: Apr. 30, 2025. [Online]. Available: https://www.qlik.com/us/products/qlik-sense

[81] Apache Softw. Found. *Apache Spark: Unified Engine for Large-scale Data Analytics*. Accessed: Apr. 30, 2025. [Online]. Available: https://spark.apache.org/

[82] Atlassian. *Jira: Issue & Project Tracking Software*. Accessed: Apr. 30, 2025. [Online]. Available: https://www.atlassian.com/software/jira

[83] Slack Technol. *Slack: AI Work Management & Productivity Tools*. Accessed: Apr. 30, 2025. [Online]. Available: https://slack.com/

[84] Great Expectations. *Great Expectations: Have Confidence in Your Data, No Matter What*. Accessed: Apr. 30, 2025. [Online]. Available: https://greatexpectations.io/

[85] Apache Softw. Found. *Apache Kafka*. Accessed: Apr. 30, 2025. [Online]. Available: https://kafka.apache.org/

[86] E. Cesario, F. Folino, G. Manco, and L. Pontieri, "An incremental clustering scheme for duplicate detection in large databases," in *Proc. 9th Int. Database Eng. Appl. Symp. (IDEAS05)*, Montreal, QC, Canada, 2005, pp. 89–95.

[87] L. Ehrlinger and W. Wöß, "A survey of data quality measurement and monitoring tools," *Frontiers Big Data*, vol. 5, p. 3, Mar. 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdata.2022.850611

[88] V. Goasdoué, S. Nugier, D. Duquennoy, and B. Laboisse, "An evaluation framework for data quality tools," in *Proc. 12th Int. Conf. Inf. Quality (ICIQ)*. Cambridge, MA, USA: MIT, 2007, pp. 280–294.

[89] Octopai. (2024). *What is Data Deduplication*. [Online]. Available: https://www.octopai.com/glossary/data-deduplication/

[90] M. Chen. (2024). *What Is Data Deduplication? Methods and Benefits*. [Online]. Available: https://www.oracle.com/data-deduplication/

[91] L. Ehrlinger, C. Lettner, W. Fragner, G. Gsellmann, S. Nestelberger, F. Rauchenzauner, S. Schützeneder, M. Tiefengrabner, and J. Zeindl, "Data integration, management, and quality: From basic research to industrial application," in *Proc. Database Expert Syst. Appl. (DEXA Workshops)*. Cham, Switzerland: Springer, Jan. 2022, pp. 167–178.

[92] L. Ehrlinger. (2023). *People and Data: Why Responsibility Matters and What Data Leaders Should Do About It*. [Online]. Available: https://www.cdomagazine.tech/opinion-analysis/article762772ae-15b4-11ee-a081-d7840f52ccfa.html

[93] C. Wohlin, P. Runeson, M. Hst, M.-C. Ohlsson, B. Regnell, and A. Wessln, *Experimentation in Software Engineering*. Heidelberg, Germany: Springer, 2012.

JASMIN KROPSHOFER received the bachelor's degree in computer science from Johannes Kepler University Linz (JKU), Austria, in 2025. Currently, she is pursuing the bachelor's degree in education and the master's degree in computer science with a major in information systems. Alongside her studies, she has been a Research Assistant with the Institute for Application-Oriented Knowledge Processing (FAW), since 2022. Her research interests include data quality, knowledge graphs, and data catalogs.

JOHANNES SCHROTT received the bachelor's degree in computer science from Johannes Kepler University Linz (JKU), Austria, in 2022, where he is currently pursuing the master's degree in computer science with a major in information systems. From 2021 to 2024, he was a Researcher with the Institute for Application-Oriented Knowledge Processing (FAW). Since 2023, he has been affiliated with the Software Competence Center Hagenberg GmbH. His research interests and publications cover the fields of data catalogs, metadata, ontologies, and data quality.

WOLFRAM WÖß received the Doctorate (Ph.D.) degree from Johannes Kepler University Linz (JKU), Austria, in 1996, and the Habilitation degree in applied computer science, in 2002.

He worked for a manufacturing company, from 1990 to 1993. Since 1993, he has been with Johannes Kepler University Linz (JKU). He has been the Deputy Head of the Institute for Application-Oriented Knowledge Processing (FAW), since 2004. In the course of his tenure at JKU, he has managed both industrial and publicly funded research projects. He has published on these topics in books, scientific journals, and conference proceedings. His research and teaching activities include topics in the field of intelligent information systems, integrated information systems, semantic information integration, ontologies, knowledge graphs, data modeling, data quality, data catalogs, big data, business intelligence, and data mining.

Dr. Wöß is also a member of numerous program committees. He received the Best Paper of the Year 2016 Award from *Data and Knowledge Engineering* (Elsevier) journal. He was the Program Chair of the International Conference on Warehousing and Knowledge Discovery (DaWaK 2003 and 2004) and the Chair of the International Workshop on Web Semantics (WebS), from 2003 to 2013.

LISA EHRLINGER received the Ph.D. degree in computer science from Johannes Kepler University Linz (JKU), Austria, on the topic of automated continuous data quality measurement. She is currently a Senior Researcher with the Hasso Plattner Institute (HPI), University of Potsdam. She has more than 13 years of practical experience in information technology and more than seven years of scientific experience in both fundamental and applied research projects. Her research interests and publications cover the topics of data quality, metadata management and data catalogs, knowledge graphs, and ontologies. She chaired the Quality of Databases (QDB) workshops, in 2023 and 2024, co-located with the VLDB conference, and acts as a Reviewer for international journals, such as ACM CSUR and JDIQ. She was a speaker at the Premier Chief Data Officer and Information Quality (CDOIQ) Symposium in Boston, MA, in 2019, 2020, and 2023. She has been a member of the review board, since 2023, and is also a member of the Global Editorial Board of the CDO Magazine.