




# Data Catalogs: A Systematic Literature Review and Guidelines to Implementation

Lisa Ehrlinger<sup>1,2</sup>(✉) , Johannes Schrott<sup>2</sup>, Martin Melichar<sup>2</sup>,  
Nicolas Kirchmayr<sup>3</sup>, and Wolfram Wöß<sup>2</sup>

<sup>1</sup> Software Competence Center Hagenberg GmbH, Hagenberg, Austria  
[lisa.ehrlinger@scch.at](mailto:lisa.ehrlinger@scch.at)

<sup>2</sup> Johannes Kepler University Linz, Linz, Austria  
{[lisa.ehrlinger](mailto:lisa.ehrlinger@jku.at), [johannes.schrott](mailto:johannes.schrott@jku.at), [wolfram.woess](mailto:wolfram.woess@jku.at)}@jku.at

<sup>3</sup> KTM Innovation GmbH, Wels, Austria  
[nicolas.kirchmayr@ktm.com](mailto:nicolas.kirchmayr@ktm.com)

**Abstract.** In enterprises, data is usually distributed across multiple data sources and stored in heterogeneous formats. The harmonization and integration of data is a prerequisite to leverage it for AI initiatives. Recently, data catalogs pose a promising solution to semantically classify and organize data sources across different environments and to enrich raw data with metadata. Data catalogs therefore allow to create a single, clear, and easy-accessible interface for training and testing computational models. Despite a lively discussion among practitioners, there is little research on data catalogs. In this paper, we systematically review existing literature and answer the following questions: (1) What are the conceptual components of a data catalog? and (2) Which guidelines can be recommended to implement a data catalog? The results benefit practitioners in implementing a data catalog to accelerate any AI initiative and researchers with a compilation of future research directions.

**Keywords:** Data catalog · Data integration · AI system engineering

## 1 Introduction

One of the key challenges of artificial intelligence (AI) system engineering is the integration and harmonization of data to enable high-quality analytics [5]. This paper investigates the extent to which data catalogs can address this challenge. The popularity of data catalogs is continuously increasing since 2016 and they are deemed to be “the new black in data management and analytics” [21] according to Gartner [21]. In 2020, Quimbert et al. [12] define data catalogs as tools to centrally “collect, create, and maintain metadata”, allowing for easier findability and accessibility. Consequently, they do not only bear the potential to (virtually) integrate heterogeneous data sources, but also to semantically enrich data

---

The research in this paper has been funded by BMK, BMDW, and the Province of Upper Austria in the frame of the COMET Programme managed by FFG.

© Springer Nature Switzerland AG 2021

G. Kotsis et al. (Eds.): DEXA 2021 Workshops, CCIS 1479, pp. 148–158, 2021.

[https://doi.org/10.1007/978-3-030-87101-7\\_15](https://doi.org/10.1007/978-3-030-87101-7_15)

with contextual information (i.e., metadata). Metadata is essential to support explainability in AI systems [5].

*Case Study.* The R&D department of motorbike manufacturer KTM, where heterogeneous data (e.g., sensor data from training runs with research prototype bikes) is stored in different formats and granularities. To enable deep insights into bike research and development with AI, KTM aims to deploy a data catalog to deliver high-quality data as basis for data science processes.

*State of the Art.* In recent years, several commercial data catalog tools have been developed, for example, Alation data catalog, Informatica enterprise data catalog, and Oracle cloud infrastructure data catalog [2, 21]. However, despite a vital discussion among practitioners and several commercial tools, there is little research on data catalogs and to the best of our knowledge no other systematic literature review. In 2020, Labadie et al. [9] express the need for further research on data catalogs, specifically with respect to its implementation.

*Contribution.* In this paper, we contribute with a systematic literature review (SLR) on data catalogs to identify (1) necessary and optional conceptual components and (2) guidelines to implement a data catalog. The results offer a consolidated view on what constitutes a data catalog (with respect to its components) and consequently facilitate more research on the topic. For practitioners, this papers provides best practices on how to implement a data catalog.

*Structure.* This paper follows the classic IMRAD structure with Sect. 1 being the Introduction, Sect. 2 describing the research Method, Sect. 3 the Results of our study, and Sect. 4 concludes with a Discussion and future work.

## 2 Research Method

Our systematic literature review is based on Kitchenham [8]. First, we identified the need for a review on the topic of “data catalog”, followed by the development of a review protocol including research questions and search criteria.

### 2.1 Research Questions

The two major aims of this survey are to identify the necessary and optional components of which a data catalog consists and to identify guidelines on how to implement a data catalog. According to these objectives, we formulated the following two research questions:

(RQ1) What are the conceptual components of a data catalog?

(RQ2) Which guidelines can be recommended to implement a data catalog?

### 2.2 Search Strategy

For the literature review, we queried the most common digital libraries as outlined in Table 1. Since literature on the topic of “data catalog” is rare, we added

the term “data cataloging” to our search expression, which describes the process of creating a data catalog [15]. We also included the British and American English spelling for each term. Consequently, the following search expression

(“data catalog”  $\vee$  “data catalogue”  $\vee$  “data cataloguing”  $\vee$  “data cataloging”)

has been applied to the scope of title and abstract, whenever setting the scope was possible. We filtered all papers published before 2000 since according to Gartner [21], data catalogs gained their popularity in 2016 and it continuously increased since then. The exact search expression applied to each of the digital library is shown in Table 1. For Google Scholar, the restriction “-VizieR”<sup>1</sup> was added, since a lot of results about the VizieR data catalog were delivered, which were of no relevance, e.g., information about astronomical data.

**Table 1.** Overview on digital libraries with exact search expressions

Source	Search expression	Scope	Additional restrictions
ACM Digital Library <sup>a</sup>	acmdlTitle: (+ (“data catalog” “data catalogue” “data cataloguing” “data cataloging”)) OR recordAbstract: (+ (“data catalog” “data catalogue” “data cataloguing” “data cataloging”))	Title, abstract	–
Google Scholar <sup>b</sup>	allintitle: “data catalog” OR “data catalogue” OR “data cataloguing” OR “data cataloging”	Title	-VizieR
IEEE Xplore <sup>c</sup>	“data catalog” OR “data catalogue” OR “data cataloguing” OR “data cataloging”	All metadata	–
ResearchGate <sup>d</sup>	“data catalog” OR “data catalogue” OR “data cataloguing” OR “data cataloging”	–	–
Science Direct <sup>e</sup>	“data catalog” OR “data catalogue” OR “data cataloguing” OR “data cataloging”	Title, abstract, keyword	–
Springer Link <sup>f</sup>	“data catalog” OR “data catalogue” OR “data cataloguing” OR “data cataloging”	Full text	Discipline computer science + availability filter

<sup>a</sup><https://dl.acm.org>

<sup>b</sup><https://scholar.google.com>

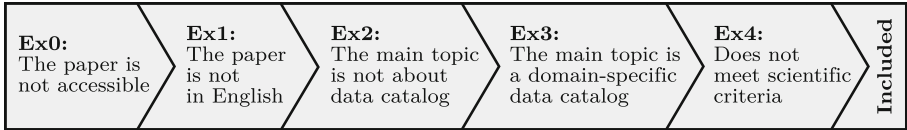
<sup>c</sup><https://ieeexplore.ieee.org/Xplore/home.jsp>

<sup>d</sup><https://www.researchgate.net/search/publication>

<sup>e</sup><https://www.sciencedirect.com>

<sup>f</sup><https://link.springer.com>

<sup>1</sup> VizieR is an online data catalog for astronomical data: <http://vizier.u-strasbg.fr>.



**Fig. 1.** Overview and order of exclusion criteria

**Table 2.** Number of found, excluded, and included publications

Source	No. of papers	Exclusion criteria					Included	
		Ex0	Ex1	Ex2	Ex3	Ex4	All	Uniques
ACM Digital Library	241	1	0	228	10	0	2	
Google Scholar	218	14	14	163	20	0	7	
IEEE Xplore	78	4	0	62	7	0	5	
ResearchGate	100	73	3	14	3	4	3	
Science Direct	54	4	0	46	4	0	0	
Springer Link	468	0 <sup>a</sup>	0	464	3	0	1	
<b>Total</b>	<b>1,159</b>	<b>96</b>	<b>17</b>	<b>977</b>	<b>47</b>	<b>4</b>	<b>18</b>	<b>11</b>

<sup>a</sup>As seen in Table 1, not accessible results have been filtered when querying Springer.

### 2.3 Paper Selection Process

To select papers that are suitable to answer our research questions, we reduced the total number of identified papers with five predefined exclusion criteria (Ex), which were checked sequentially as shown in Fig. 1. All result records that were not removed by any of the exclusion criteria were included in the search result.

## 3 Results from the Literature Review

Across all libraries, 1,159 publications (including duplicates) were found on Feb. 16, 2021 using the search terms from Table 1. Table 2 shows the number of papers excluded and those that were selected to answer our research questions.

Our research questions can be answered based on the content of the eleven papers that remain in the SLR. In addition to the two research questions, Sect. 3.1 provides an overview on the domains in which data catalogs are currently used, compiled from all papers filtered by Ex3.

### 3.1 Overview on Data Catalog Implementations in Practice

From the 47 papers filtered by Ex3, 27 discuss data catalogs that provide open data of various domains. Most papers deal with government data, scientific research data, or geospatial data, but also educational or biological/medical data can be found. Although these systems are called “data catalog”, they follow a

different approach: instead of managing data and its metadata, they provide data of a specific domain to the public. The remaining 20 papers present data catalogs as we understand them, but are limited to a specific application (e.g., a wind park) and do not cover aspects relevant to answer our research questions.

### 3.2 Components of a Data Catalog

None of the investigated papers clearly lists the conceptual parts of a data catalog. Thus, in accordance to Aristotle’s “the whole is greater than the sum of its parts”, we identified the following components as most relevant by investigating all of the eleven papers: (1) metadata management, (2) business context, (3) data responsibility roles, and (4) the FAIR principles. We describe these components and their appearance in the single papers in the following paragraphs.

**Metadata Management.** Data catalogs “collect, create and maintain metadata” [12], which is why, metadata management is the quintessence of a data catalog. Metadata is “data that defines or describes other data” [6], e.g., data quality constraints, usage statistics, or access control [15]. Metadata can be created manually or automatically (e.g., information about data lineage) [15]. While Quimbert et al. [12] classify metadata into three general categories (as originally proposed by Riley [14]), Seshadri and Shanmugam [15] distinguish between eight types of metadata, which can be mapped to the categories as shown in Table 3.

**Table 3.** Classification of metadata by Quimbert et al. and Seshadri and Shanmugam

Metadata categories by Quimbert et al. [12]	Data context variables and data attributes by Seshadri and Shanmugam [15]
<i>Descriptive metadata</i> like title, description, or information about the authors support a user in finding and classifying resources	Despite data quality ratings, a <i>data quality</i> attribute can contain subcategories like data formats or data ranges as an example The reliability of a dataset is represented by <i>reliability</i> attributes
<i>Administrative metadata</i> (also termed “technical information”) like file format, text encoding as well as information about access rights and data provenance	<i>Data lineage</i> represents the dataflow through the entire organization or company <i>Technical context</i> variables provide technical details of a given data set <i>Data sensitivity and accessibility</i> attributes mark sensitive data as such and also provides access restrictions
<i>Structural metadata</i> describes how files or parts of resources relate to each other	<i>Data system relationships</i> context variables hold information about the data origin <i>Data linkage and relationships</i> context variables describe relationships among the data <i>Business context</i> variables represent relationships between data and business domains

To enable the linkage of data across different (heterogeneous) data sources, a metadata schema (also: metadata standard or data documentation) is

required [1, 16], which is defined by “a *set of elements* connected by some *structure*” [13]. For interoperability, also metadata standards from external institutions can be used to enhance a corporate-built metadata schema [12]. In this respect, also data provenance plays a crucial role since it contains information about the source of the data and all transformations it went through [17].

Early approaches to cataloging metadata are often based on XML, e.g., the work by Jensen et al. [7] from 2006, which implements a domain-specific schema based on XML. Since traditional data models are often too less expressive to model the complexity of metadata for a specific domain, ontologies (as the most expressive data model [4]) are recommended by different papers for the implementation of the metadata schema (cf. [2, 12]). There exist several public ontologies, which address specific aspects of the data catalog metadata, e.g., the DCPAC (Data Catalog Provenance, and Access Control) ontology for data lineage and accessibility, which utilizes several other ontologies including DCAT (Data Catalog Vocabulary)<sup>2</sup> and PROV-O (PROV Ontology)<sup>3</sup>, both being W3C recommendations [2]. Other ontologies commonly used for data catalogs are ISO9115, DataCite, Dublin Core Metadata Initiative, CERIF, and schema.org.

**Business Context.** As indicated by [9] and [18], the actual target group of a data catalog are typically business users and not just data or IT specialists. To achieve better workflows and data usage, one of the main foci of building a data catalog lies in the business context of the data. There are two different suggestions how to achieve the implementation of business context: it is either possible to enrich the metadata (cf. Table 3 classification by [15]) with additional business context attributes (cf. [15]), or to choose the more general path by establishing a company-wide *business glossary* [9]. A business glossary can be defined as “a central repository that contains key business terms whose names and definitions have been agreed upon by cross-functional subject matter experts” [20].

**Data Responsibility Roles.** There is a wide agreement that data is only as useful as its quality or reliability [15, 22]. One of the main reasons for poor data quality is the lack of responsibility employees feel they have for a specific data set (i.e., unclear role assignment between IT and domain experts) [22]. Barbosa and Sena [1] go one step further and state that the success of a data catalog depends on the people maintaining it. Thus, one crucial aspect for the implementation of a data catalog is the assignment of responsible persons to the data [9]. Despite the traditional data expert roles (e.g., data architects), which are responsible for modeling the data, new less specialized roles that use the data to reach company goals are assigned in the context of data catalogs [9]. Labadie et al. [9] identify the *data steward* as most important data catalog role for companies. For Kurth et al. [11], establishing responsibility rules, particularly data stewardship, is one of the main tools for successful metadata maintenance and governance.

<sup>2</sup> <https://www.w3.org/TR/vocab-dcat> (Apr. 2021).

<sup>3</sup> <https://www.w3.org/TR/prov-o> (Apr. 2021).

**FAIR Principles.** The FAIR Principles<sup>4</sup> have been proposed in 2016 by Wilkinson et al. [19] and gained recent popularity in the enterprise context through the term “data democratization” [9]. The acronym FAIR stands for Findability, Accessibility, Interopability, and Re-use. Each term represents a category of guiding principles, where each principle defines specific characteristics of the data to fulfill FAIR [19]. The principles are designed to be “concise, domain-independent, high-level” [19] considerations for the publishing of data.

As described in [19], the connection between metadata, data management, and the FAIR principles is tight: each of the principles provides guidelines for desired characteristics of data, metadata, or both of them. Therefore, the quality of data as well as metadata directly affects the fulfillment of the FAIR principles.

The market analysis of data catalogs by Labadie et al. [9] identifies nine different function groups of data catalogs, which implement specific aspects of FAIR. For example, the “data search and tagging” group relate to the “findable” principle, whereas the data “analytics and workflows” group make use of the “accessible” and “reusable” [9]. Due to brevity, we refer to [9] for details on the function groups and the extent to which they address the FAIR principles.

### 3.3 Guidelines to Implement a Data Catalog

From the small number of scientific papers on data catalogs in general, we identified only three papers that were dedicated to implementation suggestions (this lack was already outlined in [9]). Wang [18] point out that the *definition of a metadata schema* is the first necessary step towards implementing a data catalog. A company should decide whether (partly) reusing an existing public metadata schema is possible, and only develop a completely new schema if none is available [18]. Seshadri and Shanmugam [15] recommend the following 8-step solution for implementing a data catalog, where step 1–5 effectively refer to the definition of a metadata schema:

1. Initially, a company/organization *defines data context variables*, which contain data-system relationships, business context, technical context, data lineage as well as linkage information.
2. The second step covers the *definition of data attributes*, which represent the quality, sensitivity, accessibility, and reliability of data.
3. Third, the authors suggest the *tagging of data*, where it is decided which metadata (i.e., data attributes and context variables) is attached to data at a particular level, e.g., column-level, entity-level, or data-set-level.
4. Next, *rules should be defined*, which regulate the data access or audits. For more flexibility, external business rule engines could be used and the rules can also be applied on multiple hierarchy-levels in analogy to the metadata.
5. After the previous steps have been accomplished, the *final data catalog schema* can be assembled into one enterprise data model, i.e., ontology.
6. Eventually, the data *catalog can be populated* with data.

<sup>4</sup> <https://www.go-fair.org/fair-principles> (Apr. 2021).

7. After the catalog is populated, it can be *exposed to the users*.
8. The final and ideally ongoing step is to take all the *feedback, revisions, and reviews* to improve the data catalog.

On a more general level, Labadie et al. [9] distinguish between two different approaches for the creation of a metadata schema: the *top-down approach*, where the structure is defined first, and the data imported in a second step, and the *bottom-up approach*, in which the schema is developed according to the analysis of imported data [9]. In terms of practical implementation, Labadie et al. [9] again distinguish between two contrasting approaches: the *data supply-driven approach* (also input-oriented approach), in which the requirements of the users who will provide and maintain data in the data catalog are prioritized, and the *data demand approach*, where the focus is on the output of the data catalog and prioritizes the requirements of end users who consume data from the data catalog. Three case studies in [9] show the connection between the two modeling approaches (top-down and bottom-up) and the two implementation approaches (data supply-driven and data demand). The top-down approach is typically conducted by users who maintain the data catalog, and therefore combined with the data supply-driven implementation approach, whereas the bottom-up modeling approach first considers the available data as it is used and therefore combined naturally with the data demand approach. It is pointed out that a combination of both sides and an agile iterative approach is also possible [9].

Lee and Sohn [10] propose a semi-automated method to create the metadata schema: the tag-based dynamic data catalog (DaDDCat). With DaDDCat, users are requested to annotate web resources (e.g., web pages, images, videos) with tags (i.e., a set of words) that are then used to automatically built an ontology.

One of the main challenges in the implementation of a data catalog is metadata interoperability across an entire organization. Kurth et al. [11] recommend the following two measures to address this challenge: (1) establish an enterprise-wide consensus on metadata mapping decisions, which prevents duplicate work by different teams, and (2) establish data stewardship to govern the data.

## 4 Discussion and Outlook

In this paper, we performed a SLR to (1) identify the main conceptual components of a data catalog and to (2) provide guidelines for its implementation.

*(RQ1) Main Components.* We answer (RQ1) by compiling the main conceptual components for a data catalog, which are: (1) effective metadata management, (2) the incorporation of business context either to the metadata or as separate business glossary, (3) the assignment of dedicated data responsibility roles, and (4) the adherence to the FAIR principles. We conclude that the major distinction of data catalogs to traditional data management or integration projects is on the one hand the commitment to use ontologies for describing the metadata, and on the other hand, the dedicated incorporation of business users with newly defined roles, such as the data steward.



*(RQ2) Data Catalog Implementation.* Sect. 3 indicates that the definition of a metadata schema (or ontology) is the key challenge in implementing a data catalog. In addition to fitting organizational needs, the metadata schema should fulfill the FAIR principles and adhere to common standards. Interestingly, none of the existing implementation suggestions incorporates the assignment of data responsibility roles. Due to the inherent importance of this conceptual component, we promote the following high-level process to implement a data catalog:

1. Assignment of data responsibility roles to stakeholders that contribute to the definition of the metadata schema or ontology.
2. Definition of a metadata schema (cf. steps 1–5 by [15]).
3. Population of data catalog schema with data (cf. step 6 by [15]).
4. Assignment of data responsibility roles to technical and business users for updates and continuous maintenance of the metadata.
5. Continuous improvement according to revisions and reviews (cf. step 8 by [15]).

We claim that it is necessary to divide the role assignment: in step (1), responsibility roles are assigned for the metadata schema modeling phase, and in step (4), responsibility roles are assigned for the daily use and maintenance of the metadata. Although these role assignments may overlap, they are often disjoint in practice, e.g., IT people are more involved in the data modeling phase, whereas business users without a global view on the data might maintain specific parts of the data on a daily basis.

*Open Issues for Practitioners.* According to Dibowski et al. [2], main data catalog vendors do not support the usage of existing public ontologies, but restrict the use to proprietary metadata schemas. In order to enhance interoperability and adhere to the FAIR principles, existing data catalogs should allow the incorporation of standardized public ontologies, such as DCAT or schema.org.

*Open Issues for Researchers.* In our SLR, we identified the following three topics for future research: (1) automated data catalog creation, (2) data stewardship in data catalog literature, and (3) data quality in data catalogs.

We did not find any attempt to automatically create the metadata schema of a data catalog, which would be specifically interesting with for bottom-up approaches. Most use cases with bottom-up approaches are restricted to the manual analysis of existing data sources [9] and do not address automated schema extraction, as, e.g., suggested in [3]. Barbosa and Sena [1] even state that this step cannot be automated. Considering the high human effort of schema modeling (cf. [9]), we claim that a scientific evaluation of this statement is needed.

As already pointed out in the discussion of (RQ2), current data catalog implementation approaches do not address the topic of data stewardship sufficiently. Considering the importance of the topic for organizational needs as shown in [9], the lack of data stewardship in data catalog literature indicates a gap between real-world business needs and research, which should be closed in future work.

Seshadri and Shanmugam [15] highlight the importance of data quality for data catalog projects. Metadata can be used to determine the quality of data in aggregated metrics. In our ongoing research, we plan to integrate the concept of automated data quality monitoring [3] with tools like DQ-MeeRkat<sup>5</sup> into an existing data catalog implementation at KTM Innovations GmbH.

## References

1. Barbosa, E.B.d.M., Sena, G.d.: Scientific data dissemination a data catalogue to assist research organizations. *Ciência da Informação* **37**, 19–25 (04 2008)
2. Dibowski, H., et al.: Using semantic technologies to manage a data lake: data catalog, provenance and access control, p. 17 (2020)
3. Ehrlinger, L., Wöß, W.: Automated data quality monitoring. In: Talburt, J.R. (ed.) *Proceedings of the 22nd MIT International Conference on Information Quality (ICIQ 2017)*, Little Rock, AR, USA, pp. 15.1–15.9 (2017)
4. Feilmayr, C., Wöß, W.: An analysis of ontologies and their success factors for application to business. *Data Knowl. Eng.* **101**, 1–23 (2016)
5. Fischer, L., et al.: AI system engineering-key challenges and lessons learned. *Mach. Learn. Knowl. Extr.* **3**(1), 56–83 (2021)
6. Data Quality - Part 8: Information and Data Quality Concepts and Measuring. Standard, International Organization for Standardization, Switzerland (2015)
7. Jensen, S., et al.: A hybrid XML-relational grid metadata catalog. In: *International Conference on Parallel Processing Workshops (ICPPW 2006)*, pp. 8–24 (2006)
8. Kitchenham, B.: Procedures for performing systematic reviews, p. 33 (2004)
9. Labadie, C., et al.: Fair enough? Enhancing the usage of enterprise data with data catalogs. In: *2020 IEEE 22nd Conference on Business Informatics (CBI)*, vol. 1, pp. 201–210, June 2020
10. Lee, H.J., Sohn, M.: Construction of tag-based dynamic data catalog (TaDDCaT) using ontology. In: *2012 15th International Conference on Network-Based Information Systems*, pp. 697–702 (2012). <https://doi.org/10.1109/NBiS.2012.116>
11. Martin Kurth, David Ruddy, N.R.: Repurposing MARC metadata: using digital project experience to develop a metadata management design. *Library Hi Tech* **22**(2), 153–165 (2004). <https://doi.org/10.1108/07378830410524585>
12. Quimbert, E., Jeffery, K., Martens, C., Martin, P., Zhao, Z.: Data cataloguing. In: Zhao, Z., Hellström, M. (eds.) *Towards Interoperable Research Infrastructures for Environmental and Earth Sciences*. LNCS, vol. 12003, pp. 140–161. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-52829-4\\_8](https://doi.org/10.1007/978-3-030-52829-4_8)
13. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. *VLDB J.* **10**(4), 334–350 (2001)
14. Riley, J.: Understanding metadata: what is metadata, and what is it for? National Information Standards Organization (NISO) (2017). [https://groups.niso.org/apps/group\\_public/download.php/17446/Understanding%20Met%E2%80%A6](https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Met%E2%80%A6)
15. Shanmugam, S., Seshadri, G.: Aspects of data cataloguing for enterprise data platforms. In: *IEEE 2nd International Conference on Big Data Security on Cloud (Big-DataSecurity)*, *IEEE International Conference on High Performance and Smart Computing (HPSC)*, and *IEEE International Conference on Intelligent Data and Security (IDS)*, pp. 134–139 (2016)

<sup>5</sup> <https://github.com/lisehr/dq-meerkat>.

16. Skopal, T., et al.: Improving findability of open data beyond data catalogs. In: Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services, pp. 413–417. ACM (2019)
17. Vicknair, C.: Research issues in data provenance. In: Proceedings of the 48th Annual Southeast Regional Conference. ACM SE 2010, Association for Computing Machinery, New York (2010). <https://doi.org/10.1145/1900008.1900037>
18. Wang, X.: An analysis of the benefits and issues in the development of an enterprise data catalogue. Master's thesis, School of Information Management, Victoria Business School, Victoria University of Wellington (2014)
19. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**(1), 160018 (2016)
20. Winningham, S.: Knowledge nugget: business glossary vs. data dictionaries (2019). <https://web.stanford.edu/dept/pres-provost/cgi-bin/dg/wordpress/knowledge-nugget-business-glossary-vs-data-dictionaries>
21. Zaidi, E., et al.: Data catalogs are the new black in data management and analytics (2017). <https://www.gartner.com/en/documents/3837968/data-catalogs-are-the-new-black-in-data-management-and-a>
22. Zhu, H., et al.: Data and information quality research: its evolution and future. In: Computing Handbook: Information Systems and Information Technology, pp. 16.1–16.20. Chapman and Hall/CRC, London (2014)