Unlocking Data Quality for Responsible Artificial Intelligence

Dajana Marnika
Gimnazija Vladimira Nazora, Perivoj Vladimira Nazora 3/2, 23000 Zadar, Croatia
Faculty of Informatics and Digital Technologies, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia
dajana.marnika@student.uniri.hr

Maja Gligora Marković
Faculty of Medicine, University of Rijeka
Department of Biomedical Informatics
Braće Branchetta 20, 51000 Rijeka, Croatia
majagm@medri.uniri.hr

Božidar Kovačić
Faculty of Informatics and Digital Technologies, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia
bkovacic@inf.uniri.hr

Abstract. *Artificial intelligence (AI) has been developing and advancing continuously for more than 70 years. The potential of AI is conditioned by the increase in computing power and memory, the progress of algorithms in machine and deep learning, but primarily - the availability of data. A particularly important role in ensuring data quality is played by metadata, which provide contextual information about the data, such as source, structure, creation time, format, and ownership. The aim of this paper is to investigate which types of metadata represent data quality parameters in intelligent systems and AI. The research is based on a theoretical analysis of relevant scientific literature indexed in the Web of Science (WOS) and SCOPUS databases, with the goal of identifying key metadata attributes that influence the quality and usability of data in AI contexts. This approach allows for a critical synthesis of existing knowledge and insights into how metadata contribute to improved outcomes in AI model development and application.*

Keywords. data, metadata, data quality, artificial intelligence

1 Introduction

Machines created by humans have long been capable of performing various types of more difficult and laborious human tasks. However, with the aim of greater productivity and greater ability to perform tasks, humans have been trying to inject human intelligence into machines, which is the original motive for AI. (Jiang et al., 2022)

There are many definitions of the term AI itself. Not so long ago in 1950, Alan Turing posed the philosophical question "Can machines think?" in his essay Computing Machinery and Intelligence, laying the foundations for the further development of computers. The pioneering definition of AI was devised by John McCarthy in 1955, stating that every aspect of learning or any other feature of intelligence can in principle be described so precisely that a machine can be built to simulate it. According to him, artificial intelligence can be broken down into algorithmic processes, i.e. a human precisely defines intelligent behavior, and a machine (AI) does not imitate a human but behaves as if it had intelligence and solves a given problem. A decade later, a similar definition was given by Marvin Minsky (Minsky, n.d.), according to whom AI is the science of how to make machines do things that would require intelligence if they were to be done by humans. Playing chess was the first practical application of AI. Initial forecasts predicted the rapid growth of artificial intelligence comparable to human intelligence, but the complexity of this process soon became apparent. (Leksikografski zavod Miroslav Krleža, n.d.)

With the development of machine learning, the focus has shifted from logical reasoning to statistics and data (Šuman, 2021), meaning that the goal was no longer for a machine to think like a human, but to learn from the data it receives. The definition has expanded over the years to include optimization of rational action, creativity, adaptation, social influence, and ethical use. It is expected that an AI system will be able to achieve human performance and be rational in doing the 'right thing' (Russell & Norvig, 2010) given the available data.

Because of this data dependency, artificial intelligence is inextricably linked to the development and spread of the phenomenon known as Big Data. The term Big Data (Cai & Zhu, 2015; Abdou, 2020) refers to data that is so large and complex that it exceeds the processing capabilities of traditional data management systems and software. The volume and diversity of data on which AI systems learn poses new challenges, not only in processing and storage, but also in the protection, interpretation, and ethical management of that data.

Doug Laney (Laney, 2001) first introduced the concept and described it with 3V characteristics: Volume, Velocity and Variety. With the development of technology and the growth of complexity of data systems, this model has been expanded and now includes 56V (Abdou, 2020) that reflect the comprehensiveness of modern data environments.

Following this development, the Organisation for Economic Co-operation and Development (OECD) (Grobelnik, 2024) defines artificial intelligence as a machine system that, for explicit or implicit purposes, infers from the input data it receives how to generate outputs such as predictions, content, recommendations, or decisions that can affect physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptability after implementation.

The OECD definition of AI reveals an important link between the definition of AI and the practical challenges in its application, particularly in the context of the importance of data quality management. This link is illustrated in Table 1.

Table 1. Relationship between elements of the OECD definition and the importance of data accuracy

Definition element	Importance of data
... from the input it receives it concludes ...	If the input data is incorrect, the system's inference may be incorrect or incomplete. Inconsistent data confuses models and makes it difficult to learn patterns.
... how to generate outputs such as predictions, content, recommendations or decisions ...	Inaccurate and inconsistent data reduce the reliability of system outputs. They can lead to poor or harmful decisions, especially in high-risk systems.
... can affect physical or virtual environments ...	The consequences of inaccurate, incomplete, and inconsistent data are becoming real and measurable – affecting people and systems.
... in levels of autonomy and adaptability ...	If trained on poor-quality data, adaptive systems can adopt incorrect or contradictory patterns of behavior. By adopting errors, they can reinforce them over time.”

The importance of data in information management has been recognized for decades, and it is precisely their massiveness, diversity, and accessibility that have enabled the development and training of advanced AI models in complex, real-world environments. Recent advances in AI have further emphasized the role of data quality in building sustainable *data ecosystems* (Mohammed et el., 2025). However, metadata management and quality parameters are still not systematically regulated or consistently applied in practice.

This paper consists of five parts: the first, introductory part with explanations of basic terms, the second part, which presents the methodology of the conducted research, the third and fourth parts represent a discussion based on selected reviewed papers in order to answer the research questions, while the last part presents the conclusions of the conducted research and future research.

2 Research methodology

The existing literature deals with the topics of data quality and metadata management in a fragmented manner, therefore there is a need for a systematic insight into the challenges, concepts and existing approaches related to data management, data quality and metadata within the context of AI systems. This research (Kitchenham & Charters, 2007) used a literature review in three phases: planning the review, conducting the review and discussing the studies. Through these phases, the results of the work will be presented.

Following the problem of lack of standardization and methodological uniformity that makes it difficult to evaluate the quality of datasets (which can have direct consequences on the performance and reliability of models), we posed the central research question in this article: What are the key metadata that define data quality in machine and deep learning?

In order to answer the question posed, additional research questions were also posed: 1. How is data quality in intelligent systems defined by academic papers and regulatory frameworks? and 2. What are the best practices for collecting, storing, and tracking metadata related to data quality in intelligent systems?

In the preliminary literature search, we used Scopus and Web of Science as primary scientific databases. We supplemented the publications by analyzing references and citations, with an additional search in Google Scholar. Official documents of international institutions - the European Commission and the OECD - were also used.

Thematic areas covered by the queries are: artificial intelligence, AI system lifecycle, data and metadata management, data quality, and ethical and regulatory aspects of AI. The keywords and combinations used in the search were:

- "data quality" OR "data accuracy" OR "data completeness" OR "data consistency" OR "data integrity" OR "data reliability" OR "data veracity"
- "metadata" OR "metadata management" OR "metadata standard*" OR "metadata quality" OR "semantic metadata" OR "provenance metadata"
- "data management" OR "data governance" OR "data stewardship" OR "data curation" OR "data lifecycle" OR "FAIR data" OR "data trustworthiness"
- "artificial intelligence" OR "AI" OR "machine learning" OR "ML" OR "deep learning" OR "neural network*" OR "AI system*" OR "AI model*".

By using the OR operator, we tried to include synonyms and related terms within each query, while by connecting them with the AND operator, we ensured that only papers were found that cover all four key areas, thus filtering out unnecessary results. The number of papers obtained by this search is 57 in the Scopus database and 20 in WOS. With the aim of further narrowing the search, we defined the criteria for inclusion and exclusion of papers for review as shown in Table 2.

Table 2. Inclusion and exclusion criteria for papers

Switching on	Shutdown
Publications between 2013 and 2025.	Publications older than 2013.
English language	They are not in English.
Paper or conference paper	Other types of documents (e.g. books)
Subject area: computer science	Other subject areas
Open access	Open access is not available.

The mentioned criteria reduced the number of papers to 28 in the Scopus database and 7 in WOS. By reviewing the titles and authors of the papers, duplication was observed, i.e. all 7 papers published in WOS were also published in Scopus, which resulted in a total number of papers for review, which is 28. The further selection process was carried out on the basis of the titles and abstracts, and irrelevant papers were excluded, which brought the number of papers to 23. Potentially relevant papers were then analyzed in detail, which reduced their number to 11.

In order to ensure methodological consistency, a formal protocol was developed that included: definition of key terms, database search strategies, list of used and excluded sources with associated justification, and description of the selection methodology. The protocol was formulated before the search itself began and used as a reference frame throughout the process. If necessary, it was subjected to minimal adaptations with clearly documented changes, in accordance with the recommendations of good practice in conducting systematic reviews.

3 Challenges in data preparation and quality

Data is a key aspect of achieving technological progress, but its use is hampered by the wide range of data formats, the lack of interoperability between tools, and the difficulty of discovering and combining data sets (Akhtar et al., 2024). In all three main phases of AI development – design, development, and production – data quality is the foundation of every step (Daswin & Alahakoon, 2022). Contradictory and inconsistent data can seriously undermine the reliability and effectiveness of models, directly affecting their ability to act responsibly and transparently. In order for AI systems to meet these requirements, it is crucial to ensure the consistency, clarity, and accuracy of data throughout the development lifecycle. In an attempt to provide a more comprehensive framework for solving AI tasks, the more precisely defined phases are: dataset selection, preprocessing, feature engineering, and deployment, with metadata recorded throughout all phases (Venkataramanan et al., 2024).

In order to address the challenges of data preparation and quality – the European Commission (2019) has adopted a framework for trustworthy artificial intelligence that includes seven key requirements: 1. human agency and oversight; 2. technical robustness and security; 3. privacy and data management; 4. transparency; 5. diversity, non-discrimination and fairness; 6. social and environmental well-being; and 7. accountability. In achieving the above requirements, especially those related to aspects of data management, transparency and accountability – metadata plays a key role. They enable documentation of the origin, context

and transformations of data, thus becoming an indispensable tool for ensuring the verifiability and monitoring of model's decisions. Without adequate metadata management, it is difficult to achieve full transparency in the operation of models and ensure responsible development and application of AI systems.

Following table 3 (based on (Gröge, 2022)) presents the European Commission's requirements for the reliable application of AI systematized through three basic categories of challenges: data management in the context of artificial intelligence, data governance, and democratization of data access. These categories are also conceptually linked to the Commission's requirements. The table is based on the model from (Gröge, 2022), but is supplemented with additional categories to more precisely capture the challenges arising from the application of AI systems in a real-world context.

Table 3. Data challenges of artificial intelligence

The challenge of data management in the context of artificial intelligence	**The challenge of data management**	**The challenge of democratizing data for AI**
1. Data modeling 2. Metadata management 3. Data architecture 4. Data quality 5. Integration of data from different sources 6. Methodology for managing text, images, video and sensor data	1. Data ownership 2. Data management 3. Privacy and data protection 4. Ethical use of data and algorithms 5. Management mechanisms for the life cycle of AI models	1. Data security 2. Data Engineering 3. Data discovery and exploration 4. User training 5. Access control 6. Personalization and contextualization of data access

The above challenges in data management are deeply interconnected and together form the basis for the successful development of artificial intelligence. Data quality can be described as *fitness for use* (Corte et al., 2024), where it depends on proper modeling, integration and clearly defined metadata. Assessing the level of data quality is crucial for deciding on its suitability in the process of making accurate and reliable decisions. The data architecture must enable the efficient management of all these components. The processing of different types of data (text, images, sensors) adds particular complexity, which requires harmonized methodologies that support all the previous aspects in a single system. The challenges listed in the second category form the regulatory and ethical framework for the responsible use of data and artificial intelligence. Metadata should not be seen only as technical additions to content (Zhan & Hai, 2024), they can in themselves threaten privacy if shared without appropriate protection. This raises important questions about how to design secure data exchange systems. Ultimately, the challenges listed in the third category form the operational and technical basis of data management in AI systems, and their harmonized connectivity enables data access that is efficient, secure, and user-friendly.

The FAIR principles (Findability, Accessibility, Interoperability and Reusability) (Jackson et al., 2024) should be the fundamental framework for metadata management. By using clear identifiers and descriptive labels, metadata must be findable and accessible through standardized protocols. They should be interoperable and reusable. Such compliance directly contributes to greater transparency, accountability and trustworthiness of AI systems, enabling not only the understanding of model outputs, but also the reuse and evaluation of the data on which the model was trained. Without the systematic application of the FAIR principles through metadata, it is difficult to imagine the realization of the principles of responsible and ethical artificial intelligence advocated by the European Commission.

Summarizing the dimensions of data quality, four dimensions can be distinguished – intrinsic, contextual, representational and accessibility dimensions. The intrinsic dimension can be assessed by measuring the internal attributes or characteristics of the data based on given references, and also measures missing values and redundant cases. The contextual dimension ensures that the data is aligned with the needs and objectives of the projects. The representational dimension assesses the formats and structures of the data, for example whether the data is concisely and consistently presented, but also interpretable. The accessibility dimension assesses the extent to which all or part of the data can be retrieved, and additionally allows users to use and share the data with security controls.

The lack of standardized metrics and overlapping dimensions stem from the variability and complexity of data – they are a challenge in data management. Defining standards is essential to ensure consistency, quality and reliability of data throughout the model lifecycle.

In the pursuit of standards, one solution is to introduce datasheets for datasets. The idea is that each dataset would be accompanied by a datasheet that documents its motivation, composition, collection process, recommended uses, etc. Their use would increase transparency and accountability within the machine learning community, mitigate unwanted social biases in machine learning models, facilitate greater

reproducibility of machine learning results, and help researchers and practitioners select more appropriate datasets for their chosen tasks. (Gebru et al., 2018) Ultimately, the adoption of dataset datasheets would foster a more responsible and ethical approach to machine learning research, ensuring that the datasets we use align with our values and objectives.

All the aforementioned dimensions and challenges in data management indicate that systematic metadata management is a key prerequisite for high-quality, transparent, and responsible use of AI systems. The implementation of standards such as the FAIR principles, the adoption of datasheets for datasets, and compliance with regulatory and ethical frameworks enable not only reproducibility and findability of data but also greater reliability and ethical behavior of algorithms themselves. This clearly links the need for strategic planning and data quality control, forming the foundation for the responsible deployment of AI, which is the main thesis of this work.

4 Metadata Management in Data Lakes: Ensuring Quality and Interoperability for AI Models

For a datasheet to be truly useful, it must contain clearly defined metadata. Their importance grows as the amount of data grows. It stems from the fact that AI models depend on the data from which they learn and on which they make decisions. Therefore – the data, i.e. its origin, quality, structure and meaning, must be precisely recorded and understood. In this context, the concept of a data lake emerges as a solution to the problems of big data heterogeneity – storing data in a way that prevents the lake from turning into a swamp, i.e. into an unusable data lake. (Sawadogo Pegdwendé et al., 2019)

The authors offer various strategies for metadata management. The underlying thesis is that metadata storage should be part of an organization's broader metadata management strategy, which in turn should be part of an organization's broader data management policy (Sheldon, 2023).

The characteristics of metadata quality, according to (Hillman & Thomas, 2004) are: completeness, accuracy, provenance, compliance with expectations, logical consistency and coherence, timeliness, and availability. Furthermore, it should be possible (although for most data it is not) to measure their attributes – semantic and syntactic structure, as well as the data values themselves – by automated systems (Hillman & Thomas, 2004).

Building upon this challenge, the FAIR principles (Findability, Accessibility, Interoperability, Reusability) provide a conceptual framework for assessing metadata quality, while the AIDRIN (Hiniduma et. al., 2024) framework translates these principles into measurable indicators. For instance, Findability can be quantified by metadata completeness, whereas Interoperability can be assessed by consistency and standardization across datasets. In this way, AIDRIN enables concrete, quantitative evaluation of the degree to which datasets adhere to FAIR principles, bridging the gap between abstract guidelines and practical implementation.

Semantic metadata enrichment is also highlighted by (Sawadogo Pegdwendé et al., 2019) as the first of six core functionalities that a metadata system within a data lake should provide. Semantic enrichment should enable the assignment of contextual tags to data using ontologies, thereby increasing their comprehensibility and revealing potential relationships between data sets. Other functionalities are: data indexing, link generation and preservation, polymorphism, versioning, and data tracking (Sawadogo Pegdwendé et al., 2019). Data indexing provides efficient access to data through structured indexes, which is particularly useful for managing textual and semi-structured data. Link generation and preservation identifies similarities and existing relationships between data sets and enables the discovery of clusters of related data. Data polymorphism refers to the storage of multiple versions of the same data in different formats, which facilitates analysis and avoids multiple preprocessing. Data versioning enables changes to be tracked, previous states to be preserved, and analysis to be repeated. Finally, usage tracking records user interactions with data, which contributes to security, monitoring, and anomaly detection.

In order to make certain data lake model as efficient as possible and to achieve the previously mentioned features, metadata should first be divided into three groups: intra-object, inter-object and global (Sawadogo Pegdwendé et al., 2019). Intra-object metadata is obtained from the file system: object title, size, last modified date, access path, etc., i.e. it is associated with a specific object. Inter-object metadata explains the relationships between at least two objects, where each object can simultaneously belong to multiple groups. Such groups can be automatically derived from semantic metadata such as tags and business categories. Global metadata potentially refers to the entire data lake.

To ensure consistent application of these principles in practice, it is recommended to establish a specialized team responsible for metadata management and storage, and to establish a centralized repository that allows easier monitoring and access to information or (primary) data. (Sheldon, 2023)

The Common Metadata Framework (CMF) proposes a metadata system that would record metadata for all stages and datasets, enabling the search for the optimal execution path (Koomthanam et al., 2024). One of the proposed solutions is the collection of metadata from open platforms such as Papers-with-Code, OpenML and

Hugging Face (Venkataramanan et al., 2024), although the proponents themselves see how integrating and unifying different terminologies and data formats from these diverse sources is a challenge. In this context, the Croissant method (Akhtar et al., 2024) offers a standardized, layered metadata format enabling consistent recording of basic data, resources, structure and semantic information in a way that promotes interoperability and easier unification of data from different sources. The Croissant represents a practical solution to overcome the problem of heterogeneity of formats and terminologies, facilitating the automated integration and search of metadata in AI ecosystems.

The added value of metadata exploitation is also demonstrated by the ProbSAP system (Wang et al., 2023), which uses metadata clustering to address the problems of unbalanced and multidimensional data sets in the context of predictive analytics. By integrating scalable metadata-based data clustering and an optimized predictive model (XGBoost), ProbSAP demonstrates how proper processing and structuring of metadata can directly contribute to increasing the accuracy and reliability of analytical results. Such an approach confirms the importance of metadata not only as a technical description of data, but also as an active component in improving the performance of intelligent decision-making systems.

These conceptual frameworks can be further illustrated with examples from different research domains, demonstrating how metadata are directly linked to data quality. Leipzig et al. (2020) emphasize that standards such as DICOM, EML, MIAME and CodeMeta underpin reproducible computational research by ensuring data provenance, context and structural integrity. Řezník et al. (2022) focus on geo(metadata) and repositories, demonstrating through examples such as OGC CSW, open data portals, Schema.org and Linked Open Data how proper documentation and semantic annotation directly improve data findability and reusability. Grant et al. (2024) show that spatial metadata, when integrated into machine learning models, enhance the objectivity and accuracy of geological drill core quantification, thus revealing a direct link between metadata and the reliability of analytical methods. Taken together, these contributions highlight that across domains metadata are consistently recognized as a cornerstone of data quality – whether through reproducibility, findability, or the precision of analytical processing.

It is the responsibility of the data author not only to enable the reuse of large-scale metadata but also to ensure its reproducibility. Emphasis must be placed on input data, as they carry most metadata standards, and on descriptive standards (metadata) to provide context, provenance, authenticity, and the data lifecycle. Sandve et al. (2013) identified the most common sources of reproducibility failures: lack of workflow frameworks, missing platform and software dependencies, manual data manipulation or web-based steps, lack of versioning, absence of intermediate or plotted data, and insufficient literate programming or context that can derail a reproducible analysis.

Despite the progress in the development of tools and standards, there is still no universal solution that would allow compiling a list of specific quality assurance techniques that would be applicable across a wide range of domains and data types. The large volume and different types of data are fundamental difficulties. Quality criteria (Gebru et al., 2018) must be considered based on the specific tasks of AI model development or different stages of the development process. Instead of a single comprehensive solution, the creation of a program roadmap for project managers has been proposed (Hillman & Thomas, 2004) but the time required and operational costs are also major constraints for selecting suitable roadmaps (Gebru et al., 2018). This challenge is further complicated by the fact that quality requirements vary not only across industries but also within different stages of the AI system lifecycle. Therefore, approaches must be flexible and context-specific rather than universally predefined.

5 Conclusion

Today, in a digital environment enriched by AI and Big Data, metadata is becoming an indispensable tool for managing complex data systems. With the exponential growth of data volume, AI systems increasingly depend on clearly defined metadata that enables understanding of the origin, structure, and quality of data. Without metadata, it is difficult to ensure the reliability, transparency, and ethical application of AI models in dynamic and distributed environments.

Theoretical analyses and applied studies confirm that metadata are essential for reproducibility, findability, and precise analytical outcomes across different domains. Moreover, the responsibility of data authors to ensure proper metadata management is critical for reproducibility and trustworthy AI outcomes.

With the sentence "An algorithm is only as good as the data it works with." (Corte et al., 2024) we can summarize the views presented in this work. Without reliable, representative and quality data, even the most advanced algorithms cannot deliver reliable results. We can conclude that adopting a strategic plan that includes analyzing the data infrastructure, understanding the methods and sources of data collection, and creating a plan for data cleaning is a fundamental step before launching an AI model.

The implementation of ethical standards and enforcement of data protection laws on a global scale

should be a fundamental demand of the world's leading statesmen and governments. In order to ensure the responsible use of AI, i.e. privacy and user trust in digital ecosystems, it is necessary to review it.

Systematic metadata management is a key prerequisite for quality data management, easier searching and finding of resources, and making informed decisions in conditions of exponential growth and complexity of data. Future research by the same author(s) will include a systematic review of existing data quality management models, with data storage as a fundamental activity in working with various AI models.

Acknowledgments

The work was created as part of the University of Rijeka project entitled “Learning analytics in e-learning systems based on interactive data visualization assisted by data mining”. The project ID is uniri-iskusni-drustv-23-270.

References

Abou-El-Ela Abdou, H. (2020.). Fifty-six big data V'scharacteristics and proposed strategies to overcome security and privacy challenges (BD2). *Journal of Information Security*, 11(4): 304–328.doi:10.4236/jis.2020.114019

Akhtar, M., Benjelloun, O., Conforti, C., Foschini, L., Gijsbers, P., Giner-Miguelez, J., … Zhang, L. (2024.). Croissant: A metadata format for ML-ready datasets. Workshop on Data Management forn End-to-End Machine Learning (DEEM 24), Santiago, Chile. ACM, NewYork, NY, USA, doi: 10.1145/3650203.3663326

Cai, L., & Zhu, Y. (2015.). The challenges of data quality and data quality assessment in the big data era. Data Science Journal, 14(2): 1–10.doi: 10.5334/dsj-2015-002

Corte, C., Sanz, C., Etcheverry, L., & Marotta, A. (2024.). Data quality management for responsible AI in data lakes. In Proceedings of the VLDB 2024 Workshop on Trust and Data Accountability (TaDA 2024). VLDB Endowment. Retrieved from https://www.vldb.org/workshops/2024/proceedings/TaDA/TaDA.13.pdf

Daswin, D. S., & Alahakoon, D. (2022.). An artificial intelligence life cycle: From conception to production. Patterns, 3(6), Article 100489.doi: 10.1016/j.patter.2022.100489

European Commission, High-Level Expert Group on Artificial Intelligence. (2019). Ethics and guidelines for trustworthy AI. Retrieved from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé, H., & Crawford, K. (2018.). Datasheets for datasets. arXiv. https://arxiv.org/abs/1803.09010

Grant, L. J. C., Massot-Campos, M., Coggon, R. M., Thornton, B., Rotondo, F. C., Harris, M., Evans, A. D., & Teagle, D. A. H. (2024). Leveraging spatial metadata in machine learning for improved objective quantification of geological drill core. Earth and Space Science, 11(3), e2023EA003220. doi: 10.1029/2023EA003220

Gröge, C. (2022.). Industry experiences on the data challenges of AI and the call for a data ecosystem for industrial enterprises [Lecture]. Retrieved from https://christophgroeger.de/download/Groeger_There_Is_No_AI_Without_Data.pdf

Grobelnik, M., Perset, K., & Russell, S. (2024.). What is AI? Can you make a clear distinction between AI and non-AI systems? OECD.AI. Retrieved from https://oecd.ai/en/wonk/definition

Hillmann, D., Thomas, B. (2004.). Metadata in Practice. Chicago: ALA Editions.

Hiniduma, K., Byna, S., Bez J. L., Madduri, R. (2024.).AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI. SSDBM '24: Proceedings of the 36th International Conference on Scientific and Statistical Database Management Article No.: 7, Pages 1 – 12. doi: 10.1145/3676288.3676296

Jackson, S., Khan, S., Cummings, N., Hodson, J., & de Witt, S. (2024.). FAIR-MAST: A fusion device data management system. SoftwareX, 27: Article 101869. doi: 10.1016/j.softx.2024.101869

Jiang, Y., Li, X., Luo, H., Yin, S., & Kaynak, O. (2022.). Quo vadis artificial intelligence. Discover Artificial Intelligence, 2(4): 1–19. doi: 10.1007/s44163-022-00022-8

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering (EBSE Technical Report EBSE-2007-01). Keele University.

Koomthanam, A. J., Tripathy, A., Serebryakov, S., Nayak, G., Foltin, M., & Bhattacharya, S. (2024.). Common metadata framework: Integrated for

trustworthy artificial intelligence pipelines. IEEE Internet Computing, 28(3): 37–44. doi: 10.1109/MIC.2024.3377170

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, 6. Retrieved from https://studylib.net/doc/8647594/3d-data-management--controlling-data-volume--velocity--an...

Leipzig, J., Nüst, D., Hoyt, C. T., Soiland-Reyes, S., Ram, K., & Greenberg, J. (2021). The role of metadata in reproducible computational research. Patterns, 2(9), 100322. doi: 10.1016/j.patter.2021.100322

Leksikografski zavod Miroslav Krleža. Umjetna inteligencija. (n.d.). In Hrvatska enciklopedija. Retrieved from https://www.enciklopedija.hr/clanak/umjetna-inteligencija

Minsky, M. (n.d.). Marvin Minsky. In Encyclopaedia Britannica. Retrieved from https://www.britannica.com/biography/Marvin-Minsky

Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2022.). The effects of data quality on machine learning performance on tabular data. Information Systems, 132: Article 102549. doi: 10.1016/j.is.2025.102549

Řezník, T., Raes, L., Stott, A., De Lathouwer, B., Perego,

A., Charvát, K., & Kafka, Š. (2022). Improving the documentation and findability of data services and repositories: A review of (meta)data management approaches. Computers & Geosciences, 169, 105194. doi: 10.1016/j.cageo.2022.105194

Russel, S. J., & Norvig, P. (2010). Artificial intelligence: A modern approach (3rd ed.). Pearson Sandve, G. K., Nekrutenko, A., Taylor, J., Hovig, E. (2013.). Ten Simple Rules for Reproducible Computational Research. PLoS Comput Biol 9(10): e1003285.doi: 10.1371/journal.pcbi.1003285

Sawadogo, P. N., Scholly, É., Favre, C., Ferey, É., Loudcher, S., & Darmont, J. (2019.). Metadata systems for data lakes: Models and features. In: Welzer, T., et al. New Trends in Databases and Information Systems. ADBIS 2019. Communications in Computer and Information Science, vol 1064. Springer, Cham.doi: 10.1007/978-3-030-30278-8_43

Sheldon, R. (2023.). 6 best practices for metadata storage and management. TechTarget. Retrieved from https://www.techtarget.com/searchstorage/feature/6-best-practices-for-metadata-storage-and-management

Šuman, S. (2021). Pregled metoda obrade prirodnih jezika i strojnog prevođenja. Zbornik Veleučilišta u Rijeci, 9(1), 137–152. Retrieved from https://hrcak.srce.hr/file/377721

Venkataramanan, R., Tripathy, A. K., Serebryakov, S., Annmary, J., Shah, A., Bhattacharya, S., Foltin, M., Faraboschi, P., Roy, K., & Sheth, A. (2024.). Constructing a metadata knowledge graph as an atlas for demystifying AI pipeline optimization. Front. Big Data, 7: Article 1476506.doi: 10.3389/fdata.2024.1476506

Wang, X. N., Zhao, Y. B., Li, C., & Ren, P. (2023.). ProbSAP: A comprehensive and high-performance system for student academic performance prediction. Pattern Recognition, 137: Article 109309.doi: 10.1016/j.patter.2022.100489

Zhan, D., & Hai, R. (2024.). Will sharing metadata leak privacy? IEEE 40th International Conference on Data Engineering Workshops (ICDEW), Utrecht, Netherlands, 2024, 317–323. doi: 10.1109/ICDEW61823.2024.00047