Metadata Quality Dimensions for Big Data Use Cases

Widad Elouataoui[1][a], Imane El Alaoui[2][b] and Youssef Gahi[1][c]
[1] *Laboratoire des Sciences de l'Ingénieur, University of Ibn Tofail, Kénitra, Morocco*
[2] *Laboratoire des Systèmes de Télécommunications et Ingénierie de la Décision, University of Ibn Tofail, Kénitra, Morocco*

Keywords: Metadata Quality, Big Data, Metadata Management, Metadata Quality Dimensions, Machine Learning.

Abstract: Even though more than two decades have elapsed since the Big data big bang, there are still big data problems that are not addressed. Indeed, the emergence of big data has raised several challenges related to data analysis, data quality, and all activities involving data processing. These challenges have also affected metadata management as a crucial component of big data ecosystems. Thus, many metadata management approaches for big data environments have been suggested recently. However, a metadata management policy could not be effective if the metadata is of low quality. Metadata quality directly impacts data quality and, thus, the reliability of data analysis. Moreover, enhancing metadata quality in big data environments allows optimizing data processing time and cost, one of the biggest concerns of data managers. Thus, this paper aims to contribute to the ongoing discussion regarding metadata quality enhancement. Therefore, we highlight the metadata quality issues raised by the particular characteristics of big data. We then provide recommendations to overcome the presented quality issues and conclude with some possible future work directions.

1 INTRODUCTION

Nowadays, data is considered one of the most valuable resources companies could rely on to improve their business processes. Data analytics allows companies to collect practical knowledge about their customers and their business strategy and supports their decision-making. Achieving such objectives requires a robust data management policy. However, a data management policy could not be effective if it does not include metadata management since metadata is crucial. Indeed, managing metadata is highly required for a complete understanding of data content. Metadata could be defined as data about data (Immonen et al., 2015). It summarizes all relevant information about data such as quality, resource type, provenance, and other technical details. However, metadata could not be used efficiently if it is not well structured and of high quality. Indeed, the gathered metadata may contain errors such as empty fields, duplicated records, or inconsistent values. Therefore, ensuring metadata quality has received much attention from companies and researchers. Also, multiple approaches have been suggested in the literature to assess and improve metadata quality (Bruce and Hillmann, 2004), (Király, 2017).

With the emergence of big data, new challenges related to metadata and data management have been raised. Managing big data involves handling many unstructured and heterogeneous data that traditional data management tools could not process. Thus, new data analytics techniques and processing approaches have been introduced to manage big data (El Alaoui et al., 2019a), (Alaoui and Gahi, 2019). Also, big data requires a special treatment related to its particular characteristics, also known as big data 7V's (Kapil et al., 2016), (Alaoui et al., 2019b) (Table 1).

[a] https://orcid.org/0000-0002-2968-2389
[b] https://orcid.org/0000-0003-4428-0000
[c] https://orcid.org/0000-0001-8010-9206

Elouataoui, W., El Alaoui, I. and Gahi, Y.
Metadata Quality Dimensions for Big Data Use Cases.
DOI: 10.5220/0010737400003101
In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML 2021)*, pages 488-495
ISBN: 978-989-758-559-3

Table 1: Big Data Characteristics.

Characteristic	Meaning
Volume	It is an essential characteristic that refers to the vast amount of data generated and processed in big data systems.
Variety	Big data incorporates heterogeneous data sources and includes different data types that could be structured, semi-structured or unstructured.
Velocity	Refers to the high speed at which data is generated and processed.
Veracity	Refers to the accuracy and the truthfulness of the generated data.
Value	Refers to the insights and the business value that data provide.
Variability	Sheds light on the dynamic aspect of data. It refers to the frequency at which the meaning of data changes.
Visualization	It refers to the process of presenting the extracted information in a readable form, more comfortable to understand.

The collected metadata in big data environments is usually not well structured, inconsistent, and low quality because of big data characteristics. Such metadata is useless and may even bias data analytics results. Hence, enhancing metadata quality in big data environments is of a high priority and should be considered by data owners.

Therefore, this paper aims to shed light on the metadata quality challenges raised by the emergence of big data, a less discussed issue in the existing literature. Also, some solutions to overcome these challenges will be suggested.

This paper is divided into four parts: Section 2 describes the metadata quality dimensions introduced in the literature. Section 3 reviews all related works that have tackled metadata quality issues and metadata management in the big data era. Section 4 highlights the metadata quality issues raised by the emergence of big data and proposes some recommendations deal with the problems presented. Finally, conclusions will be made, and future research directions will be discussed.

2 METADATA QUALITY DIMENSIONS

Metadata quality could not be addressed without considering its main characteristics, also known as Metadata Quality Dimensions. Metadata quality dimensions were first defined by Bruce and Hillmann that have defined seven quality dimensions: accuracy, consistency, completeness, conformance to expectations, timeliness, accessibility, and provenance. Later, other studies have extended these basic dimensions to include more metrics such as shareability, extendibility, and versionability [8] [9] [10]. Table II describes the most common metadata quality dimensions discussed in the literature that we grouped into four aspects: Usability, Reliability, Availability, and Interoperability.

Table 2: Metadata Quality Dimensions.

Quality Aspect	Metadata Quality Dimension	Meaning
Usability	Completeness	Assures that there are no missing attributes and all the expected attributes have values.
	Consistency	The extent to which metadata is coherent and aligned with metadata standards and schemas.
	Usefulness	The extent to which metadata is valuable and relevant for its intended use.
Reliability	Accuracy	The degree to which metadata is correct and reliable.
	Timeliness	Refers to how recent and up-to-date the metadata is.
Availability	Accessibility	The extent to which metadata is available and easily accessible.
	Security	It consists of ensuring that access to metadata is appropriately restricted.
Interoperability	Shareability	The extent to which metadata could be effectively used out of its local environment.
	Discoverability	The extent to which metadata is visible and can be easily found.
	Extendibility	The extent to which the metadata may be easily extended [9].
	Versionability	The extent to which a new version may be easily created [9].

One of the most common questions that come up is: "Which of these metadata quality dimensions is the most important? ". We believe that a definitive answer to this question could not be given since the most crucial quality dimension could differ from one context to another. Indeed, the prioritization of these dimensions mainly depends on the intended use of data, the context, and the organizational policy. Thus, several approaches to assess and improve metadata quality have been suggested in the literature. Each has focused on the dimensions deemed to be most relevant to the study's context and requirements.

In the next section, we survey the available approaches and works that have tackled metadata quality and metadata management in the Big data era.

3 RELATED WORK

Enhancing metadata quality has long been addressed by the literature. Indeed, the first paper that has introduced metadata quality dates back to 2004 [11], when Bruce and Hillmann have defined seven quality dimensions related to metadata: accuracy, consistency, completeness, conformance to expectations, timeliness, accessibility, and provenance. The authors have also defined three levels to address metadata quality: the semantic level, the syntactic level, and the data values themselves. Likewise, Steve et al. in [12] have defined a taxonomy of 38 information quality dimensions divided into three categories (Intrinsic, Relational and Reputational). Moreover, they have suggested an approach to conceptualize and measure metadata quality while considering data use.

Later, there have been many efforts to implement metadata quality dimensions. Ochoa and Duval in [13] have proposed metrics and measurement methods that assess the seven metadata quality dimensions previously defined in [11]. For the same purpose, the authors in [14] have described the formulas to measure and assess five metadata quality metrics (completeness, weighted completeness, accuracy, the richness of information, and accessibility). The proposed metrics were implemented and applied to three public government data repositories. Also, Király and Buchler, in [15], have suggested an open-source implementation to measure some metadata quality metrics such as completeness, uniqueness, and multilingualism.

Other authors have even suggested relevant frameworks to assess and improve metadata quality in different domains. The authors in [9] have defined new metadata quality dimensions such as extendibility and versionability. They have also suggested a quality framework that allows assessing biomedical metadata quality. In research and open science, Kubler et al. in [16] have developed a quality framework that will enable users to evaluate the metadata quality of open data portals using the Analytic Hierarchy Process (AHP). In the same area, Király in [17] have proposed a flexible metadata quality assessment framework that works with several metadata schemas and can also support new schemas. Always in the field of digital libraries, other metadata quality assessment frameworks were suggested in [18] [19] [20].

All the research mentioned above has made a significant contribution to metadata quality enhancement. However, they have not addressed the metadata quality issues that could be raised in big data environments. Indeed, given the particular characteristics of big data, ensuring metadata quality in a big data project is more challenging and requires a more specific and in-depth study. Data in big data environments also goes through multiple processing steps known as the big data value chain (BDVC), leading to a loss and degradation of metadata quality. Moreover, because of big data characteristics, the collected metadata in big data environments is usually not well structured, inconsistent, and low quality. Thus, enhancing metadata quality is highly required for effective and reliable big data analysis. However, very few initiatives have been conducted to address metadata quality in a big data context.

In [21], the authors have identified some of the top quality issues related to metadata used in the Dryad data repository. In this study, the authors have also shown how low metadata quality could have a negative impact on the results of data analytics. This work was focused on enhancing metadata quality in research metadata repositories. Likewise, authors in [22] have suggested an approach to assess metadata quality dimensions such as completeness, multilingualism, and reusability. The proposed method was implemented in Europeana, a digital platform for cultural heritage qualified as Big Data. In [23], a novel big data quality approach encompassing all big data processing phases (input, throughput, output) has been suggested. In this paper, the authors have defined the quality dimensions that should be considered for the data source, data, and metadata.

Table 3: Metadata Quality Approaches

Ref	Main Findings	For Big Data?	Metadata quality Dimensions	Application Domain
[11]	New Metadata Quality Dimensions	No	accuracy, consistency, completeness, conformance to expectations, timeliness, accessibility, provenance	Generic
[12]	A taxonomy of 38 information quality dimensions divided into three categories: (Intrinsic, Relational, and Reputational)	No	completeness, redundancy, consistency, accuracy, completeness	Generic
[13]	Metrics and measurement methods to assess the metadata quality dimensions	No	accuracy, consistency, completeness, conformance to expectations, timeliness, accessibility, provenance	Generic
[14]	Formulas to measure the metadata quality dimensions	No	completeness, weighted completeness, accuracy, the richness of information, accessibility	Generic
[15]	An open-source implementation to measure metadata quality	No	completeness, uniqueness, and multilingualism	Generic
[9]	A quality framework to assess biomedical metadata quality	No	extendibility and versionability	Health
[16]	A quality framework to assess metadata quality in open data portals	No	existence, conformance, retrievability, accuracy, accessibility	Digital libraries
[17]	An extensible metadata quality assessment framework that supports multiple metadata schemas	No	measurement flexibility	Digital libraries
[18]	A study to show the correlation between the use of geospatial datasets and the quality of metadata	No	existence, conformance, accessibility	Digital libraries
[19]	A metadata analyzer that measures metadata quality based on ontology concepts and the terms used in the metadata	No	term coverage, semantic specificity	Digital libraries
[21]	Illustrating some of the primary data quality issues associated with the use of metadata in the Dryad Repository	Yes	completeness, consistency	Digital libraries
[22]	A novel approach to assess metadata quality in Europeana	Yes	completeness, multilingualism, and reusability	Digital libraries
[23]	A global big data quality approach to enhance the quality of data source, data, and metadata as well	Yes	complexity, completeness, usability, linkability, consistency, validity	Generic

Other issues related to metadata management in big data contexts, such as metadata storage and processing, have been addressed in [24] [25] [26]. However, this research has not tackled metadata's quality aspect and has focused on the metadata system's functional architecture. A summary of the reviewed studies is described in Table 3.

To the best of our knowledge, the few available studies that have tackled metadata quality in a big data context are limited to a particular big data application domain (e.g., research repositories, Europeana) and do not deal with the subject whole. Thus, to address metadata quality issues in big data projects, we highlight in the next section how the particular characteristics of big data could impact the quality of metadata. Also, some recommendations to overcome these quality issues will be suggested.

4 METADATA QUALITY ISSUES AND SOLUTIONS IN BIG DATA SYSTEMS

This section highlights how big data could impact the most common metadata quality dimensions: Usefulness, Completeness, Accuracy, Consistency, Shareability, and Timeliness. Moreover, we propose some solutions to overcome the presented issues.

4.1 Usefulness

Managing big data consists of handling a large volume of data and metadata, which generally goes through several processing steps, such as data acquisition, data pre-processing, data storage, data analysis, and data visualization, known as Big Data Value Chain (BDVC). Indeed, in a big data era, metadata is used to store basic descriptions of data and save information related to data processing, such as data analysis results and data quality assessment measures. Thus, when data is collected, the gathered metadata attributes may be associated with the original context of data use and, thus, not relevant to the current environment. Therefore, the accumulated metadata attributes must be cleaned and filtered to keep only useful and valuable information. Once the outside metadata elements are determined, metadata can be cleansed using outlier detection techniques, parsing, statistical clustering, etc. Besides, in a big data context, data are usually gathered from multiple data sources. This could lead to metadata redundancy since data sources use different standards and terminologies to rename the metadata attributes. Thus, a redundancy check must be performed on the selected parameters before embarking on data analysis. There are two types of metadata redundancy:

- **Semantic Metadata Redundancy:** Refers to information redundancy that can be found in metadata attributes. To avoid this kind of redundancy, metadata mapping techniques can align metadata models of two different systems. Also, it is highly recommended to use metadata standards and file naming convention frameworks to automate metadata mapping.
- **Syntactic Metadata Redundancy:** Refers to string similarities that could be found within metadata values. To address syntactic redundancy, similarity measures can be applied. Moreover, other methodologies suggested in [8] [27] to assess metadata records' similarity can be used.

4.2 Completeness

In big data environments, the collected raw data are usually cleaned, reduced, and transformed into a helpful format before being analyzed. This process is also known as data ingestion. During this phase, metadata must be completed with the appropriate contextual information. The gathered metadata is usually insufficient for the intended purpose and lacks contextual information related to the data source and the collection process. Moreover, some missing information may not be specified in the original context because they are considered assumed or not helpful [20]. There are two layers of metadata completeness. The first layer consists of ensuring that the required attributes are not missing. For this purpose, we suggest defining a metadata model that describes all the essential information that needs to be incorporated throughout the metadata. It is beyond this paper's scope to provide a detailed description of this model since it mainly depends on the intended use of the data and the organizational policy. However, we can introduce some general guiding questions as follows:

- Does the available metadata provide all the required descriptive information of the data?
- Does the available metadata provide all the information required by the different processing phases that the data goes through?
- Does the available metadata provide all the information required by the data quality approach adopted?
- Does the available metadata provide all the information required by the organizational policy?

The second layer of metadata completeness ensures that metadata values describe the first layer's attributes completely and exhaustively. For this purpose, the authors in [23] have suggested attributing a clarity and completeness score to each parameter. (e.g., 0 description missing, 1 Insufficient description, 2 Complete description).

4.3 Accuracy

Big data comes from multiple data sources that are not always credible and may contain inaccurate and not factual values (e.g., social media data). This calls into question the reliability of the information provided in the metadata. Thus, knowing the provenance of metadata could already offer an excellent basis to make a quality judgment. Therefore, information like which organization the metadata comes from, how well they are an expert on metadata standards and classifications, what transformations were applied to metadata should be considered when assessing metadata accuracy [11]. Accuracy is usually evaluated by measuring how values match with an accurate model. Thus, a metadata set model could be defined from an open metadata portal or similar corroborative metadata set values. The metadata set model could then be trained and compared to a metadata collection using machine learning and record linkage algorithms. Given the enormous volume of the gathered metadata in big data environments, metadata accuracy could not be directly assessed. Thus, some data inspection procedures could be performed, such as sampling and profiling techniques. As mentioned earlier, metadata is also used to store information related to data processing and data quality assessment. This information is usually quantified using approximate values rather than exact real ones. Indeed, processing a large volume of data is costly in time, money, and effort. Therefore, several quality approaches are confined to analyze a representative sample rather than the entirety of data. The results of the sample quality assessment are then generalized to the whole data. The type of sampling method used to select the data to evaluate can impact the assessment results. Thus, to ensure high accuracy, metadata that contains such results should also include an uncertainty score based on the conditions under which the data analysis and assessment were made.

4.4 Consistency

One of the most common characteristics of big data is variety. This characteristic refers to the heterogeneous aspect of big data from various data sources and incorporates different data types. This aspect also applies to the gathered metadata since the data sources use different metadata standards and terminologies depending on the data and the organizational policy. Thus, one of the most critical issues to address when handling big data is to convert the gathered metadata that is usually unstructured and heterogeneous into uniform and consistent metadata. Also, not all the collected metadata attributes are aligned with metadata standards and schemas.

Moreover, inconsistency may be related to metadata values that do not use the same formats and range of values (e.g., inconsistencies with the structure of date and time attributes), creating some anomalies and bias data analysis. Therefore, it is highly recommended before collecting data to define the standards and schemas used, especially for critical variables. Also, the impact of such anomalies on the potential use of data should be determined.

4.5 Shareability

In [10], Shreeves et al. have observed that even if metadata is of high quality within a local database, this quality may not be maintained when metadata is combined in a federated environment. Later, in [28], Shreeves et al. have introduced "shareability" as a crucial metadata quality dimension to cross-domain resource discovery and promote search interoperability. Thus, shareable metadata is metadata that does not lose its quality and could be effectively used outside its local environment. Ensuring metadata shareability would significantly contribute, especially in a big data era that consists mainly of aggregating data from multiple resources. Given that metadata is also used to store information related to data processing, reusing the obtained results will save time and cost. For this purpose, metadata should use a controlled vocabulary and should be conformed to metadata standards. To enhance metadata shareability, it is highly recommended to use ontologies while structuring metadata. Indeed, ontologies provide a formal representation of metadata attributes with their properties and the relationships that hold between them. Thus, using ontologies allows presenting metadata in a more readable and machine-processible format, promoting metadata discoverability and enabling automatic metadata processing. Also, shared metadata must be well documented and must provide the appropriate context of use. Moreover, content should be kept as short as possible and optimized to keep only the pertinent information.

4.6 Timeliness

One of the essential characteristics of big data is variability. This characteristic sheds light on the dynamic aspect of continuously updated data as new information becomes available such as IoT and social media. Therefore, metadata should also be regularly updated on every change of data. Indeed, outdated metadata is useless and may even confuse the obtained results. Also, some metadata are static and do not change over time. Therefore, metadata that needs to be updated should be determined when establishing the metadata model. Another aspect of metadata timeliness could occur when the collection of data and metadata are not synchronized. This problem becomes serious in the case of real-time data when time-sensitive analysis should be performed. Thus, an assessment of the risks associated with metadata unavailability and the use of outdated metadata should be performed.

It is worth noting that metadata quality dimensions are strongly related to each other. For example, adding additional information to improve completeness may lead to increase metadata redundancy. It is recommended to find the right balance to ensure that any dimension is impacted heavily in such a case. Thus, focusing only on one dimension without considering its correlation with the others may not be a practical approach for supporting metadata quality. Therefore, for a successful metadata quality assessment and improvement, data managers should also consider the existing dependencies between the metadata quality dimensions while prioritizing metadata quality dimensions. This would also help data managers to define the causes that may degrade a specific dimension. Thus, improving metadata quality is not limited to enhance the quality of metadata dimensions. It is a whole process that involves other components such as the intended use of data, the business requirements, the organizational policy, and the data owners.

5 CONCLUSIONS

Ensuring metadata quality is of great importance since it directly impacts data quality and, thus, the extracted insights' reliability. Therefore, several approaches have been suggested in the literature to assess and improve metadata quality. With the emergence of big data, new challenges related to metadata quality have been raised by big data's particular characteristics, known as 7 V's. To the best of our knowledge, no studies have been conducted to address the impact of big data 7V is on the different metadata quality dimensions. Thus, this work's purpose was to highlight the quality issues related to metadata in big data environments. This study analyzed each big data characteristic's impact on the most common metadata quality dimensions. Thus, six metadata quality dimensions have been addressed: accuracy, usefulness, completeness, consistency, shareability, and timeliness. Also, some recommendations to address the raised issues have been suggested. As future work, we aim to propose a novel quality framework for big data that addresses the metadata quality issues raised in this paper and implements the suggested solutions while considering the different factors that could impact metadata quality, including the organizational policy project context and the business requirements.

REFERENCES

A. Immonen, P. Paakkonen, and E. Ovaska, "Evaluating the Quality of Social Media Data in Big Data Architecture," IEEE Access, vol. 3, pp. 2028–2043, 2015.

T. Bruce and D. Hillmann, "The Continuum of Metadata Quality: Defining, Expressing, Exploiting," ALA Ed., Jan. 2004.

P. Király, "Towards an extensible measurement of metadata quality," presented at the Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage, Jun. 2017.

I. El Alaoui, Y. Gahi, and R. Messoussi, "Big Data Quality Metrics for Sentiment Analysis Approaches," in Proceedings of the 2019 International Conference on Big Data Engineering (BDE 2019) - BDE 2019, Hong Kong, Hong Kong, 2019, pp. 36–43.

I. E. Alaoui and Y. Gahi, "The Impact of Big Data Quality on Sentiment Analysis Approaches," Procedia Comput. Sci., vol. 160, pp. 803–810, 2019.

G. Kapil, A. Agrawal, and Prof. R. Khan, "A study of big data characteristics," in International Conference on Communication and Electronics Systems, Oct. 2016, p. 4.

I. E. Alaoui, Y. Gahi, and R. Messoussi, "Full Consideration of Big Data Characteristics in Sentiment Analysis Context," in 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, Apr. 2019, pp. 126–130.

M. Foulonneau, "Information redundancy across metadata collections," Inf. Process. Manag., vol. 43, no. 3, pp. 740–751, May 2007.

C. McMahon and S. Denaxas, "A novel framework for assessing metadata quality in epidemiological and public health research settings," AMIA Summits Transl. Sci. Proc., vol. 2016, pp. 199–208, Jul. 2016.

S. Shreeves, E. Knutson, B. Stvilia, C. Palmer, M. Twidale, and T. Cole, "Is 'Quality' Metadata 'Shareable' Metadata? The Implications of Local Metadata Practices for Federated Collections," Dec. 2010.

T. R. Bruce and D. I. Hillmann, "The Continuum of Metadata Quality: Defining, Expressing, Exploiting," ALA Editions, 2004.

B. Stvilia, L. Gasser, M. Twidale, and L. Smith, "A framework for information quality assessment," JASIST, vol. 58, pp. 1720–1733, Oct. 2007.

X. Ochoa and erik duval, "Automatic evaluation of metadata quality in digital libraries," Int J Digit. Libr., vol. 10, pp. 67–91, Aug. 2009.

K. J. Reiche and E. Höfig, "Implementation of Metadata Quality Metrics and Application on Public Government Data," in 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops, Jul. 2013, pp. 236–241.

P. Király and M. Büchler, "Measuring Completeness as Metadata Quality Metric in Europeana," in 2018 IEEE International Conference on Big Data, 2018.

S. Kubler, J. Robert, S. Neumaier, J. Umbrich, and Y. Le Traon, "Comparison of metadata quality in open data portals using the Analytic Hierarchy Process," Gov. Inf. Q., vol. 35, no. 1, pp. 13–29, Jan. 2018.

P. Király, "Towards an extensible measurement of metadata quality," Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017, June 2017.

A. Quarati, M. De Martino, and S. Rosim, "Geospatial Open Data Usage and Metadata Quality," ISPRS Int. J. Geo-Inf., vol. 10, no. 1, Art. no. 1, Jan. 2021.

B. Inácio, J. Ferreira, and F. Couto, "Metadata Analyser: Measuring Metadata Quality," 2017.

A. Tani, L. Candela, and D. Castelli, "Dealing with metadata quality: The legacy of digital library efforts," Inf. Process. Manag., vol. 49, no. 6, pp. 1194–1205, Nov. 2013.

D. Rousidis, E. Garoufallou, P. Balatsoukas, and M.-A. Sicilia, "Metadata for Big Data: A preliminary investigation of metadata quality issues in research data repositories," Inf. Serv. Use, vol. 34, pp. 279–286, Dec. 2014.

P. Király, "Measuring Metadata Quality," PhD thesis, Georg-August, Gottingen, 2019.

The UNECE Big Data Quality Task Team, "A Suggested Framework for the Quality of Big Data." Dec. 2014.

R.-M. Holom, K. Rafetseder, S. Kritzinger, and H. Sehrschön, “Metadata management in a big data infrastructure,” Procedia Manuf., vol. 42, pp. 375–382, Jan. 2020.

M. Golosova, V. Aulov, and A. Kaida, "Metadata handling for Big Data projects," J. Phys. Conf. Ser., vol. 1117, p. 012007, Nov. 2018.

Smith, K., Seligman, L., Rosenthal, A., Kurcz, C., Greer, M., Macheret, C., et al., "Big Metadata: The Need for Principled Metadata Management in Big Data Ecosystems," in Proceedings of Workshop on Data analytics in the Cloud, 2014.

E. N. Borges, K. Becker, C. A. Heuser, and R. Galante, "An Automatic Approach for Duplicate Bibliographic Metadata Identification Using Classification," in 2011 30th International Conference of the Chilean Computer Science Society, Nov. 2011, pp. 47–53.

S. L. Shreeves, J. Riley, and L. Milewicz, "Moving towards shareable metadata," First Monday, Aug. 2006.