Model for the evaluation of metadata quality: Proposal for open science management in Cuba

Lisandra Díaz de la Paz
Universidad Central "Marta Abreu" de Las Villas, Cuba.
Email: ldp@uclv.edu.cu

Alberto Taboada Crispí
Universidad Central "Marta Abreu" de Las Villas, Cuba.
Email: ataboada@uclv.edu.cu

Amed Abel Leiva Mederos
Universidad Central "Marta Abreu" de Las Villas, Cuba.
Email: amed@uclv.edu.cu

ABSTRACT

The evaluation of metadata quality is of vital importance in the management of open science (OS) in Cuba. In the metadata used in open-access computational systems, unresolved quality problems such as incompleteness of records, ambiguous author names, null values, inconsistency in the use of data exchange formats, and the non-adoption of procedures for metadata quality management are detected. Therefore, this paper proposes a model for the evaluation of metadata quality associated with the management of OS in Cuba. This model is constituted by four stages. Stage 1 refers to the measurement of the identified quality dimensions. Stage 2 corresponds to data cleaning and standardization. Stage 3 corresponds to data integration. Stage 4 deals with data disambiguation based on open access criteria and standards. As a result, completeness at the record level and accuracy at the author name level were identified as the main dimensions of quality. Possible duplicate elements were detected for subsequent integration. A case study is presented with two variant solutions, one for grouping synonymous author names and the other for disambiguating synonymous and homonymous author names, thus, laying the foundations for the interoperability of computational systems.

Keywords: metadata quality, open science, disambiguation of author names, data integration, data cleansing

How to cite: Paz, L. D. de la, Crispí, A. T., & Mederos, A. A. L. (2024). Model for the evaluation of metadata quality: Proposal for open science management in Cuba. In M.J. Peralta González (Ed.), Generation and Transfer of Knowledge for Digital Transformation, II International Symposium, SITIC 2023, Santa Clara, 15-17 November 2023, proceedings. *Advanced Notes in Information Science, volume 6* (pp. 99-113). Pro-Metrics: Tallinn, Estonia. DOI: 10.47909/978-9916-9974-5-1.97.

1. INTRODUCTION

In Project 3, called "Information and Communication Technologies (ICT) in support of educational processes and knowledge management in higher education (ELINF)," a set of actions are being carried out to achieve interoperability among the main open source computer systems: the teaching-learning platform Moodle, which works with learning object metadata; the system for the Automation of Libraries and Documentation Centers (ABCD), whose cataloging module works with metadata using the MARC 21 format (in transit to resource description and access); the institutional repository DSpace, which works with Dublin Core metadata; and the VIVO system for research data management that uses the ontology web language and resource description framework for the management of its internal ontology. The main objective of the project is to form a network where actors in Cuban education and research can search for virtual services needed for their professional tasks, access information for scientific research, and exchange publications and data by working together in an integrated environment that supports open science (OS; Goovaerts et al., 2016). As indicated by FOSTER (2018) and echoed by Meneses-Placeres et al.

(2022), OS includes open access, open research data, open peer review, and OS policies, which are complemented by other more concrete components such as open research practices, reproducible research, open source software, and open licenses. OS complies with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles; for these reasons, the computer systems used in this context must comply with these principles, and for this purpose, metadata formats are used that allow communication and interoperability between them.

In the metadata used in these open-access computer systems, unresolved quality problems such as incompleteness of records, ambiguous author names, null values, inconsistency in the use of data exchange formats, and the non-adoption of procedures for metadata quality management are detected. Therefore, it is of vital importance that the data involved are of adequate quality to be able to interoperate and that the searches performed on them provide relevant information that allows successful information retrieval.

2. METHODOLOGY

As part of the 3 ELINF project, a set of actions are being carried out to achieve interoperability among the main computer systems used in the cataloging of bibliographic records, in the teaching-learning process, and in the institutional repositories of six entities in the country. These entities are the University of Pinar del Río (UPR), the University of Informatics Sciences (UCI), the Central University "Marta Abreu" of Las Villas (UCLV), the University of Camagüey (UC), the University of Holguín (UHO), and the University of Oriente (UO). This paper proposes a model for the evaluation of metadata quality associated

with OS management in Cuba, consisting of four stages. The first stage refers to the measurement of the identified quality dimensions. The second stage corresponds to data cleaning and standardization. The third stage concerns data integration. The fourth stage deals with data disambiguation based on open access criteria and standards, as shown in Figure 1.

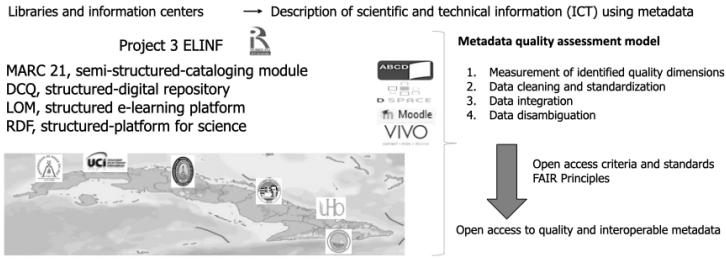

Figure 1. Model for the evaluation of metadata quality in OS management in Cuba.

2.1. Measurement of Identified Quality Dimensions

In the measurement process, the quality dimensions affected by problems identified in the status reconstruction phase and their corresponding metrics should be selected. The measurement is used to establish references, with the purpose of enabling a quality diagnosis (Batini et al., 2009; Batini & Scannapieco, 2016). To perform quality measurement, it is necessary, once the dimensions of interest have been identified, to use metrics that best fit them, as these indicate the degree of quality that the metadata possess.

Data are described through multiple dimensions that are usually grouped into frameworks. These dimensions vary from one framework to another depending on

Table 1. Main metadata quality dimensions identified.

DIMENSION	DEFINITION IN THE CONTEXT OF METADATA	METRICS
Completeness	Extent to which metadata records store all the information necessary to have a global representation of the described object.	$Q_{comp} = \frac{\sum_{i=1}^{N} P(i)}{N} (1)$ $Q_{wcomp} = \frac{\sum_{i=1}^{N} \alpha_i \times P(i)}{\sum_{i=1}^{N} \alpha_i} (2)$
Accuracy	The degree to which the metadata values are "correct" and describe the actual object is defined as a measure of the proximity of a data value *v* to some other value *v'*, which is considered correct.	Syntactic accuracy: $1 - \frac{\sqrt{\sum_{i=1}^{n} d(campo_i)^2}}{\sum_{i=1}^{n} d(campo_i)} (3)$ Semantic accuracy: $\frac{\sum_{i=1}^{N} tf1vector_i * tf2vector_i}{\sqrt{\sum_{i=1}^{N} tf1vector_i^2 * \sum_{i=1}^{N} tf2vector_i^2}} (4)$

the context of analysis of the dimensions (Moges et al., 2013). Bruce and Hillmann's (2004) framework presents seven domain-independent dimensions for the purpose of improving its applicability. These dimensions are completeness (Comp), accuracy (Exac), logical consistency and coherence (CLCo), timeliness (Opor), conformance to expectations (CoEx), accessibility (Acce), and provenance (Proc). In Díaz de la Paz et al. (2021), correspondence between these seven dimensions and several metadata quality frameworks is established for the period from 2009 to 2019. Figure 2 shows that completeness, accuracy, logical consistency, and coherence are the most analyzed, and among them, completeness is present in all the frameworks reviewed.

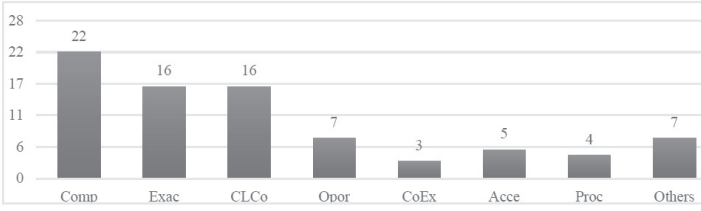

Figure 2. Usability of Bruce and Hillman's dimensions in other frameworks. Source: Based on Díaz de la Paz et al. (2021).

Table 1 shows the main dimensions identified in the context of the 3 ELINF project and the metrics used to assess the quality of metadata in the computational systems involved. The metrics are taken from Ochoa and Duval (2011, 2006).

2.2. Data Cleaning and Standardization

Data cleaning and standardization refers to the correction in the data of possible errors, for example, incomplete data,

duplicates, inconsistent formats in terms of description, abbreviations and units of measurement, missing input data, or data that violate the integrity constraints of the system. The cleaning stage is one of the most important, as it ensures the quality of the data. At this stage, anomalies detected in the data integration process should be corrected. Data anomalies can be classified into syntactic anomalies, semantic anomalies, and context anomalies (López Porrero, 2011).

- Syntactic anomalies: These anomalies refer to lexical errors, formatting and domain errors, and non-standardization of information.
- Semantic anomalies: These anomalies include violations in row integrity constraints, contradictions in data values that violate some kind of dependency between them, and duplicate rows and invalid rows.
- Context anomalies: These anomalies include missing data values in rows or missing complete rows that exist in the mini-world and have not been represented.

The problems present in a single source are compounded when you need to integrate data from multiple sources. Each source may contain erroneous data, and data may have different representations in each source or be represented in one or the other in a contradictory way or mixed with other data because the sources are created, developed, and maintained independently and for very specific purposes. This increases the level of heterogeneity in the database management systems, data models, and data source designs used.

2.3. Data Integration

Data integration refers to the process of combining data from different heterogeneous sources with each other

in order to provide the user with a unified view of the data that is clean, free of anomalies, and of the required quality.

Data integration is a non-trivial, multi-faceted, and, in many cases, autonomously impossible process. There are several heterogeneity conflicts which need to be addressed by a possible solution:

- Syntactic heterogeneity: The language used in the two sources may differ even if they are semantically identical. For example, in one source, you can have students and in the other pupils.
- Structural heterogeneity: The types and structure of the data may also vary. For example, the salary can be represented as a dollar value in a single table or have a reference to the identifier of a pay scale stored in a different table.
- Representational heterogeneity: Different sources may implement different models for their data, different levels of normalization may appear, and data represented as a single element may exist in multiple elements in another source.
- Semantic heterogeneity: The opposite of syntactic problems, where two semantically different objects are referred to in the same way.

These are some of the fundamental problems of all forms of data integration. Data integration provides a mechanism for joining data from different sources into a single schema. Integration takes place in two stages:

- Homogenization: Transformation of the information from the original format of the natural sources to the

format and data model of the target system takes place in this stage.

- Integration: The retrieved information is aggregated and organized into the target system.

2.4. Data Disambiguation

The disambiguation of data with textual information such as names of persons, institutions, words, and so on is an open problem of natural language processing, which includes the identification of when that data in question present polysemy (multiple meanings), and it is desired to match each data in the correct context.

There are dissimilar approaches to address this problem. In Díaz-de-la-Paz et al. (2022), a framework is presented that combines an ontological approach with deep-learning techniques for personal author name disambiguation based on the gold-standard LAGOS-AND dataset (Zhang et al., 2021). One of the fundamental steps at this stage is duplicate item detection and removal. There is a direct proportionality between metadata quality assessment and OS management; the higher the metadata quality, the greater the success of retrieving disambiguated and interoperable data in the context of OS in Cuba.

3. RESULTS AND DISCUSSION

One of the case studies where metadata quality is evaluated is the cataloging module of the ABCD suite, for which two databases are used; one extracted from the UCLV catalog (BD_UCLV) and the other from the UCI catalog (BD_UCI), with 18,745 and 13,807 records, respectively, collected at the end of 2016. The "MARCQuality" tool

is used to issue a quality criterion for these datasets. Figure 3 shows the fields cataloged in each bibliographic record, organized from lowest to highest according to their label, and calculates the completeness as expressed in metric 2 of Table 1, using the particle swarm-based optimization method described in Díaz de la Paz et al. (2021).

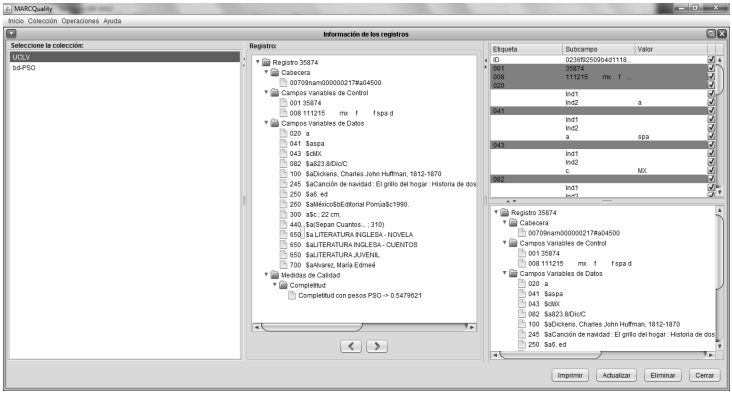

Figure 3. Bibliographic record belonging to the BD_UCLV collection.

Then, the possible duplicated elements are listed, and two options to mark are observed; one is the "Main" record (which is marked by default according to the completeness values calculated but can be changed by the specialist), and the other is the "Select" option to mark those records to be integrated. The integrated record is displayed in the right pane once the conflicts detected during the merging are resolved. If there are no conflicts between the records or if they could be solved automatically, the integrated record is displayed in the right panel. Otherwise, a dialog box is displayed to ask the user how to proceed in cases where there is a conflict (see Figure 4).

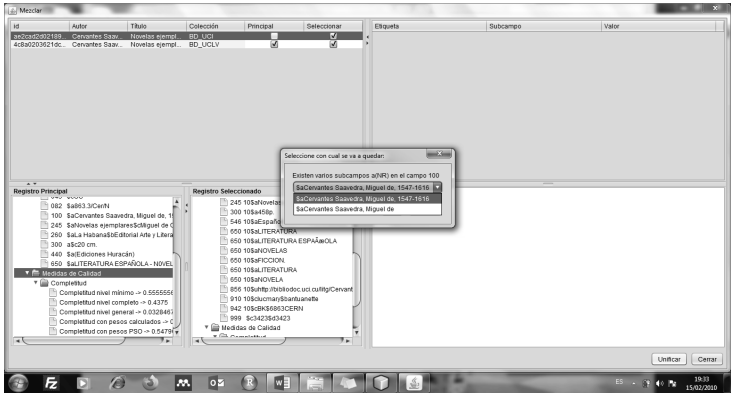

Figure 4. Resolution of conflicts detected during the data integration process.

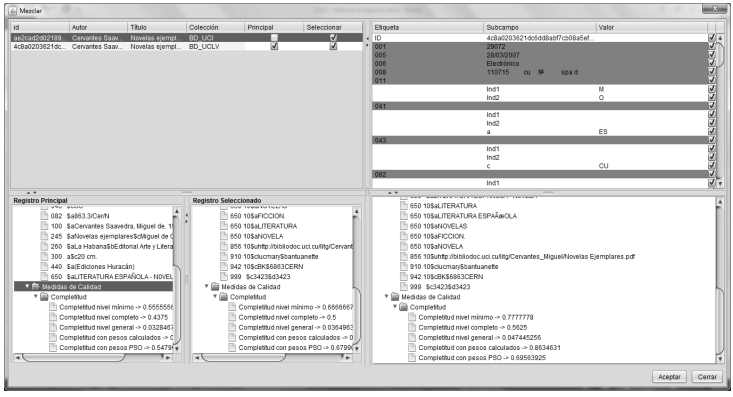

Figure 5. Example of an integrated record when resolving conflicts.

When the integration is performed, the resulting record has a completeness value equal to or higher than the highest completeness value of the separate records. Figure 5 shows that the completeness criterion of the record marked as the main record is 0.5479, that of the selected record is 0.6799, and that obtained by integrating the records is approximately 0.6956. Therefore, its quality is improved in

terms of the completeness dimension. These changes can be made persistent in the corresponding collection if the specialist desires to do so.

Another functionality provided by the MARCQuality tool is the standardization of synonymous author names in the "Standardize authors" option. Figure 6 shows a list of authors and co-authors grouped according to the unsupervised grouping method based on DBSCAN density.

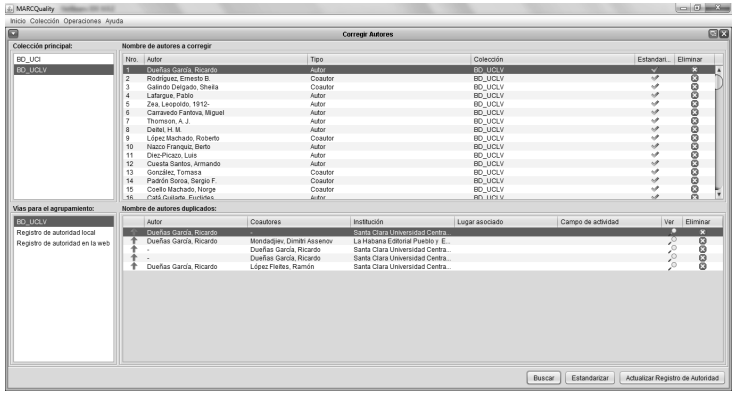

Figure 6. List of authors and co-authors with their possible name variants.

To disambiguate the names of homonymous authors (names that are spelled the same but represent different persons), the hybrid framework proposed by the authors of this paper (Díaz-de-la-Paz et al., 2022), inspired by the method presented in Zhang et al. (2021), is used. Rules and queries that increase the semantic rigor in the construction of the weighted co-authorship network are added to this proposal. This framework combines the ontological approach with several deep-learning network models. For the validation of this framework, a comparison was performed with respect to different data portions of the

LAGOS-AND dataset; the hybrid approach obtains a good performance for gated recurrent unit (GRU), whose F1 and recall value improves by more than 9% and 17% to the MAG Author ID and Name Similarity datasets, respectively, noting that long short-term memory improves the accuracy value by 7%, as illustrated in Díaz-de-la-Paz et al. (2022). Therefore, in the present work, the hybrid solution of AND ontology and GRU deep neural network model is used, which shows the best results for personal author name disambiguation.

4. CONCLUSIONS

This paper proposed a model for the evaluation of the quality of metadata associated with OS management in Cuba. The most relevant quality dimensions to be addressed in this context were determined, as well as how to measure them. The proposal was evaluated in an instrumental case study, which allowed knowing the degree of completeness and accuracy present in the analyzed datasets; possible duplicated elements were detected for their subsequent integration. A case study was presented with two variants of the solution, one to group the names of synonymous authors and the other to disambiguate the names of authors with polysemy. This served as a basis for the interoperability of the computational systems.

REFERENCES

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*(3), 1–52. https://doi.org/10.1145/1541880.1541883

Batini, C., & Scannapieco, M. (2016). *Data and information quality. Dimensions, principles and techniques*. Springer International Publishing.

Bruce, T. R., & Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. I. Hillmann & E. L. Westbrooks (Eds.), *Metadata in practice*. American Library Association.

Díaz-de-la-Paz, L., Concepción-Pérez, L., Portal-Díaz, J. A., Taboada-Crispi, A., & Leiva-Mederos, A. A. (2022). Framework for author name disambiguation in scientific papers using an ontological approach and deep learning. . In B. Villazón-Terrazas, F. Ortiz-Rodriguez, S. Tiwari, M.-A. Sicilia, & D. Martín-Moncunill (Eds.), *Communications in computer and information science* (vol. 1686, pp. 216233). Springer International Publishing.

Díaz de la Paz, L., Riestra Collado, F. N., García Mendoza, J. L., González González, L. M., Leiva Mederos, A. A., & Taboada Crispi, A. (2021). Weights estimation in the completeness measurement of bibliographic metadata. *Computación y Sistemas, 25*(1), 117–128. https://doi.org/10.13053/cys-25-1-3355

FOSTER. (2018). Manual de Capacitación sobre Ciencia Abierta.

Goovaerts, M., Ciudad Ricardo, F. A., & Benitez Erice, D. (2016). Desarrollo de una red virtual de investigación y educación para la información científico en Cuba. In *Congreso Internacional de Información Info'2016* (pp. 118).

López Porrero, B. (2011). *Limpieza de datos: reemplazo de valores ausentes y estandarización*. Universidad Central "Marta Abreu" de Las Villas.

Meneses-Placeres, G., Álvarez Reinaldo, L. A., & Machado Rivero, M. O. (2022). Revisión de las Prácticas de Ciencia Abierta en América Latina y el Caribe. *Revista Cubana de Transformación Digital*, *3*(1).

Moges, H.-T., Dejaeger, K., Lemahieu, W., & Baesens, B. (2013). A multidimensional analysis of data quality for credit risk management: New insights and challenges. *Information & Management*, *50*(1), 43–58. https://doi.org/10.1016/j.im.2012.10.001

Ochoa, X, & Duval, E. (2011). Learnometrics: metrics for learning objects. In *Proceedings of the 1st international conference on learning analytics and knowledge* (pp. 1–8). ACM.

Ochoa, X., & Duval, E. (2006). Quality metrics for learning object metadata. In *Proceedings of World conference on educational multimedia, hypermedia and telecommunications* (pp. 1004–1011). AACE.

Zhang, L., Lu, W., & Yang, J. (2021). LAGOS-AND: A large gold standard dataset for scholarly author name disambiguation. *arXiv preprint arXiv:2104.01821*, 1–27.