On Metadata Quality in Sceiba, a Platform for Quality Control and Monitoring of Cuban Scientific Publications

Eduardo Arencibia[1](✉), Rafael Martinez[1], Yohannis Marti-Lahera[2], and Marc Goovaerts[3]

[1] CRAI at University of Pinar del Río "Hermanos Saíz Montes de Oca", 300 Martí, Pinar del Río 20100, Pinar del Rio, Cuba
{eduardo.arencibia,rafael.martinez}@upr.edu.cu

[2] Central Library of Havana University, San Lazaro and L, 10400 Havana, Cuba
yohannis@dict.uh.cu

[3] Hasselt University, Martelarenlaan 42, 3500 Hasselt, Belgium
marc.goovaerts@uhasselt.be

Abstract. It is introduced a platform for quality control and monitoring of Cuban scientific publications named Sceiba. To this end, it needs to collect scientific publications comprehensively at the national level. Metadata quality is crucial for Sceiba interoperability and development. This paper exposes how metadata quality is assured and enhanced in Sceiba. The metadata aggregation pipeline is worked out to collect, transform, store and expose metadata on Persons, Organizations, Sources, and Scientific Publications. Raw data transformation into Sceiba's internal metadata models includes cleaning, disambiguation, deduplication, entity linking, validation, standardization, and enrichment using a semi-automated approach aligned with the findability, accessibility, interoperability, and reusability principles. To meet the requirements of metadata quality in Sceiba, a three-layer structure for metadata is used, including 1) discovery metadata, which allows the discovery of relevant scientific publications by browsing or query, 2) contextual metadata, which allows a) rich information on persons, organizations and other aspects associated with publications, b) interoperation among common metadata formats used in Current Research Information Systems, journals systems or Institutional Repositories; 3) detailed metadata, which is specific to the domain of scientific publication evaluation. The example provided shows how the metadata quality is improved in the Identification System for Cuban Research Organizations, one of Sceiba´s component applications.

Keywords: Current research information system · Metadata quality · Scientific publication quality

1 Introduction

Metadata topics, usually understood as data about data, are receiving a lot of attention in the realm of Information Systems research. Metadata can be defined as "structured,

E. Garoufallou et al. (Eds.): MTSR 2021, CCIS 1537, pp. 106–113, 2022.
https://doi.org/10.1007/978-3-030-98876-0_9

encoded data that describe characteristics of information bearing entities to aid in the identification, discovery, assessment, and management of the described entities" [8].

The use of metadata models and standards are key to Current Research Information Systems (CRIS), especially in achieving higher levels of interoperability with internal and external systems. At the same time, it is needed to assure metadata quality in this endeavour. Wiley [11] and Allen [1] define metadata quality criteria: completeness, accuracy, consistency, standardization, machine-processable, and timely. Also, FAIR (Findability, Accessibility, Interoperability, and Reusability) principles [12] must be considered since they are crucial for metadata quality in CRIS.

Empirical studies [2–4] state that metadata quality should be enhanced by cleaning, disambiguation, deduplication, enrichment and validation of metadata. These processes are related to metadata curation that should be carried out after metadata collection. The peril of ignoring metadata standards and its quality in a CRIS have several implications in the performance of research organizations [9]. Even research assessment can be affected by metadata quality due to the need for all institutional research outputs to be collected and described in a standardized way in a single system [3] at regional, national and institutional levels.

A VLIR-UOS[1] Joint project entitled "Improving quality control and monitoring of scientific publications on national and institutional levels" was launched to address this and other issues related to scientific publications. The project is developed through the cooperation of six universities from Cuba, Belgium and Peru: University of Havana, University of Pinar del Rio, National Agrarian University of La Molina, University of San Ignacio Loyola, Hasselt University and Antwerp University. With the general objective of "Enhancing the quality of scientific publications as part of the research output", in Cuba the project faces the problem of setting up a system capable of gathering comprehensively the research output metadata at national level. The Sceiba[2] platform aims to be the answer to this problem. Metadata quality is a key element to consider by the platform.

This paper introduces the Sceiba platform, focusing on the processes by which metadata quality is assured. Section 2 gives a general description of the structure, the metadata model and the metadata aggregation pipeline of Sceiba. Section 3 exposes how metadata quality is ensured and enhanced in Sceiba, using as an example the application Identification System for Cuban Research Organizations. Final considerations, main challenges and further developments are presented in Sect. 4.

2 Sceiba Structure, Pipeline and Metadata Model

The Sceiba platform is powered by Invenio[3], an open-source framework to build repositories. It follows the next-generation repositories principles from COAR[4]. Sceiba emerges

[1] Vlaamse Interuniversitaire Raad - Universitaire Ontwikkelingssamenwerking' (VLIR-UOS), more information about the project can be found in https://www.vliruos.be/en/projects/project/22?pid=4202.

[2] Sceiba is a word that arises from the combination of the Latin "sci" and Ceiba, a leafy tree considered sacred by several Cuban traditions.

[3] https://invenio.readthedocs.io/en/latest/.

[4] https://www.coar-repositories.org/news-updates/what-we-do/next-generation-repositories/.

as an open system, acting as a framework to build applications for evaluating and monitoring scientific publications. The platform collects and manages scientific publication metadata and metadata linked to identification systems for organizations and persons. Metadata standardization relies on using controlled vocabularies and persistent identifiers where possible.

Sceiba is divided into the following components:

- Sceiba Core: manages scientific publications and main sources.
- Organizations Identification System: manages research organizations profiles.
- Persons Identification System: manages research related persons' profiles.
- System for Controlled vocabularies: manages vocabularies related to research data and metadata.
- Tools for monitoring and evaluation

Sceiba applies a three-layer Metadata Architecture, as proposed by Jeffery [6], to ensure the quality of metadata. Sceiba's feeding sources use heterogeneous metadata standards and schemas like DCMI (DC terms), Qualified DC, CERIF or ontologies. Others, like domestic developed systems, do not assume international and recognized standards. The metadata standards used in the discovery metadata (first layer) have the advantage of enabling the easy linkage of large numbers of scientific publications. However, they insufficiently describe the relationships between those publications, persons and organizations involved in publications as research outputs. "The syntax of flat metadata standards is often insufficiently formal, the semantics presented are rudimentary, they do not handle multilingualism well, they do not respect referential integrity, and they do not handle temporal relationships well" [13].

Because of the disadvantages of flat metadata standards, it was chosen to add contextual metadata (second layer) that offers structured relationships inspired by the CERIF [7] and GRID models[5], mainly based on persistent identifiers usage (see Fig. 1). The contextual metadata allows rich information on many publications' aspects, including the required metadata fields about the context, provenance, organizations, and persons. Also, detailed metadata from the domain (third layer) is needed, with the use of rich semantics in the contextual metadata layer and the ability to crosswalk from one semantic term to another. The domain metadata layer is oriented, but not limited to, the quality of scientific publications or criteria related to their visibility and impact.

The three-layer metadata architecture and metadata quality have a significant impact on the metadata aggregation process implementation. Therefore, an aggregation metadata pipeline (see Fig. 2) is in development with four general stages:

- Collection of data from primary and secondary sources with heterogeneous metadata models and standards.
- Transformation of raw data into Sceiba's internal metadata models. This stage includes processes like cleaning, disambiguation, deduplication, entity linking, validation, standardization, and enrichment using a semi-automated approach aligned to FAIR principles.

[5] https://grid.ac/format.

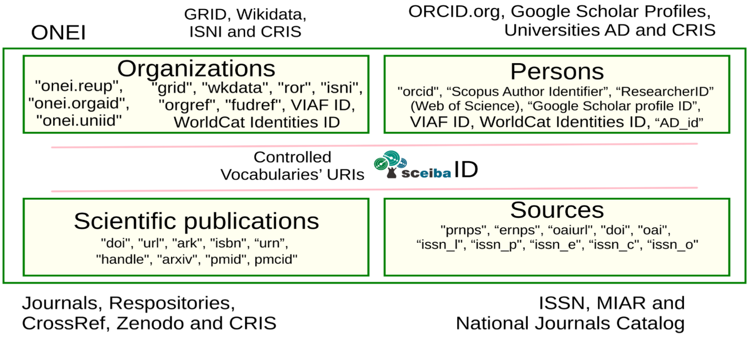
ONEI
GRID, Wikidata,
ISNI and CRIS
ORCID.org, Google Scholar Profiles,
Universities AD and CRIS
Organizations
"onei.reup",
"onei.orgaid",
"onei.uniid"
"grid", "wkdata", "ror", "isni",
"orgref", "fudref", VIAF ID,
WorldCat Identities ID
Persons
"orcid", "Scopus Author Identifier", "ResearcherID"
(Web of Science), "Google Scholar profile ID",
VIAF ID, WorldCat Identities ID, "AD_id"
Controlled
Vocabularies' URIs
sceiba ID
Scientific publications
"doi", "url", "ark", "isbn", "urn",
"handle", "arxiv", "pmid", pmcid"
Sources
"prnps", "ernps", "oaiurl", "doi", "oai",
"issn_l", "issn_p", "issn_e", "issn_c", "issn_o"
Journals, Respositories,
CrossRef, Zenodo and CRIS
ISSN, MIAR and
National Journals Catalog

Fig. 1. Identifiers in Sceiba

- Storing the metadata considering the most probable scenarios for recovering by persistent identifiers, text fields, and relationships between publications and other entities included metadata model (see Fig. 3).
- Exposure of metadata using standards to guarantee interoperability and reusability.

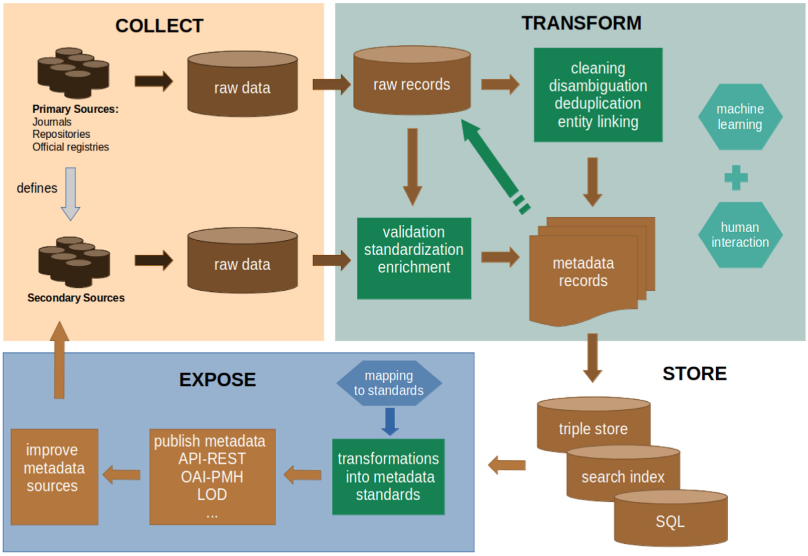

Fig. 2. Sceiba metadata aggregation pipeline

The Sceiba metadata model involves entities like Persons, Organizations, Sources, and Scientific Publications (see Fig. 3). Sceiba Core works as an aggregator at the national level and therefore requires, in each metadata record, additional source information from the original content providers to be encoded. Provenance-related metadata also ensures compatibility with OpenAIRE[6].

All records of each entity have persistent identifiers, brought from original sources if they exist or added in the enrichment processes. In addition, Sceiba also assigns unique identifiers intended to be persistent as long as the platform lives. By working in this way, an instance with different identifiers in different sources is unified in Sceiba. Relationships are established using Sceiba ID (See Fig. 1 for an example of the persistent identifiers used in different sources that are incorporated into Sceiba).

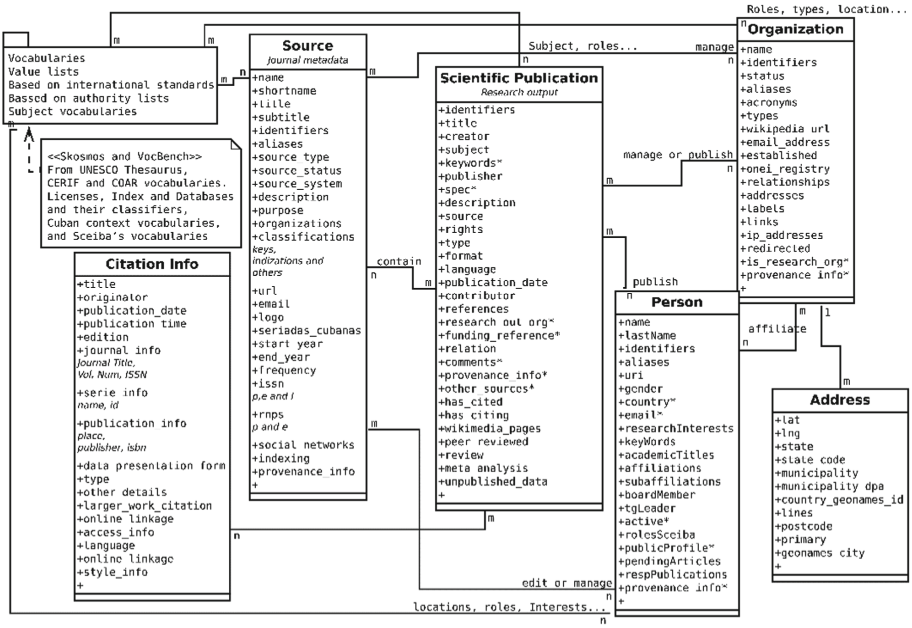

Fig. 3. Main entities in Sceiba platform.

Note: Fields with an asterisk (*) are those following other standards than the main used for each entity.

3 Enhancing the Metadata Quality in Sceiba Organizations

Sceiba includes the development of an Identification System for Cuban Organizations called Sceiba Organizations (see Sceiba components in Sect. 2). It aims to enable connections between organization records in various systems. This application component only includes officially registered research organizations as listed by the National Office of Statistics and Information of Cuba (ONEI, Spanish acronym). The data is collected

[6] https://guiasopenaire4.readthedocs.io/es/latest/use_of_oai_pmh.html#formato-de-los-metadatos.

automatically from public Microsoft Excel documents, and cross-walked into the Sceiba data model. There are several types of organizations in the official ONEI registry (primary and authority source), so those that are of interest in the context of scientific publications are selected by a cleaning process.

Constraints emerge from the use of the ONEI source: although the data is accessed openly, international metadata standards or internationally recognized persistent identifiers are not used. Besides, its structure is not intuitive and is dispersed over several files. Because of these and other issues, it was needed to add a contextual metadata layer. This second layer was developed using the GRID[7] data model, the Cuban organizations context and the wikidata registries. The project is considering the integration of ROR's data[8], as GRID is passing the torch to ROR. Both are a great inspiration for this Sceiba's component application.

Disambiguation has been more labour intensive for organizations' metadata of coming from ONEI and GRID, because of the initial absence of persistent identifiers in the ONEI metadata. The enrichment will come from other sources such as Wikidata, ROR and ISNI. Wikidata is gaining popularity in libraries as an open and collaborative global platform for sharing and exchanging metadata [10]. The Sceiba organizations identification system is able to collect data from every Cuban research organization, and with more options possible when a Wikidata ID is available. For instance, the Sceiba integration with Wikidata allows to expose statistical graphs with data from Wikipedia about the organizations and link them to more details in Scholia[9] website.

Enrich metadata through curation is a process that can't be fully automated, therefore, to put a human in the loop, user interface was designed to allow actions such as duplicate detection, disambiguation and enrichment of records. The user interface allows selecting a master organization, searches for possible duplicates and disambiguates and merges fields when applicable.

The algorithms for duplicate detection are based on external identifiers. In case of any match they are considered as the same organization. Sceiba keeps ONEI codes, already transformed into URIs and links unequivocally with common persistent identifiers used internationally for organizations identification (see Fig. 1), when possible. Offering this way a service to identify them more easily henceforth. Therefore, if reviewers find inconsistency they can correct them through the curation user interface.

The project is working on an approach which combines rule-based, machine learning and manual approach to connect heterogeneous author affiliations in scientific publications to known research organizations. Thus, possible duplicates of organizations will be detected applying the parametrized finite-state graphs method proposed by Gálvez & Moya-Anegón [5] and through human processing. Using this mixed disambiguation method would reduce the amount of manual reviewing to the most difficult cases, increase the precision of disambiguation in organization-scientific publication relationships and facilitate more accuracy in control and monitoring of scientific publication at institutional and national levels.

[7] https://grid.ac/.

[8] https://ror.org/.

[9] https://scholia.toolforge.org/.

Updating organization information and managing organizational hierarchies is a challenging issue during enrichment processes. Subordination relationships are represented between organizations already included in the ONEI registry. How to get a deeper and comprehensive representation of organizational hierarchy is a pending task in the project. Much more work needs to be done to clarify workflows and methods.

Also, the self-update procedure by organizations to improve the content curation of metadata is still in development. Sceiba proposes a follow-up report on the completeness of organizations' metadata. Organizations will be required to complete the mandatory and recommended metadata according to the Sceiba metadata model on a periodic basis.

The quality control process in the transformation of the data, seeks to ensure not only that it is complete but also that its syntactic and semantic value, and its overall compliance with the aforementioned quality metadata criteria and FAIR principles (e.g. the use of the OpenAire validator to confirm that it complies with OpenAire guidelines) is realized. Thanks to the use of international standards and this FAIRification workflow, quality metadata related to Cuban organizations will be reusable, looking for improving records in domestic systems and feeding other organization identification systems (e.g. ROR) to improve Cuban organizations visibility on those international databases.

4 Challenges and Further Work

The paper focuses on the challenges about metadata quality. Improving the quality of metadata will always be essential to achieve Sceiba's objectives. It means the further development of the Sceiba metadata model to include other entities such as projects and other research outputs besides scientific publications, going deeper on details in the domain of research systems, the improvement of curation and transformation operations of metadata and the exposure of metadata for reuse in the context of open data and open science. Crucial in this process will be cooperation with data creators on improving records and metadata.

The project is also developing policies and workflows for quality control and monitoring of scientific publications, taking in account the specificity of Spanish-speaking countries like Cuba and Peru, with a large scientific production that is not taken in account in the international citation databases. A vision has been worked out about policy and guidelines to ensure the sustainability and adoption of Sceiba principles for quality control and monitoring of Cuban scientific publications at the national and institutional levels. The policy and guidelines will be the subject of another paper.

The challenges for the project will be to address the development of the platform, with a strong focus on metadata standards and quality, while implementing the specific policies and workflows developed by Sceiba.

Acknowledgements. The work of the Sceiba project was supported by the 'Vlaamse Interuniversitaire Raad - Universitaire Ontwikkelingssamenwerking' (VLIR-UOS), Belgium. The authors are team members of the Sceiba project. They like to thank Sadia Van Cauwenbergh (Hasselt University) and Raf Guns (Antwerp University) for their suggestions on the article and to the Sceiba team of the University of Pinar del Rio for their contribution to the development of the Sceiba platform.

References

1. Allen, R.: Metadata for social science datasets. In: Rich Search and Discovery for Research Datasets: Building the Next Generation of Scholarly Infrastructure, pp. 40–52. Sage (2020)
2. Alma'aitah, W.Z.A., Talib, A.Z., Osman, M.A.: Opportunities and challenges in enhancing access to metadata of cultural heritage collections: a survey. Artif. Intell. Rev. **53**(5), 3621–3646 (2020)
3. Bryant, R., Clements, A., Castro, P., de Cantrell, J., Dortmund, A., Fransen, J., et. al.: Practices and patterns in research information management: findings from a global survey (2020). https://doi.org/10.25333/BGFG-D241
4. Fernandes, S., Pinto, M.J.: From the institutional repository to a CRIS system. Qual. Quant. Methods Libr. **7**(3), 481–487 (2019)
5. Galvez, C., Moya-Anegón, F.: The unification of institutional addresses applying parametrized finite-state graphs (P-FSG). Scientometrics **69**, 323–345 (2006). https://doi.org/10.1007/s11192-006-0156-3
6. Jeffery, K., Houssos, N., Jörg, B., Asserson, A.: Research Information management: the CERIF approach. Int. J. Metadata Semant. Ontol. **9**, 5–14 (2014). https://doi.org/10.1504/IJMSO.2014.059142
7. Jörg, B., Jeffery, K., Dvorak, J., Houssos, N., Asserson, A., Grootel, G., et.al.: CERIF 1.3 Full Data Model (FDM): introduction and specification (2012)
8. Ma, J.: Managing metadata for digital projects. Libr. Collect. Acquis. Tech. Serv. **30**, 17–23 (2006)
9. Schriml, L.M., Chuvochina, M., Davies, N., Eloe-Fadrosh, E.A., Finn, R.D., Hugenholtz, P., et al.: COVID-19 pandemic reveals the peril of ignoring metadata standards. Sci. Data **7**(1), 188 (2020). https://doi.org/10.1038/s41597-020-0524-5
10. Tharani, K.: Much more than a mere technology: a systematic review of Wikidata in libraries. J. Acad. Librarianship **47**(2), 102326 (2021). https://doi.org/10.1016/j.acalib.2021.102326
11. Wiley, C.: Metadata use in research data management. Bull. Assoc. Inf. Sci. Technol. **40**(6), 38–40 (2014). https://doi.org/10.1002/bult.2014.1720400612
12. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**(1), 1–9 (2016). https://doi.org/10.1038/sdata.2016.18
13. Zuiderwijk, A., Jeffery, K., Janssen, M.: The potential of metadata for linked open data and its value for users and publishers. J. e-Democracy Open Gov. **4**(2), 222–244 (2012). https://doi.org/10.29379/jedem.v4i2.138