

THE SCIENTIFIC PRODUCTION ON SEMANTIC ENRICHMENT OF METADATA USING THE OPEN KNOWLEDGE MAPS TOOL

Ana Carolina Novaes de Mendonça

Master's in Information Science, Federal University of São Carlos (UFSCar), São Carlos, São Paulo, Brazil.
Brazilian Institute of Information in Science and Technology (IBICT), Brasília, Brazil.

mnovaesana@gmail.com

<https://orcid.org/0009-0004-0285-9932>

Ana Carolina Simionato Arakaki

Postdoctoral research in Computer Science, University of São Paulo (USP), São Paulo, Brazil.

Ph.D. in Information Science, São Paulo State University (UNESP), Marília, São Paulo, Brazil.

Professor at the Faculty of Information Science, University of Brasília (UnB), Brasília, Brazil.

anaarakaki@ibict.br

<https://orcid.org/0000-0002-0140-9110>

ABSTRACT

Objective: the study addresses the semantic enrichment of metadata, a process that enhances the quality and usefulness of bibliographic data by adding contextual information and semantic relationships. This process contributes to data accessibility and interoperability, enabling the integration of different databases and facilitating the discovery of previously isolated information. The objective of the study was to describe the scientific production on the application of semantic enrichment in the context of bibliographic data. **Method:** this is a descriptive and qualitative study, using the Open Knowledge Maps tool and complementary searches in recognized scientific databases. **Results:** sixteen studies were selected, applying different methods such as ontologies, controlled vocabularies, and Linked Open Data. **Conclusions:** by adding layers of meaning, semantic enrichment of metadata enables more precise searches, facilitates the integration of diverse information sources, and broadens knowledge discovery across various domains.

Keywords: Semantic enrichment. Metadata. Bibliographic data.

1 INTRODUCTION

Semantic enrichment is a process that enhances metadata by adding more contextualized meanings, either through the creation of new metadata or the improvement of existing ones (Isaac et al., 2015). This approach represents a significant development in data management, making data more understandable, interoperable, and useful, as well as enabling its connection to other databases and the exposure of information that was previously isolated and disconnected. By providing additional layers of information to metadata, semantic enrichment can also support decision-making, given that decisions are often made under pressure or

based on incomplete and outdated data (Arakaki; Isotani, 2024; Pfeffer; Sutton, 2006).

Metadata organize, describe, and facilitate access to stored resources. According to Zeng and Qin (2008, p. 07), they correspond to information that describes any document or object, in both digital and traditional formats. When well structured, metadata enable the description of items by providing details such as author, date, format, and subject, which facilitates their retrieval and use.

The evolution of cataloging has been driven by the adoption of specialized systems that have significantly improved the organization and retrieval of bibliographic and authority records. Catalogs play a central role in Information Science, as they ensure the structuring and accessibility of bibliographic data, which are essential for the construction of organized and reliable records.

In this context, the semantic enrichment of metadata is a process that enhances the organization and retrieval of information in digital environments. In the context of bibliographic data, this practice enables the addition of structured and more meaningful semantics to metadata, favoring their interoperability and connection with other sources of knowledge. However, the literature still reveals a gap regarding scientific production on this topic, particularly with respect to the approaches adopted. This leads to the following research question: how has scientific production addressed the semantic enrichment of bibliographic metadata?

Based on this problem, this study aims to describe the scientific production on the semantic enrichment of bibliographic metadata. In doing so, it seeks to contribute to the understanding of the current state of research in the field and to provide support for future investigations and for the improvement of semantic enrichment strategies applied to bibliographic data.

2 THEORETICAL FRAMEWORKS

Data play a fundamental role in cataloging and on the Web; when they are structured and standardized, they represent the fundamental attributes of an item and can be widely accessed and made available. Santos and Sant'Ana (2013, p. 205) define data as

[...] a unit of content necessarily related to a given context and composed of the triad entity, attribute, and value, in such a way that, even if the details about the context of the content are not explicitly stated, they must be implicitly available to the user, thus allowing their full interpretation.

Therefore, by identifying the context in which data are situated, it becomes possible to understand their meaning and correctly interpret the information conveyed. This is especially important when linguistic elements are also considered, since meaning may vary depending on the context in which it is used.

Data differ as structured, semi-structured, or unstructured. Schöch (2013) highlights the importance of distinguishing these types of data. Structured data are stored in databases and follow an explicit data model, in which key/value pairs have well-defined identifiers and clear relationships. Unstructured data, such as free text, images, or audio, do not present a formal organization, making their relationships and meanings implicit. Between these two are semi-structured data, which are frequently found in eXtensible Markup Language (XML) or JSON files. They have some level of organization but do not follow a rigid schema and may vary in structure depending on the context and the rules applied.

Data take on the role of characterizing other data and are referred to as metadata. Pomerantz (2015, p. 26) states that the definition of “data about data” is not useful and therefore defines metadata as “[...] a statement about a potentially informative object.” As mentioned by Zen and Qin (2008), metadata organize and describe resources and thus serve as the basis upon which new layers of information are added.

For Gilliland (2016, p. 2), metadata are “value-added information used to organize, describe, track, and enhance access to information objects, physical items, and collections related to these objects.” Arakaki (2019), drawing on several authors, classifies metadata according to their typologies and functions, categorizing them as administrative, authentication, preservation, provenance, technical, meta-metadata, descriptive, rights, access and use, structural, and markup languages.

Given this panorama, the use of metadata goes beyond the organization and retrieval of information, assuming a strategic role in decision-making. In various contexts, the availability of structured metadata helps solve problems and define

strategies to achieve specific objectives. This becomes even more relevant in situations in which administrative decisions are made under pressure, often based on incomplete, outdated information or grounded solely in individual experience (Arakaki; Isotani, 2024; Pfeffer; Sutton, 2006).

To strengthen quality and minimize distortions in metadata, the process of metadata enrichment can be used to increase the richness of data through the addition of contexts and semantic relationships that go beyond the explicit information contained in basic metadata (Zeng, 2018).

This process can be used in various contexts, such as bibliographic, geographic, biomedical data, and so on, each with its own metadata standards and types of data. For example, in the context of bibliographic data, semantic enrichment enables, in addition to traditional information such as title, author, and creation date, the connection among different works, authors, cultural influences, gender, and subjects, creating networks of knowledge that facilitate the discovery of new materials and the construction of relationships among documents.

There are several methodologies and approaches for data enrichment that make it possible to improve their quality and discoverability. According to Zeng (2018, p. 02), the methods are

[...] using knowledge organization systems (KOS) vocabularies and other resources that have embraced Linked Open Data (LOD); transforming semi-structured data into structured data through entity-based semantic analysis and annotation to extend access points; digging into unstructured data and generating knowledge bases to support knowledge discovery; enabling one-to-many usages of LAM data across data silos while delivering intuitive user interfaces online; and making the heterogeneous contents from diverse providers semantically interoperable via shared ontology infrastructure.

In this way, the semantic enrichment of metadata is presented as an approach to ensure the quality and accessibility of information, promoting interoperability among different data sources and expanding their discoverability. The adoption of structured methodologies and specialized tools in this process enhances the organization of metadata and supports decision-making in various informational contexts. Thus, the continuous development and application of these techniques represent a significant advancement for information management, enabling precise and contextualized retrieval of bibliographic data.

3 METHODOLOGICAL PROCEDURES

The approach of this research is descriptive and qualitative, as it seeks to characterize the scientific production on the semantic enrichment of bibliographic metadata. According to Gil (2002, p. 42), descriptive research aims at “the description of the characteristics of a given population or phenomenon or then establishment of relationships between variables.” Thus, this study describes the existing literature without intending to test hypotheses or propose explanatory models.

The literature search was carried out using the Open Knowledge Maps tool, a free and open-source platform that facilitates navigation through scientific articles. The search was conducted on August 10, 2024, using the keywords “semantic enrichment” AND “metadata.” The system generated a knowledge map with the 100 most relevant articles, covering all disciplines but filtered to include only scientific articles. The selection of the 100 most relevant articles was performed automatically by the tool, and the results are organized into thematic groups (clusters) that highlight areas of concentration on the topic.

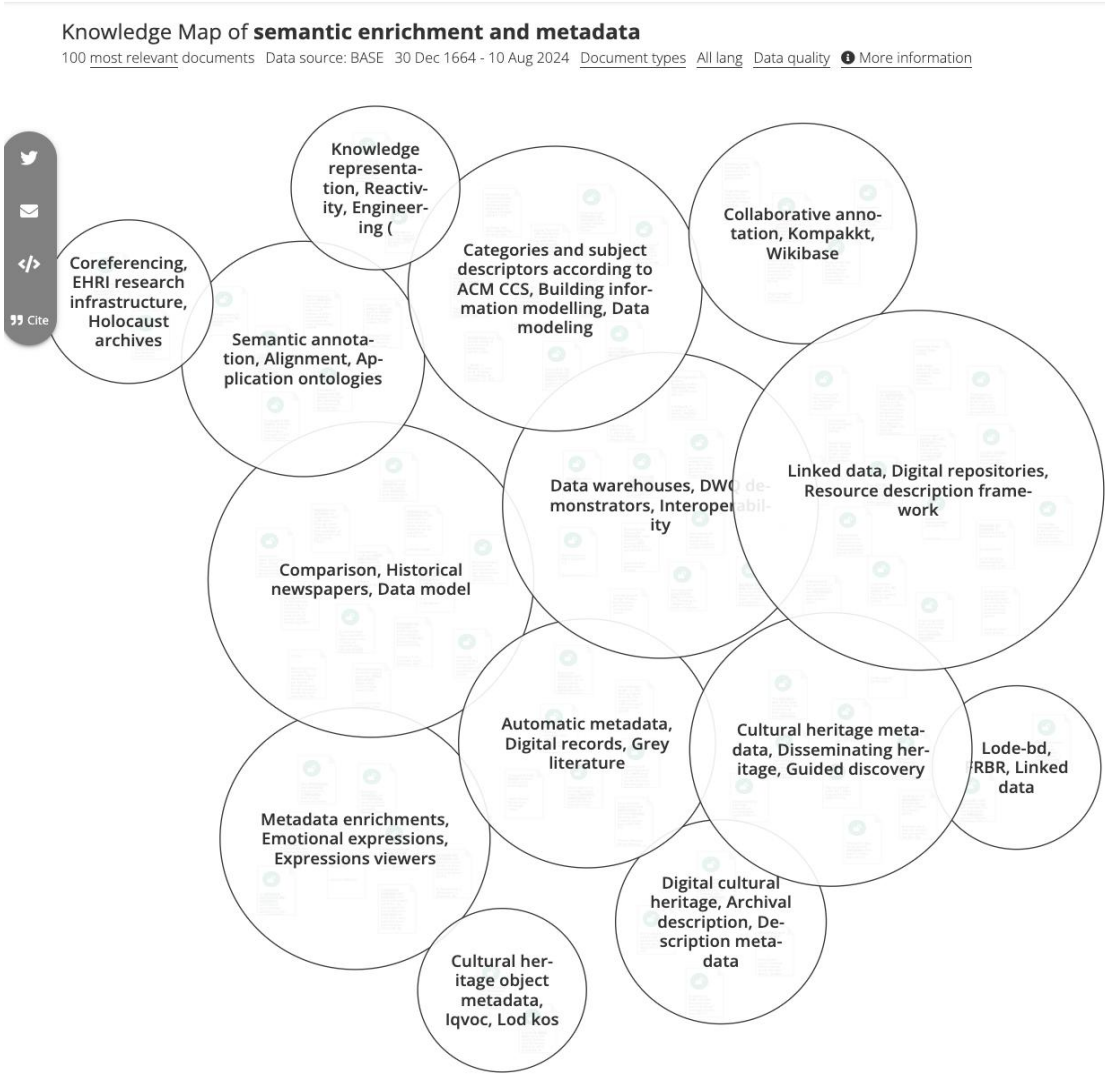
In addition to the search conducted in Open Knowledge Maps, the following databases were consulted: *Base de Dados em Ciência da Informação* (BRAPCI), Google Scholar, and Web of Science, to complement the literature review. The search was conducted using the keywords “semantic enrichment,” “data,” and “metadata,” considering only full articles, published between 2013 and 2024, and indexed in recognized scientific databases. The inclusion criteria considered articles in Portuguese and English that directly addressed the theme of semantic enrichment applied to bibliographic metadata.

4 RESULTS

Using Open Knowledge Maps, 100 articles considered the most relevant on semantic enrichment and metadata in different contexts were retrieved. The publications analyzed cover the period from 2000 to 2024, reflecting the evolution of approaches and tools used over the years. This survey highlights the diversity of applications of semantic enrichment, encompassing areas such as cultural heritage, digital libraries, biomedical data, and geospatial data, among others.

The articles were organized into 14 thematic groups, clustered according to the similarity of their approaches. Figure 1 presents the distribution of these groups, providing a clearer view of the main trends and research directions in the field.

Figure 1 - Knowledge map of semantic enrichment



Source: Open Knowledge Maps (2024).

The thematic groups (clusters) and a brief description of each of each are presented in Table 1 below.

Table 1 – Thematic groups (clusters)

Thematic group	Description
Automatic metadata, Digital records, Grey literature	Brings together documents on the automatic generation of metadata and semantic enrichment, with a focus on digital libraries and grey literature. The approaches include the use of ontologies, metadata recognition, and the migration of relational data.
Metadata of cultural heritage objects, IQVOC, LODKOS	Brings together documents on the semantic enrichment of metadata for university repositories, which include strategies to improve the use and role of repositories of cultural heritage objects.
Metadata enrichments, Emotional expressions, Expression viewers	Groups documents that address tools for the semantic enrichment of metadata, dealing with topics such as semantic search, integration of datasets, and emotional expressions in metadata systems.
Digital cultural heritage, Archival description, Descriptive metadata	Brings together documents that address improving the quality of cultural and archival metadata through crowdsourcing and digital technologies. They also discuss the Europeanization of cultural institutions and the publication of open data related to cultural heritage.
Cultural heritage metadata, Heritage dissemination, Guided discovery	Groups documents that address the enrichment of metadata for guided discovery and the integration of linked data for the management of cultural, linguistic, and musical heritage.
LODE-BD, FRBR, Linked Data	Brings together documents that assess issues related to metadata management in digital libraries, focusing on key concepts and perspectives for improving interoperability and the management of online resources through linked data and standards such as FRBR.
Linked Data, Digital repositories, Resource Description Framework (RDF)	Brings together documents on linked data and metadata in digital repositories, exploring semantic enrichment, system interoperability, and improvements in information access and management.
Data warehouses, DWQ demonstrations, Interoperability	Addresses research on metadata interoperability, data warehouses, and Data Warehouse Quality (DWQ) demonstrations, as well as semantic annotations, integration of services and data in domains such as biomedical and genomic data, authority control tools, and methods to improve interoperability between systems.
Comparison, Historical newspapers, Data model	Brings together studies focused on the semantic enrichment of scientific and historical content, exploring how metadata can be used to improve the retrieval and analysis of data in areas such as the life sciences and academic publications.
Collaborative annotation, Kompakkt, Wikibase	Addresses collaborative annotation and semantic enrichment of media, integrating open-source tools and knowledge graphs for the exchange and modeling of cultural annotations and 3D data.
Subject categories and descriptors according to ACM CCS, Information modeling, Data modeling	Brings together research on semantic enrichment and data modeling in contexts such as geographic information and cultural heritage. The studies address techniques to enhance the representation and integration of data in digital systems through semantic annotations.

Knowledge representation, Reactivity, Engineering	Brings together studies on languages for expressing reactivity and modeling dynamic behavior on the Semantic Web, focusing on knowledge representation and the engineering of interactive systems.
Semantic annotation, Alignment, Application ontologies	Brings together research on semantic annotation, ontology alignment, and their use in applications, with a focus on enriching spatial data, processing factual answers, and modeling geospatial information.
Coreferencing, EHRI research infrastructure, Holocaust archives	Addresses the semantic integration of Holocaust archives, focusing on the EHRI research infrastructure. The articles explore referencing and interoperability techniques to unify access to dispersed historical sources.

Source: Elaborated by the authors (2024).

The clusters bring together specific issues related to the semantic enrichment of metadata. For example, the cluster “Automatic metadata, Digital records, Grey literature” focuses on the automatic generation and migration of data in digital libraries and grey literature, whereas the cluster “Metadata of cultural heritage objects, IQVOC, LODKOS” explores the enrichment of metadata in cultural heritage repositories using ontologies and linked data. The cluster “Metadata enrichments, Emotional expressions, Expression viewers,” in turn, discusses the integration of emotional expressions into viewers and semantic search tools.

Subsequently, documents that address the enrichment of bibliographic data were selected. As a result, 16 documents address this theme, which are grouped in ascending order by date in Table 2.

Table 2 – Articles that address the enrichment of bibliographic data

Author and year	Title	Description
Shotton <i>et al.</i> (2009)	Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article.	The article explores how to enhance the accessibility and reuse of bibliographic data in scientific articles by applying semantic enhancements. Active DOIs, semantic markups, interactive figures, and downloadable data are integrated, promoting a richer and more connected research experience.
Solodovnik (2011)	Metadata issues in Digital Libraries: key concepts and perspectives.	The article highlights the importance of metadata standardization and the use of controlled vocabularies to improve the management of bibliographic data. It also emphasizes semantic enrichment, which enhances the description and connection of these data on the Web.
Schaible; Mayr (2012)	Discovering Links for Metadata Enrichment on Computer Science Papers.	The article proposes a way to add information to initial bibliographies that usually contain only basic metadata such as title, authors, and publication date. Using the Silk tool, the study connects computer science records to external sources such as DBLP and ACM, facilitating the discovery and completion of additional bibliographic data.

Vlachidis <i>et al.</i> (2013)	Automatic metadata generation in an archaeological digital library: Semantic annotation of grey literature.	The article addresses the automatic generation of rich metadata for archaeological excavation reports, using CIDOC CRM as the core ontology. Metadata are extracted and semantically enriched, improving search and interoperability of data through annotations in XML and RDF.
Leiva-Mederos <i>et al.</i> (2013)	AUTHORIS: a tool for authority control in the Semantic Web.	The article presents AUTHORIS, a tool that improves authority control and the enrichment of bibliographic data using learning rules and Linked Data. Based on standards such as Dublin Core and RDA, the tool facilitates disambiguation and the efficient reuse of bibliographic records.
Anibaldi <i>et al.</i> (2015)	Migrating bibliographic datasets to the Semantic Web: The AGRIS case.	AGRIS is a global catalog with 5 million structured bibliographic records indexed by the AGROVOC thesaurus. After its conversion to SKOS and publication as Linked Open Data (LOD), the set of records was tripled and made available as LOD.
Decourselle (2016)	Towards a Pattern-based Semantic Enrichment of Bibliographic Entities.	The article addresses challenges related to the process of metadata migration and semantic enrichment of bibliographic data.
Brisebois (2017)	A semantic metadata enrichment software ecosystem (SMESE): its prototypes for digital libraries, metadata enrichments and assisted literature reviews.	The article presents SMESE V1, a semantic metadata enrichment ecosystem for digital libraries that integrates data mining, machine learning, and a mapping ontology to enhance the description and usability of bibliographic data.
Brisebois <i>et al.</i> (2017)	A semantic metadata enrichment software ecosystem based on topic metadata enrichments.	The article presents the design, implementation, and evaluation of the SATD (Scalable Annotation-based Topic Detection) model and algorithm using web metadata, linked open data, and matching rules.
Candela <i>et al.</i> (2018)	A linked open data framework to enhance the discoverability and impact of culture heritage.	The article proposes a framework for enriching and disambiguating places in text using Wikidata and GeoNames, successfully applied in the Miguel de Cervantes Virtual Library, demonstrating how semantic enrichment improves information retrieval.
Al-Natsheh, <i>et al.</i> (2018)	Metadata Enrichment of Multi-Disciplinary Digital Library: A Semantic-based Approach.	The article works with a standard scientific categorization and a sample of articles tagged with topics to find semantically relevant articles and enrich their metadata accordingly.
Arakaki <i>et al.</i> (2019)	Availability of metadata in linked data for digital repositories.	The article aims to discuss possibilities of relationships among datasets and the semantic enrichment of metadata in digital repositories, and to present a model for converting records into RDF.
Bellotto; Bekesi (2019)	Enriching metadata for a university repository by modelling and infrastructure: a new vocabulary server for phaidra.	This article illustrates an initial step toward the semantic enrichment of the University of Vienna's Phaidra repository.
Labropoulou <i>et al.</i> (2023)	The LexMeta Model for Lexical Resources: Theoretical and Implementation	The article presents LexMeta, based on the FRBR and META-SHARE models, which aims to improve semantic interoperability and the exchange of metadata among catalogs. Implemented in RDF/OWL and as a Wikibase

	Issues.	ontology, LexMeta facilitates the semantic enrichment and interoperability of bibliographic data.
Massariet al. (2024)	OpenCitations Meta.	OpenCitations Meta is an open bibliographic metadata database that uses Semantic Web technologies to ensure interoperability and semantic enrichment of data. It assigns persistent identifiers (OMIDs), eliminates external dependencies, and improves the reliability and performance of the data.
Ibañez; Giménez-Tol edo(2024)	Exploring named-entity recognition techniques for academic books.	The article proposes a methodology that applies named-entity recognition to academic books in order to semantically enrich the data, improving information discovery and querying through an ontological framework on the Semantic Web.

Source: Elaborated by the authors (2024).

Shotton et al. (2009) demonstrate, through the semantic enrichment of an article from *PLoS Neglected Tropical Diseases*, how Semantic Web technologies can improve access, reuse, and interoperability of scientific information. The proposal resulted in an enriched version of the article, with the integration of DOIs, term markup, interactive figures, and reusable data. Despite technical limitations such as the absence of structured metadata, the need for manual curation, lack of integration with editorial systems, and restrictions on access to sensitive data, the potential of semantic publishing was validated, as well as the need for infrastructure, standardization, and tools for its large-scale adoption.

Solodovnik (2011), in turn, discusses the importance of the standardization of bibliographic metadata in digital libraries. The author highlights the use of controlled vocabularies, ontologies, and metadata standards to improve the management of bibliographic data in digital environments and on the Semantic Web. The models LODE-BD and SWAP are analyzed as applications aimed at the publication and description of bibliographic data in digital environments and on the Semantic Web. The author identifies as challenges the fragmentation of standards, the conversion of legacy data, and the lack of unified mechanisms for data publication and integration.

Schaible and Mayr (2012) propose the use of the Silk tool to enrich bibliographic metadata by means of links to external sources such as DBLP, ACM, and the Semantic Web Conference Corpus. Two approaches are explored: the addition of links to records and the direct incorporation of data. Silk proved to be functional, although it requires prior knowledge of RDF schemas and presents difficulties in handling large volumes of data. Nonetheless, the authors emphasize the feasibility of the approach and suggest its application in other domains.

Vlachidis et al. (2013) present a method for the automatic generation of semantic metadata for archaeological excavation reports in grey literature from the Archaeological Data Service (OASIS), integrating CIDOC CRM, CRM-EH, and Natural Language Processing (NLP). The methodology comprises the stages of preprocessing, information extraction, and conversion of data to RDF format, ensuring compliance with semantic standards. The evaluation carried out by experts confirmed the accuracy and retrieval capacity of the information, showing that the method improves access to and integration of archaeological data.

Subsequently, Leiva-Mederos et al. (2013) present AUTHORIS, a tool based on Linked Data and Semantic Web technologies for the generation and conversion of authority records, with the aim of improving authority control and promoting interoperability between bibliographic and non-bibliographic entities. Tested with records from 32 libraries and publishers, the tool demonstrated satisfactory performance, with a higher occurrence of errors in the ISAD(G) and EAD formats. The study also highlights limitations of the Virtual International Authority File (VIAF), such as restricted access, low flexibility, and the exclusion of non-library organizations, underscoring the need for open, inclusive, and collaborative models for authority control on the Web.

In the same context, Anibaldi et al. (2015) report the conversion of the International Information System for the Agricultural Sciences and Technology (AGRIS) catalog, which contains 5 million bibliographic records, to Linked Open Data (LOD) by modeling the records in RDF. The initiative involved the creation of persistent URIs, semantic enrichment, and the development of Open AGRIS, an application that integrates external data via SPARQL and Web services. The main challenges included the absence of administrative metadata to address provenance and trust, as well as limitations in semantic interoperability and in the availability of RDF services in the agricultural domain.

Decourselle (2016) proposes a pattern-oriented model to migrate and enrich bibliographic records based on the FRBR model. The approach facilitates the transformation of Machine Readable Cataloging (MARC) data into FRBR entities by identifying recurring bibliographic patterns. Implemented in the CoM3T tool, it enables the automation of part of the process, including analysis and extraction. The main challenges involve the complexity of MARC, the creation of specific rules, the

scarcity of evaluation metrics, and limitations in merging and enriching with external data.

Brisebois (2017) proposes the Semantic Metadata Enrichment Software Ecosystem (SMESE V1) for digital libraries, based on a metadata-oriented software architecture and the software product line model. The proposal integrates text and data mining components, machine learning, and ontologies to support the mapping and semantic enrichment of documents originating from the Web.

Building on the previous study, Brisebois et al. (2017) propose a software ecosystem for semantic metadata enrichment based on topic detection, called SATD. The model combines textual analysis, co-occurrence graphs, and semantic relations extracted from ontologies and thesauri, enabling the identification of topics through similarity with annotated documents. The main challenges identified concern the limitation of extraction based solely on keywords, the absence of latent co-occurrences, the lack of incorporation of structured semantic knowledge, and the difficulty of balancing accuracy with computational performance.

Candela et al. (2018), in turn, propose a framework to enrich the identification of places in bibliographic records, using resources such as Wikidata and GeoNames, with an application in the Miguel de Cervantes Virtual Library. They faced challenges such as data heterogeneity, which required adjustments to ensure compatibility among different formats; linguistic variations, which hindered the automatic matching of terms; and limitations in the availability and accuracy of external data, which affect the quality of the enrichment.

Al-Natsheh et al. (2018) apply a scientific categorization to enrich metadata in a multidisciplinary digital library. The approach combines supervised classification with semantic vectors extracted from titles and abstracts and query expansion based on synonyms obtained from the BabelNet network. The fusion of results makes it possible to assign multiple thematic labels per article, even in contexts with varied vocabularies across disciplines. Evaluated with data from the ISTEEX library, the proposal outperforms traditional methods such as supervised LDA. The main challenges involve terminological heterogeneity, the limitations of semantic networks for technical jargon, and the incompleteness of test data for evaluation.

Arakaki et al. (2019) propose a model for converting data from digital repositories to Linked Data, focusing on improving interoperability and the semantic

enrichment of metadata. The approach identifies databases, vocabularies, and standards compatible with the principles of the Semantic Web, structuring the process into stages of extraction, transformation to RDF, and enrichment.

Bellotto and Bekesi (2019) describe the semantic enrichment of the Phaidra repository at the University of Vienna through the adoption of a new controlled vocabulary in RDF aligned with LOD principles. The integration with external vocabularies and the implementation of structured query mechanisms expanded interoperability with other platforms. To ensure the consistency and evolution of the vocabulary, the authors highlight the need for maintenance guidelines, including rules for editing, term inclusion, and supporting documentation.

Labropoulou et al. (2023) present the LexMeta model for the description of lexical resources in linguistic and bibliographic catalogs, based on the FRBR and META-SHARE models. The proposal structures metadata into three main classes, allowing the representation of different versions and formats of the same resource. The model incorporates controlled vocabularies, such as LexVoc, to classify aspects such as entry type and microstructure of the content. Implementation occurs in two formats, RDF/OWL and Wikibase, with automated workflows for synchronization. The main challenges involve interoperability among heterogeneous models, the balance between flexibility and standardization of values, the need for extensibility, and adoption by communities with distinct descriptive practices.

Massari et al. (2024) present OpenCitations Meta, a bibliographic metadata database of academic publications indexed by OpenCitations. The proposal adopts its own persistent identifiers to disambiguate publications with multiple external identifiers and to include documents without DOI or PMID, expanding citation coverage. The process involves automated curation, enrichment, error correction, conversion to RDF, and the use of SPAR ontologies. The main challenges concern the absence of identifiers in certain areas of knowledge, ambiguity among multiple identifiers, low quality of source data, disambiguation of authors without ORCID, interoperability among formats, and the scaling of curation activities.

Finally, Ibañez and Giménez-Toledo (2024) explore the use of named-entity recognition (NER) to improve the querying and discovery of information in academic books in Spanish, with the aim of increasing the retrievability and visibility of content in the Humanities and Social Sciences. The methodology involves converting PDF to

XML, extracting and indexing paragraphs, and mapping the recognized entities to an ontology developed to represent books, authors, sections, paragraphs, and named entities. The entities are published as RDF triples and made available in a SPARQL endpoint. The main challenges concern the unavailability of full texts, the need for robust language models for each language, ambiguity among entities, linking with knowledge graphs, the technical complexity of mappings, and the limited adoption of ontologies in the publishing sector.

As pointed out by Zeng (2018), semantic enrichment can be carried out through different methods and strategies, reflecting the diversity of approaches applied to add meaning and context to bibliographic data. The 16 studies analyzed demonstrate the impact of applying different techniques, models, and tools.

Linked Data is one of the most frequent approaches, enabling the linking of records to external sources. Anibaldi et al. (2015) apply it to the AGRIS catalog, while Schaible et al. (2012) use the Silk tool to enrich data from scientific articles. Arakaki et al. (2019) also propose the conversion of metadata from institutional repositories with a focus on interoperability. In addition, Massari et al. (2023) develop a semantic database that centralizes metadata of scientific publications and maintains change tracking with version control in RDF.

The use of ontologies and controlled vocabularies stands out in initiatives such as that of Leiva-Mederos et al. (2013), who propose AUTHORIS for authority control, and Bellotto and Bekesi (2019), who implement the iQvoc vocabulary in the Phaidra repository to improve the semantic organization of records. Labropoulou et al. (2023) present LexMeta as a model oriented toward the description of lexical resources, aiming at interoperability between bibliographic and linguistic catalogs.

Approaches based on semantic annotation and named-entity recognition are also explored. Ibañez et al. (2024) apply these techniques to academic books, using trained language models to extract entities and structure the information in an ontology. Strategies oriented toward topic classification and labeling, such as that proposed by Al-Natsheh et al. (2018), and toward the detection of emerging topics through semantic networks, as in the model of Brisebois et al. (2017), show the capacity of semantic enrichment to expand the possibilities of discovery and reuse of scientific content.

These studies also point to challenges in the process of semantic enrichment of metadata, such as the heterogeneity of source formats, the need for alignment among different models and vocabularies, and the limited access to the full text of documents. Issues related to entity disambiguation, large-scale data curation, interoperability among systems, and the technical complexity involved in modeling and publishing data in structured formats are also highlighted.

As observed by Arakaki and Isotani (2024), decision-making often occurs based on incomplete information or individual experience, a context in which the semantic enrichment of metadata can play a key role by providing more structured and accurate support for information management.

5 CONCLUSIONS

Semantic enrichment of metadata represents an advance in improving descriptive quality and interoperability among systems. More than the simple addition of information, it involves the creation of semantic layers that make data more understandable, linkable, and reusable. As evidenced in the publications analyzed, there is a growing diversity of approaches, with applications ranging from cultural heritage and geospatial data to biomedical domains, demonstrating the interdisciplinary nature of this practice.

In the bibliographic context, semantic enrichment contributes to improving information retrieval by enabling more precise and context-sensitive searches. Enriched records provide more complete data, supporting deeper and more well-founded analyses. This qualification of metadata benefits decision-making in scientific, administrative, and cultural environments.

In addition, the adoption of tools based on ontologies and Linked Data promotes the integration of heterogeneous sources and improves information management, especially in view of the growing volume of data. In this sense, semantic enrichment is consolidated as a relevant process for making information more accessible, interoperable, and meaningful. Ethical issues involved in the semantic enrichment process, although relevant, will not be addressed at this time.

REFERENCES

AL-NATSHEH, H. T. *et al.* Metadata enrichment of multi-disciplinary digital library: a semantic-based approach. In: INTERNATIONAL CONFERENCE ON THEORY AND PRACTICE OF DIGITAL LIBRARIES, 2018, Cham. **Proceedings...** Cham: Springer International Publishing, 2018. p. 32–43. Available at: https://doi.org/10.1007/978-3-030-00066-0_4. Accessed on: 10 dez. 2025.

ANIBALDI, S. *et al.* Migrating bibliographic datasets to the Semantic Web: The AGRIS case. **Semantic Web**, Amstredã, v. 6, n. 2, p. 113–120, 2015. Available at: <https://doi.org/10.3233/SW-140148>. Accessed on: 10 dez. 2025.

ARAKAKI, A. C. S.; ISOTANI, S. Metadados e a tomada de decisão com base em evidências. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA – WIDAT, 6., 2023, Brasília. **Anais...** Brasília, v. 6, 2023. DOI: <https://doi.org/10.22477/vi.widat.52>. Available at: <https://labcotec.ibict.br/widat/index.php/widat2023/article/view/52>. Accessed on: 10 dez. 2025.

ARAKAKI, F. A. *et al.* Availability of metadata in linked data for digital repositories. **Anales de Documentación**, Murcia, v. 22, n. 2, 2019. Available at: <https://doi.org/10.6018/analesdoc.327441>. Accessed on: 10 dez. 2025.

BELLOTTO, A.; BEKESI, J. Enriching metadata for a university repository by modelling and infrastructure: a new vocabulary server for Phaidra. **Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare**, Viena, v. 72, n. 2, p. 446–459, 2019. Available at: <https://journals.univie.ac.at/index.php/voebm/article/view/5403> Accessed on: 10 dez. 2025.

BRISEBOIS, R. **A Semantic Metadata Enrichment Software Ecosystem (SMESE):** its prototypes for Digital Libraries, metadata enrichments and assisted literature reviews. 2017. Tese (Doutorado) – École de technologie supérieure, Montreal, 2017. Available at: <https://espace.etsmtl.ca/id/eprint/1967/> . 1 Accessed on: 10 dez. 2025.

BRISEBOIS, R.; ABRAN, A.; NADEMBEGA, A.; N'TECHOBO, P. A semantic metadata enrichment software ecosystem based on topic metadata enrichment. **International Journal of Data Mining & Knowledge Management Process**, Chennai, v. 7, n. 3, p. 1–23, 2017. DOI: <https://doi.org/10.5121/ijdkp.2017.7301> . Accessed on: 10 dez. 2025.

CANDELA, G. *et al.* A linked open data framework to enhance the discoverability and impact of cultural heritage. **Journal of Information Science**, Londres, v. 45, n. 6, p. 756–766, 2019. DOI: <https://doi.org/10.1177/0165551518812658> . Accessed on: 10 dez. 2025.

DECOURSELLE, J. Towards a pattern-based semantic enrichment of bibliographic entities. **IEEE TCDL Bulletin**, Piscataway, v. 12, n. 2, 2016. Available at: <http://www.ieee-tcdl.org/Bulletin/v12n2/papers/decourselle.pdf>. Accessed on: 10 dez. 2025.

IBAÑEZ, P. C.; GIMÉNEZ-TOLEDO, E. Exploring named-entity recognition techniques for academic books. **Learned Publishing**, Oxford, v. 37, n. 3, p. e1610, 2024. DOI: <https://doi.org/10.1002/leap.1610>. Accessed on: 10 dez. 2025.

ISAAC, A.; MANGUINHAS, H.; STILLER, J.; CHARLES, V. **Report on enrichment and evaluation**. Haia: Europeana Task Force on Enrichment and Evaluation, 2015. Available at: http://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/FinalReport_EnrichmentEvaluation_102015.pdf. Accessed on: 10 dez. 2025.

LABROPOULOU, P. *et al.* The Lexmeta Model for Lexical Resources: Theoretical and Implementation Issues. **Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje**, Zagreb, v. 49, n. 1, p. 137–154, 2023. Available at: <https://hrcak.srce.hr/292655>. Accessed on: 10 dez. 2025..

LEIVA-MEDEROS, A. *et al.* Authoris: a tool for authority control in the semantic web. **Library Hi Tech**, Bradford, v. 31, n. 3, p. 536–553, 2013. DOI: <https://doi.org/10.1108/LHT-12-2012-0135>. Accessed on: 10 dez. 2025.

MASSARI, A. *et al.* OpenCitations Meta. **Quantitative Science Studies**, Cambridge, v. 5, n. 1, p. 50–75, 2024. DOI: https://doi.org/10.1162/qss_a_00292. Accessed on: 10 dez. 2025.

OPEN KNOWLEDGE FOUNDATION. **Open definition**. Cambridge, [s. d.]. Disponível em: <https://opendefinition.org>. Accessed on: 10 dez. 2025.

OPEN KNOWLEDGE MAPS. **Knowledge map for research on semantic enrichment and metadata**. Viena, 2024. Available at: <https://openknowledgemaps.org/map/3e05c9f29d39a3d6d1f55bd1fd610acf?area=1>. Accessed on: 10 dez. 2025.

POMERANTZ, J. **Metadata**. Cambridge: The MIT Press, 2015. 239 p.

SANTOS, P. L. V. A. C.; SANT'ANA, R. C. G. Dado e granularidade na perspectiva da informação e tecnologia: uma interpretação pela ciência da informação. **Ciência da Informação**, Brasília, v. 42, n. 2, 2013. Available at: <https://brapci.inf.br/index.php/res/v/17173>. Accessed on: 10 dez. 2025.

SCHAIBLE, J.; MAYR, P. Discovering links for metadata enrichment on computer science papers. **arXiv**, Ithaca, NY, 2012. (Preprint). Available at: <https://arxiv.org/abs/1212.3677>. Accessed on: 10 dez. 2025.

SHOTTON, D. *et al.* Adventures in semantic publishing: exemplar semantic enhancements of a research article. **PLoS Computational Biology**, São Francisco, v. 5, n. 4, p. e1000361, 2009. DOI: <https://doi.org/10.1371/journal.pcbi.1000361>. Accessed on: 10 dez. 2025.

SOLODOVNIK, I. Metadata issues in digital libraries: key concepts and perspectives. **JLIS.it: Italian Journal of Library, Archives and Information Science**, Florença, v. 2, n.

2, p. 1–27, 2011. DOI: <https://doi.org/10.4403/jlis.it-4663> . Accessed on: 10 dez. 2025.

VLACHIDIS, A. *et al.* Automatic metadata generation in an archaeological digital library: semantic annotation of grey literature. In: PRZEPIÓRKOWSKI, A.; PIASECKI, M.; JASSEM, K.; FUGLEWICZ, P. (ed.). **Computational Linguistics**. Heidelberg: Springer, 2013. p. 187–202. (Studies in Computational Intelligence, v. 458). DOI: https://doi.org/10.1007/978-3-642-34399-5_10 . Accessed on: 10 dez. 2025.

ZENG, M. L. Semantic enrichment for enhancing LAM data and supporting digital humanities. **Profesional de la Información**, [S. l.], v. 28, n. 1, 2019. DOI: <https://doi.org/10.3145/epi.2019.ene.03>. Accessed on: 10 dez. 2025.

ZENG, M. L.; QIN, J. **Metadata**. Nova Iorque: Neal-Schuman Publishers, 2008.

ARTICLE NOTES AND CREDITS

- Acknowledgements: The authors would like to thank the Brazilian Institute of Information in Science and Technology (IBICT).
- Funding: This study was funded by the Brazilian Institute of Information in Science and Technology (IBICT) and the National Council for Scientific and Technological Development (CNPq).
- Conflicts of interest: Not applicable.
- Ethical approval: Not applicable.
- Data and materials availability: Not applicable.
- Manuscript published as a preprint: The manuscript was originally published as a full paper in the *National Seminar on Cataloging and Technology* (SNCat) in 2024. It subsequently underwent a new double-blind peer review process and received content revisions and updates. In addition, it has already been published in Portuguese in the journal *CIExpress*.
- Author contributions:

Contributions	MENDONÇA, A. C. N.	ARAKAKI, A. C. S
Study design	X	
Conceptualization	X	X
Metodology	X	X
Data collectio / investigation	X	
Data Curation	X	
Data analysis	X	
Discussion of the results	X	
Visualization (graphs, tables and others)	X	
Original draft	X	
Revision and final edit	X	X
Supervision and administration	X	X

- License to Use

The authors grant Ciência da Informação Express – CIExpress exclusive rights of first publication, with the work simultaneously licensed under the Creative Commons Attribution License (CC BY) 4.0 International. This license allows third parties to remix, adapt, and build upon the published work, attributing proper credit for authorship and initial publication in this journal.

- Publisher

Federal University of Lavras (UFLA).

The ideas expressed in this article are his own responsibility and do not necessarily represent the opinion of the editors or the university

Editor of the communication and scientific dissemination channel *Ciência da Informação Express*

Nivaldo Calixto Ribeiro, Universidade Federal de Lavras (UFLA).

Linguistics Reviewer

From the authors.

Reference Reviewer

From the authors.

English version

Ana Carolina Novaes de Mendonça

- Historical

Received on: 2024-11-14

Accepted on: 2025-04-14

Posted on: 2025-04-15