Semantic Mapping of Bibliographic Models in National Libraries

Felipe Augusto Arakaki[1,†], Ana Carolina Simionato Arakaki[1,2,*,†] and Ana Carolina Novaes de Mendonça[3,†]

[1] *Faculty of Information Science, University of Brasilia, Campus Darcy Ribeiro - DF, 70297-400 , Brasilia, Brazil*
[2] *Graduate program in Information Science, Federal University of São Carlos (UFSCar), São Carlos, Brazil*
[3] *Brazilian Institute of Science and Technology, SAUS Q 5, L 6, Bl H, Brasília, DF, Brazil*

Abstract

This paper presents a comparative analysis of the data and metadata models used by national libraries that publish their collections following the principles of Linked Open Data (LOD). Using the Crosswalk method, the study mapped the classes and properties adopted by different institutions, identifying conceptual convergences, terminological variations, and distinct modeling strategies. The results reveal heterogeneous practices, ranging from simplified models to robust, ontology-based structures. Despite shared core entities such as Work and Person, semantic differences remain regarding the scope and level of detail. The findings support the development of semantic alignment strategies and shared vocabularies to improve interoperability.

Keywords
Linked Data, Semantic Web, Metadata, Ontology, Library.

1. Introduction

The provision of structured bibliographic data on the Web, in an open and connected manner, represents one of the main contemporary challenges faced by libraries. Although libraries have traditionally adopted well-established representation standards, such as Machine-Readable Cataloging (MARC 21) developed by the Library of Congress in the United States, many of these models were originally designed for the construction of printed and centralized catalogs, which hinders their integration into digital ecosystems guided by the principles of the Semantic Web and Linked Open Data (LOD). In this context, transforming bibliographic records into connected data requires technological adjustments and a theoretical-methodological revision of the approaches to data modeling, description, and interoperability.

The publication of linked open data requires the use of persistent identifiers, standardized vocabularies, a consistent semantic structure, and machine-readable formats. Various international organizations, such as the World Wide Web Consortium (W3C), have promoted best practices for publishing linked data, aiming to build an interoperable, accessible, and reusable Web of Data. However, in the bibliographic domain, the heterogeneity of data models adopted by different institutions, as well as the scarcity of initiatives that effectively publish their data in accordance with these principles, reveal the complexity of transitioning from legacy systems to LOD-oriented architectures.

This article investigates the issue of structuring bibliographic data in national libraries based on the principles of LOD. The aim is to identify the data and metadata models used by libraries that

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.
[1,*] Corresponding author.
[†] These authors contributed equally.
felipe.arakaki@unb.br (F. A. Arakaki); ana.arakaki@unb.br (A. C. S. Arakaki); anamendonca@ibict.br (A. C. N. Mendonça)
0000-0002-3983-2563 (F. A. Arakaki); 0000-0002-0140-9110 (A. C. S. Arakaki); 0009-0004-0285-9932 (A. C. N. Mendonça)

already publish their collections as connected data, analyzing their classes, properties, and representation strategies. The research focuses on the systematization of existing models through the application of the Crosswalk method, which enables comparative analysis and semantic mapping across the different standards in use.

The relevance of this study lies in providing support for the harmonization of bibliographic data at the international level, promoting greater interoperability, visibility, and reuse of national collections. By identifying convergences and divergences among the models in use, the study seeks to understand how libraries have been adapting their descriptive structures to the context of linked data, and which methodological paths can be followed by institutions that have not yet adopted this publication model.

2. Related works

The proposal for structuring and publishing bibliographic data on the Web is situated within the broader context of the Semantic Web and the principles of LOD [7]. The Semantic Web aims to assign meaning to data available on the internet so that it can be understood and processed by computational agents, enabling automated integration, discovery, and reuse of information. LOD, in turn, refers to a set of best practices for publishing structured data that are interlinked by URIs and described using interoperable standards such as Resource Description Framework (RDF), Web Ontology Language (OWL), and SPARQL Protocol and Resource Description Framework Query Language (SPARQL) [8,9].

In the field of Library and Information Science, the adoption of these principles requires the adaptation of traditional tools that support information representation such as MARC21 and the Universal Machine-Readable Cataloging (UNIMARC), developed by the International Federation of Library Associations and Institutions (IFLA) toward an ontology-driven structure. Initiatives such as BIBFRAME, developed by the Library of Congress, represent concrete efforts in this direction by proposing a model based on entities such as Work, Instance, and Item, aligned with controlled vocabularies and Semantic Web standards. Nevertheless, the diversity of models adopted by national libraries in different countries presents challenges for interoperability and semantic mapping.

It is important to highlight the risks and benefits of simplifying schemas for interoperability purposes: while such simplification enhances machine readability, it may lead to the loss of descriptive nuances, requiring critical decisions about which characteristics of entities should be preserved [1]. In this context, ontologies play a central role as tools for conceptual structuring, enabling the explicit representation of classes, properties, and relationships between domain entities, thus promoting consistency and reuse across heterogeneous systems [10].

The present research is therefore grounded in approaches that integrate the Semantic Web, Linked Data, and metadata schema mapping, aiming to understand the models adopted by national libraries and to propose pathways for interoperability and the publication of bibliographic data as LOD.

From a methodological standpoint, several studies propose strategies for mapping between metadata schemas, highlighting the importance of preserving semantic integrity during the conversion process [2]. Additionally, the challenges of achieving semantic interoperability are underscored by the diversity of elements and structural variations adopted by different information communities [1].

The literature also highlights the central role of ontologies in the representation of bibliographic data in digital environments. Studies provide guidelines for defining classes, properties, and relationships between entities [11], while others discuss the impacts of schema simplification on metadata quality and reusability [12].

3. Methodology

This research is characterized as qualitative, exploratory and theoretical, using the Crosswalk method, proposed by the National Information Standards Organization (NISO) in 1999, for data analysis. The Crosswalk method enables interoperability between systems that use heterogeneous metadata standards. According to NISO [1] "Crosswalks provide the ability to make the content of elements defined in a metadata standard available to communities using related metadata standards".

The Crosswalk method consists of four stages: harmonization, semantic map, element-to-element mapping, and hierarchy, object and logical view. To carry out the crosswalk, Chan and Zeng [2] highlight two approaches: "absolute crosswalking" and "relative crosswalking". Absolute crosswalking refers to the exact correspondence between metadata, while relative crosswalking is used to minimize information loss by matching elements of a source schema to at least one element of a target schema.

The crosswalk process can face equivalence difficulties, such as one-to-one, one-to-many, many-to-one and one-to-one. However, it was decided to carry out a general mapping, checking the compatibility of the classes and properties proposed by each of the institutions analyzed. The crosswalk was done individually for each metadata standard, followed by the creation of a table with the crosswalk of all the standards analyzed, establishing a general overview of the mapping.

In this context, a manual crosswalk was developed to map the metadata elements used by national libraries whose catalogs share connected open data. The task involved a detailed semantic analysis of each metadata element and its contextual application. Four specialists conducted this work based on the official documentation provided by each institution.

The selection of libraries was guided by a previous survey [3] that identified eleven national libraries publishing linked open data. However, among these eleven institutions, only seven were found to explicitly provide documentation regarding their data and metadata models, which enabled their comparative analysis within the scope of this study. The libraries analyzed were: the National Library of Spain (BNE), the National Library of France (BnF), the German National Library (DNB), the Finnish National Bibliography (FENNICA), the Royal Library of the Netherlands (KB), the Library of Congress (LC), and the National Library of Medicine (NIH).

4. Results and discussion

Based on the results in Table 2 and the specialized literature, it is possible to discuss the diversity of approaches adopted by National Libraries in modeling their bibliographic data in Linked Open Data (LOD) environments. It should be noted that not all classes and proprieties identified in the models are detailed in the text, but only a selection.

The Library of Congress, represented through the BIBFRAME model, was developed to replace the MARC21 format and constitutes an entity-based structure aligned with the principles of the Semantic Web. Among the institutions analyzed, it was the library that showed the highest degree of correspondence with the other models, particularly through the adoption of core classes such as Work and Person. This indicates a baseline convergence around essential elements of bibliographic description. The BIBFRAME model currently comprises 208 classes and 148 properties, which reflects a robust and detailed semantic framework. The observed overlap suggests partial alignment between the models in use, although differences persist regarding semantic granularity and the inclusion of additional entities such as Expression, Contribution, and Dataset, which vary depending on institutional contexts and implementation strategies.

The National Library of Spain (BNE) The National Library of Spain (BNE) adopts the IFLA Library Reference Model (LRM) as the conceptual basis of its bibliographic data model. Its current structure includes 8 classes and 225 properties, reflecting a semantically rich approach oriented toward interoperability and reuse of bibliographic data [4].

Table 2
Mapping of entity classes in National Libraries with Open Data catalogs

BIBFRAME	BNE	BNF	DNB	FENNICA	KB	NIH
Agent/Person	Person	Person	Persons/ Entities	Person	-	Agent
Collection	-	-	-	-	-	Collection
-	-	Concept	-	Concept	-	-
Contribution	-	-	-	-	-	Contribution
Dataset	-	-	-	-	Dataset	Dataset
Item	Item	Item	-	-	-	-
Organization	Corporate Body	Collective agent	-	Organization	-	Agent
Place	-	Place	-	Place	-	-
Work	Work	Work	-	Work	-	-

The Bibliothèque Nationale de France (BnF) adopts a limited set of entity classes Person, Item, Place, and Work reflecting a more selective approach to semantic modeling. The main entities are grounded in the IFLA Library Reference Model (LRM). The BnF integrates different metadata standards such as Internal Format of the National Library (InterMARC)[5].

Among the national libraries analyzed, the Deutsche Nationalbibliothek (DNB) adopts a data model that groups entities under the broader class Persons/Entities, encompassing Person, Concept. The Finnish National Bibliography (FENNICA) [6] includes the classes Person, Organization, Place, Concept, and Work, which demonstrate a concern for semantic granularity and alignment with Linked Data principles.

In contrast, the Koninklijke Bibliotheek (KB) displays 7 classes and 50 properties in the mapped data. With emphasis on Dataset and Work, the limited appearance of other classes like Person or Organization may reflect a partial publication strategy or a modeling focus on specific types of bibliographic resources. This configuration constrains broader analysis of its ontological infrastructure. Lastly, the National Library of Medicine (NIH) adopts a domain-specific model oriented toward scientific and biomedical information. The mapped entities include Agent, Collection, Contribution, and Dataset, revealing a strong emphasis on collaborative authorship and data sharing.

The class mapping conducted offers valuable insight into how different national libraries and bibliographic frameworks structure knowledge and resources. It reveals not only a shared foundation based on core concepts, but also the distinctive modeling choices made by each institution to address its specific informational and operational contexts. The role of reference models such as IFLA-LRM is particularly significant, as it establishes a common baseline of entities that facilitates interoperability across diverse vocabularies and systems.

To further understand the modeling strategies adopted by these national libraries, it is essential to complement the analysis of entity classes with an examination of the properties employed. The analysis of the properties used in open linked data catalogs reveals not only differences in the granularity and descriptive scope adopted by each library but also varying levels of adherence to Semantic Web best practices. Properties such as hasPart, ISBN, and ISSN are widely present across the models analyzed, suggesting a common core of elements focused on the identification and

structuring of bibliographic relationships.

However, in the BIBFRAME model used by the Library of Congress, both ISBN and ISSN are treated as classes rather than properties. This means that these identifiers are modeled as autonomous entities with their own attributes and relationships, in line with the entity-oriented approach proposed by the model. This modeling choice reflects an effort to enhance semantic flexibility, allowing, for instance, different editions or versions of a publication to be associated with multiple identifiers while maintaining consistency across complex datasets. This conceptual distinction is critical, as it directly affects interoperability between models and the way data can be linked to other datasets on the Web of Data.

Another relevant aspect is the use of the Description property, present in the catalogs of FENNICA, KB, and NIH. This property, often associated with widely adopted vocabularies such as Dublin Core, indicates a concern with both human-readable records and semantic indexing. Meanwhile, the Date property, used by BNF, DNB, and FENNICA, underscores the importance of temporal elements in bibliographic control, which are essential for organizing editions, versions, and publication events.

Lastly, the hasPart property, recurring in all models analyzed except BIBFRAME (where the relationship is represented differently), demonstrates a broad recognition of the importance of hierarchical and compositional relationships in bibliographic records. This property is crucial for representing collections, volumes, chapters, or any composite structure of works, reinforcing the need for descriptive mechanisms that capture complex relationships among resources.

Taken together, these findings show that, although there is a minimal set of shared properties, modeling decisions such as elevating identifiers to entities or emphasizing certain descriptive properties vary according to technical and institutional contexts. This highlights the need for semantic mapping and alignment strategies to enable full interoperability between catalogs structured as Linked Open Data.

5. Conclusion

This study presented a comparative analysis of the data and metadata models used by national libraries that publish their collections according to Linked Open Data (LOD) principles. Based on the Crosswalk method, it was possible to map the classes and properties adopted by different institutions, identifying conceptual convergences, terminological variations, and specific strategies for structuring bibliographic data.

The results reveal a heterogeneous landscape in the models employed, ranging from simplified approaches, such as those of the Bibliothèque Nationale de France and the Biblioteca Nacional de España, to more complex structures like the Deutsche Nationalbibliothek. Amid this diversity, there is a growing trend toward the adoption of the BIBFRAME model, particularly among Anglophone libraries such as the Library of Congress, which indicates consistent efforts toward standardization and semantic interoperability on an international scale.

The mapping conducted revealed that, although many libraries share core classes and properties such as Work, Instance, Item, Agent, and Concept relevant semantic variations persist regarding their application, scope, and level of detail.

One of the limitations identified is the absence of formal validation of the mappings, which could be addressed in future studies with the support of specific ontology alignment tools. Despite these limitations, the findings contribute meaningfully to ongoing discussions about bibliographic data interoperability in contexts driven by the Semantic Web. The study provides concrete support for the development of shared vocabularies and harmonization strategies among bibliographic systems, assisting in national collection modernization and open access initiatives.

For future developments, it is recommended to create a unified vocabulary that can serve as a foundation for interoperability between institutions, supported by ontologies as central tools for semantic alignment across heterogeneous descriptive standards. Additionally, the use of artificial intelligence techniques such as machine learning and semantic inference offers promising potential to enhance the processes of discovery, reconciliation, and reuse of information across libraries, broadening the reach and effectiveness of data integration on the Web.

It can be concluded that harmonizing bibliographic data and metadata models is a fundamental step toward building a more connected, accessible, and standards-driven information ecosystem, significantly contributing to the strengthening of national bibliographic heritage in the era of linked data.

Acknowledgements

Acknowledgements of support from the National Council for Scientific and Technological Development (CNPq) for the projects: Linked data publishing in libraries: theoretical-methodological proposal for SIBISC, CNPq Universal nº 409407/2021-6, Connected authority data for libraries: theoretical and methodological proposal for SIBISC, CNPq Universal nº 421178/2023-0 and the Brazilian Institute of Information in Science and Technology (Ibict).

Declaration on Generative AI

The AI was used to evaluate the text and assist in the drafting of some paragraphs.

References

[1] Pierre, M. S., & LaPlant, W. P, Issues in crosswalking content metadata standards. [S.l.]: NISO Baltimore, Maryland, USA 2000.

[2] L. M. Chan and M. L. Zeng, Metadata interoperability and standardization: A study of methodology part I. D-Lib Magazine, 12(6), 1082–9873, 2006.

[3] A. F. de. Jesus, Recomendações teórico-metodológicas para a publicação de dados bibliográficos abertos e conectados (Dissertação de mestrado, Universidade Federal de São Carlos, UFSCar, São Carlos, SP), 2021. https://repositorio.ufscar.br/handle/ufscar/14228.

[4] Biblioteca Nacional de España. Ontología BNE (Rev. 2.0). 2020. Biblioteca Nacional de España. https://datos.bne.es/def/index-es.html.

[5] Bibliothèque nationale de France. (n.d.). Semantic Web and Data Model. Bibliothèque nationale de France. https://data.bnf.fr/semanticweb.

[6] Fennica. Fennica RDF data model, 2019. https://www.kiwi.fi/display/Datacatalog/Fennica+RDF+data+model.

[7] PomerantzM Metadata. USA: The MIT press essential knowledge series. 2015.

[8] T. Berners-Lee, Linked data: Design issues. 2006. https://www.w3.org/DesignIssues/LinkedData.html.

[9] B. Hyland, G. Atemezing, and B. Villazón-Terrazas, Best practices for publishing Linked Data, 2014. W3C Working Group Note. Disponível em https://www.w3.org/TR/ld-bp/.

[10] Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. (2008). Linked data on the web. In Proceedings of the 17th International Conference on World Wide Web (pp. 1265–1266). [*S.l.*]: ACM.

[11] S. Van Hooland and R. Verborgh, Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata. Facet Publishing. 2014

[12] N. F. Noy and D. L. McGuinness, Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05, 2001.