Data mapping in national libraries (abstract)

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

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

Abstract

Semantic Web technologies have provided solutions, offering significant opportunities for improving and optimizing information search and retrieval processes, especially in the bibliographic domain. Although the representation tools used in the bibliographic universe are already consolidated, many new information resources have been created, making some guidelines obsolete. Libraries are therefore faced with the challenge of revising their representation processes and tools to make them more suitable for the new technological demands. These adjustments and updates require a review and change of theoretical and methodological paradigms, such as the use of ontologies, which are essential for organizing and representing knowledge. Ontologies define structured frameworks using classes and properties, enabling libraries to make their data more interoperable and accessible. Currently, much bibliographic data is published in formats that limit interlinking with other datasets, restricting data accessibility. Berners-Lee's core principles, along with best practices from the World Wide Web Consortium (W3C), emphasize using URIs, standard protocols, and linked data to enable discovery and connection across datasets. Libraries hold vast amounts of valuable data, but existing systems often fail to make it accessible on the Web in an open, connected format. In Brazil, no institutions have yet fully implemented Linked Open Data (LOD) practices to address these challenges. Research problem: How to structure national library data based on the principles of connected open data. The aim is to map the data and identify the processes required for the publication of connected open data by libraries. The research is ongoing and is currently in the phase of identifying the data and metadata models used by each library. 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. It involves harmonization, semantic mapping, element-to-element mapping, and the organization of metadata hierarchically. The crosswalk process can encounter challenges, such as one-to-one, one-to-many, and many-to-one equivalences. To manage these challenges, a general mapping approach was adopted, focusing on the compatibility of the classes and properties proposed by each of the analyzed institutions. First, the metadata standards were mapped individually, followed by the creation of a comprehensive table that presents an overall view of all the standards. In this context, the metadata from national libraries that share linked open data were mapped. The research universe was based on a survey that identified eleven national libraries that publish linked open data, including the Bibliotheca Apostolica Vaticana (BAV), Biblioteca Nacional de España (BNE), Bibliothèque Nationale de France (BnF), British National Bibliography (BNB), Deutsche Nationalbibliothek (DNB), Finnish National Bibliography (FENNICA), Koninklijke Bibliotheek (KB), Library Information System of Swedish National Union (LIBRIS), Library of Congress (LC), National Library of Iran (NLAI), National Library of Medicine (NLM), and National Széchényi Library (NSZL).The mapping process began with the identification of each institution's metadata and its complexities. Property classes were then separated, and comparisons were made based on each schema's terms and the definitions provided by the institutions regarding classes and properties. These definitions helped with the semantic alignment of the terms. In some cases, the terminology and scope of the classes were

Proceedings of the 17th Seminar on Ontology Research in Brazil (ONTOBRAS 2024) and 8th Doctoral and Masters Consortium on Ontologies (WTDO 2024), Vitória, Brazil, October 7-10, 2024.

* Corresponding author.
† These authors contributed equally.
mnovaesana@gmail.com (A.C.N. Mendonça); felipe.arakaki@unb.br (F.A. Arakaki); fernanda.farinelli@unb.br (F. Farinelli); acsimionato@ufscar.br (A.C.S. Arakaki)
0009-0004-0285-9932 (A.C.N. Mendonça) 0000-0002-3983-2563 (F.A. Arakaki); 0000-0003-2338-8872 (F. Farinelli); 0000-0002-0140-9110 (A.C.S. Arakaki)

identical, which facilitated the creation of an absolute crosswalk. In more complex cases, where different terminologies but similar scopes of use were observed, a relative crosswalk was applied. Results: The results so far show that libraries have specialized metadata, reflecting their particularities, but varying in use and description. Preliminary results show that while most libraries share core classes, variations in terminology and structure exist. For example, BIBFRAME, presents a more granular and detailed classification system with numerous subclasses that differentiate between various bibliographic aspects. This contrasts with institutions like the National Library of Medicine (NHI), which uses more specialized classes focused on medical concepts, reflecting its domain-specific needs. Similarly, some libraries, like the Biblioteca Nacional de España (BNE), prioritize metadata classes related to biographical descriptions and associated images, utilizing external sources such as Wikipedia for these elements. On the other hand, the Koninklijke Bibliotheek (KB) in the Netherlands emphasizes classes related to access types, such as those governing the distribution and access of digital resources. In terms of properties, there are also significant variations. For example, the National Library of Medicine uses properties like accrual periodicity to indicate the frequency of item additions, while the Koninklijke Bibliotheek defines accessType properties to specify the type of access permitted for collections. Final Considerations: In conclusion, while libraries share common metadata elements, the variation in their use, structure, and descriptions highlights the influence of each institution's unique context and standards. For effective harmonization of metadata between different systems, it would be beneficial to develop a common vocabulary or mappings that explain these variations in scope, as future work on this mapping.

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

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).