XXIV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – XXIV ENANCIB

ISSN 2177-3688

GT 8 – Informação e Tecnologia

CROSSWALK EM BIBLIOTECAS NACIONAIS NO CONTEXTO DO *LINKED DATA*

METADATA CROSSWALK IN NATIONAL LIBRARIES IN THE CONTEXT OF LINKED DATA

Ana Carolina Simionato Arakaki - Universidade Federal de São Carlos (UFSCar) e Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
Felipe Augusto Arakaki - Universidade de Brasília (UnB) e Universidade Federal de São Carlos (UFSCar)

Modalidade: Resumo Expandido

Resumo: A abertura de dados é um desafio para as unidades de informação, como as bibliotecas. Existe uma demanda crescente para que esses dados sejam abertos e conectados, permitindo acesso, identificação, uso e reuso. Nesse contexto, o problema da pesquisa consiste em como estruturar os dados das bibliotecas nacionais com base nos princípios dos dados abertos conectados. Com o objetivo de identificar os processos necessários para a publicação de dados abertos conectados por bibliotecas. A pesquisa está em andamento e atualmente está na fase de identificação dos modelos de dados e metadados utilizados por cada biblioteca. Os metadados serão mapeados e comparados para entender quais informações são utilizadas na descrição dos recursos informacionais das bibliotecas e estabelecer um conjunto mínimo de informações que devem ser descritas, além de identificar as principais ontologias utilizadas. Acredita-se que os processos estabelecidos poderão contribuir com diretrizes teórico-metodológicas para que as bibliotecas brasileiras possam publicar seus dados abertos conectados de forma mais efetiva.

Palavras-chave: metadados; linked data; crosswalk; bibliotecas nacionais; dados abertos conectados.

Abstract: Data openness is a challenge for information units such as libraries. There is a growing demand for this data to be open and connected, allowing access, identification, use and reuse. In this context, the research problem is how to structure the data of national libraries based on the principles of connected open data. With the aim of identifying the processes necessary for the publication of connected open data by libraries. The research is ongoing and is currently in the phase of identifying the data models and metadata used by each library. The metadata will be mapped and compared to understand what information is used to describe the libraries' information resources and establish a minimum set of information that should be described, as well as identifying the main ontologies used. It is believed that the processes established will be able to contribute theoretical and methodological guidelines so that Brazilian libraries can publish their connected open data more effectively.

Keywords: metadata; linked data; crosswalk; national libraries; linked open data.

1 INTRODUÇÃO

Com a disposição das tecnologias advindas da *Web* Semântica, faz com que os ambientes informacionais digitais, especialmente do domínio bibliográfico, tenham mais possibilidades para melhorias e otimização no processo de busca e recuperação da informação. Apesar dos instrumentos para representação utilizados na Biblioteconomia já serem consolidados e estabelecidos, muitos recursos informacionais foram sendo criados e tornando as orientações obsoletas. Desse modo, a Biblioteconomia possui como desafio, a revisitação dos processos e instrumentos para representação, sendo mais adequados e condizentes às novas demandas tecnológicas disponíveis. Nesse viés, para estas adequações e atualizações, é necessário uma revisão e alteração dos paradigmas teóricos e metodológicos.

Outro ponto a ser considerado, é que os dados bibliográficos e de autoridade são publicados na *Web*, por meio de plataformas de disponibilização de registros e pelos catálogos. Contudo, eles ainda estão sendo publicados em formatos monolíticos e que não possuem uma estrutura possível de conexão entre outros dados e por essa razão, os instrumentos de representação e organização da informação também precisam convergir para as boas práticas de publicação de dados na *Web*. Para isso e em âmbito geral, o *World Wide Web Consortium* (W3C) apresentou dois documentos que propõem boas práticas para publicação de dados no ambiente *Web*. O primeiro documento trata das boas práticas para publicação de dados em formato aberto, ou seja, apresenta orientações e requisitos para que instituições de qualquer tipo possam publicar seus dados de forma aberta (Lóscio; Burle; Calegari, 2017). Já o outro documento, trata das boas práticas para publicação de dados em *Linked Data*, ou seja, apresenta orientações para conectar dados que estão publicados na Web (Hyland; Atemezing; Villazón-Terrazas, 2014).

O *Linked Data* é fundamentado pelos conceitos da *Web* Semântica e é definido como um conjunto de boas práticas para o estabelecimento de tecnologias e orientações para publicação de dados conectados e ao considerar esses dados publicados em formato aberto é denominado de *Linked Open Data*. Estas duas orientações são complementares e juntas promovem a publicação de dados abertos conectados de forma padronizada.

A partir dessa contextualização, pondera-se que as bibliotecas possuem uma grande quantidade de dados sobre seu acervo. Contudo, esses dados não estão disponíveis e

acessíveis para serem utilizados por aplicações tecnológicas na *Web*, ou ainda, pelos próprios usuários, prejudicando o compartilhamento de dados. Além desses desafios, é destacado também a dificuldade dos sistemas de informação utilizados pelas bibliotecas, que não estão preparados para a disponibilização desses dados de forma aberta na *Web*.

Ressalta-se que os instrumentos existentes e utilizados na Biblioteconomia para representar os recursos informacionais, estão relacionados ao processo de reprodução e construção de fichas catalográficas, isto é, possuem uma estrutura estabelecida para um catálogo em papel. Apesar das iniciativas para que esses dados sejam legíveis por máquinas, há algumas iniciativas para publicação dos dados sobre o acervo de forma aberta e conectada, entretanto, desconhece instituições no Brasil que publicam os dados de bibliotecas a partir dos princípios de dados abertos conectados.

Diante deste contexto, o problema da pesquisa consiste em como estruturar os dados das bibliotecas nacionais com base nos princípios dos dados abertos conectados. As bibliotecas nacionais podem ser conceituadas como responsáveis pela execução da política governamental de captação, guarda, preservação e difusão da produção intelectual do País. Assim, com o objetivo de identificar os processos necessários para a publicação de dados abertos conectados por bibliotecas. A pesquisa está em andamento e atualmente está na fase de identificação dos modelos de dados e metadados utilizados por cada biblioteca.

2 REFERENCIAL TEÓRICO

Berners-Lee, Hendler e Lassila (2001), ao publicarem o artigo "*The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities*", estabelecem os princípios para a *Web* Semântica. Dessa forma, as questões de interoperabilidade são respaldadas pelos recursos da *Web*. Nesse artigo, os autores afirmam que a *Web* Semântica seria uma *Web* na qual os computadores poderiam compreender o contexto das pessoas para interpretar o significado das informações. Além disso, eles complementam que: "A *Web* Semântica não é uma *Web* separada, mas uma extensão da atual, na qual as informações possuem um significado bem definido, permitindo que computadores e pessoas trabalhem em cooperação" (Berners-Lee; Hendler; Lassila, 2001).

Em 2006, Shadbolt, Hall e Berners-Lee publicaram "*The Semantic Web Revisited*" (Shadbolt; Berners-Lee; Hall, 2006). Por meio de pesquisas, publicações e do desenvolvimento

da Web Semântica, surgiram conceitos de dados conectados. Nesse mesmo ano, Tim Berners-Lee, inventor da *World Wide Web*, propôs o conceito de *Linked Data* em um artigo intitulado "*Linked Data*", publicado na revista científica *Scientific American*. O objetivo de Berners-Lee era tornar a *Web* mais inteligente, permitindo que os dados fossem conectados e acessados de forma eficiente por humanos e máquinas (Berners-Lee, 2006).

A *Web* Semântica não se resume apenas a disponibilizar dados na *Web*. Trata-se de estabelecer conexões de forma que tanto pessoas quanto máquinas possam explorar a *Web* de dados. Com o *Linked Data*, quando se possui um pequeno conjunto de dados, é possível encontrar outros dados relacionados (Berners-Lee, 2006).

Tais tecnologias abarcavam desde linguagens para a estruturação e a representação dos conteúdos como o *eXtensible Markup Language* (XML) e o *Resource Description Framework* (RDF), passando por protocolos para recuperação de dados como o SPARQL *Protocol and RDF Query Language (*SPARQL) e linguagens para a construção de ontologias como a *Web Ontology Language* (OWL), chegando a estruturas básicas para a identificação única de conteúdos como o *Uniform Resource Identifier* (URI).

Em 2007, foi fundada a *Open Data Consortium*, uma organização sem fins lucrativos, com o objetivo de promover a publicação de dados abertos e conectados na *Web*. Desde então, diversas outras organizações e iniciativas surgiram para fomentar e incentivar a adoção do *Linked Data,* incluindo a *World Wide Web Foundation* e o *Linked Data Research Centre*.

No ano de 2009, Bizer, Heath e Berners-Lee utilizaram o termo *Linked Open Data*, traduzido como Dados Abertos Conectados, definindo-o como um conjunto de melhores práticas para a publicação aberta e interconexão de conjuntos de dados estruturados na *Web*, com o propósito de criar uma *Web* de Dados (Bizer *et al.*, 2008).

A partir desta perspectiva, é notável que as bibliotecas estão isoladas em termos de intercâmbio de dados, uma vez que os dados são recolhidos principalmente de bibliotecas para bibliotecas. O processo de intercâmbio e utilização conjunta de dados com instituições não bibliotecárias ainda está a dar os primeiros passos e não utilizam de formatos interoperáveis, isto deve-se principalmente ao fraco nível de ligação entre os conjuntos de dados das bibliotecas e os dados de outros domínios.

Van Hooland e Verborgh (2014) afirmam que, ao estruturar e reduzir os campos de metadados a uma dimensão mínima, tornamos esses campos mais interoperáveis com as máquinas, mas também nos tornamos cada vez mais dependentes dos esquemas quando

precisamos interpretar nossos próprios metadados ou os de outras pessoas. Nesse sentido, os autores também expõem que:

> Estas diferenças estão relacionadas com a medida em que o modelo utilizado para representar um objeto e os seus metadados é considerado adequado ou não. Quando queremos disponibilizar os recursos e os seus metadados de forma estruturada na *Web*, temos de decidir primeiro quais são as suas características mais importantes para serem representadas. Ao fazê-lo, fazemos uma abstração da realidade através do desenvolvimento de um modelo (Van Hooland; Verborgh, 2014, p. 11).

Ressalta-se assim, que a adoção do *Linked Data* em bibliotecas é necessária para promover a abertura dos dados e permitir que agentes computacionais consumam essas informações. Isso não apenas melhora o processo de navegação do usuário além do catálogo que está consultando, mas também cria uma rede interconectada entre diversas instituições, facilitando a recuperação e o acesso aos dados.

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa é caracterizada como qualitativa, exploratória de cunho teórico, que utiliza como método para análise dos dados o método *crosswalk* proposto pela *National Information Standards Organization* (NISO) em 1999 (Ballinger, 2015).

O método *crosswalk* é utilizado como processo para viabilizar a interoperabilidade entre sistemas que utilizam padrões de metadados heterogêneos. De acordo com a NISO, os "*Crosswalks* fornecem a capacidade de fazer o conteúdo de elementos definidos em um padrão de metadados disponíveis para as comunidades que utilizam padrões de metadados relacionados" (Chan; Zeng, 2006). O método estabelece quatro etapas, que são: 1) harmonização, 2) mapa semântico, 3) mapeamento elemento a elemento, e 4) hierarquia, objeto e visão lógica.

Para realização do *crosswalk*, há duas possibilidades, o "*crosswalking* absoluto" e o "*crosswalking* relativo". O "*crosswalking* absoluto" é a correspondência exata entre os metadados, ou seja, a semântica de um metadado é exatamente a mesma do metadado do outro padrão que está sendo analisado. A correspondência garante equivalência dos elementos. Quando isso não ocorre no processo de correspondência dos metadados, não há o *crosswalking*, o que resulta em perdas de informações. Para minimizar esse problema, as autoras sugerem a realização do "*crosswalking* relativo", usado para corresponder os

elementos de um esquema de fonte de pelo menos um elemento de um esquema de destino (Pierre; LaPlant, 2000).

O processo de *crosswalk* pode apresentar algumas dificuldades em diferentes graus de equivalência como: um-para-um, um-para-muitos, muitos-para-um e um-para-nenhum. A equivalência um-para-um corresponde a um metadado corresponde a apenas um metadado no outro padrão. Na equivalência um-para-muitos significa que um metadado do primeiro padrão, pode ter diversos metadados com contexto similar, fazendo com que a correspondência final possa apresentar diversos metadados correspondidos. Na equivalência muitos-para-um significa que muitos metadados correspondem a apenas um metadado no segundo padrão. Por fim, a equivalência um-para-nenhum representa que um metadado não teve nenhum metadado correspondido com o segundo padrão (Pierre; LaPlant, 2000).

Nesse contexto, optou-se em realizar o *crosswalk* dos metadados identificados nas Bibliotecas Nacionais que possuem catálogos que compartilham dados abertos conectados. O Universo de pesquisa foi definido a partir do levantamento realizado (Jesus, 2021) que identificou onze bibliotecas nacionais que fazem a publicação os dados abertos conectados: *Bibliotheca Apostolica Vaticana* (BAV); *Biblioteca Nacional de Espanã* (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); e *National Széchényi Library* (NSZL).

4 RESULTADOS PARCIAIS

A pesquisa encontra-se em andamento e está no processo de identificação dos modelos de dados e os metadados utilizados em cada biblioteca.

A *Biblioteca Nacional de Espanã* (BNE) adota um modelo de dados com 8 classes, 51 propriedades do objeto e 174 propriedades de dados, indicando uma estrutura detalhada. Em contraste, a *Deutsche Nationalbibliothek* (DNB) adota um modelo ainda mais complexo, com 131 classes, 53 propriedades de dados e 23 propriedades de anotações, refletindo uma abordagem robusta e abrangente na organização e descrição dos acervos.

A *Bibliothèque Nationale de France* (BnF) apresenta 42 propriedades, demonstrando uma abordagem menos complexa, mas ainda assim substancial. A *British National*

Bibliography (BNB), *Library Information System of Swedish National Union* (LIBRIS) e a *Library of Congress* (LC) utilizam o BIBFRAME como padrão, indicando uma tendência de padronização e facilitação da interoperabilidade entre diferentes sistemas.

A *Finnish National Bibliography* (FENNICA) utiliza o modelo de dados OCLC *WorldCat Linked Data* e a separação entre Obra e Instância do BIBFRAME 2.0, com 9 classes e 36 propriedades. Esta adoção de padrões reconhecidos internacionalmente facilita a interoperabilidade e a integração com outros sistemas de informação.

A *British National Bibliography* (BNB), *Library Information System of Swedish National Union* (LIBRIS) e a *Library of Congress* (LC) adotam o BIBFRAME como padrão, indicando uma tendência de padronização que pode simplificar a integração e a interoperabilidade entre diferentes bibliotecas. A *Koninklijke Bibliotheek* (KB) dos Países Baixos possui um modelo mais compacto, com 4 classes e 22 propriedades, mostrando uma abordagem específica e direcionada para suas necessidades.

A *National Széchényi Library* (NSZL) da Hungria faz uma adaptação do HUNMARC para *Dublin Core*, FOAF e SKOS, evidenciando a flexibilidade e a adaptabilidade dos modelos de dados para atender às necessidades específicas de cada instituição.

As bibliotecas *Bibliotheca Apostolica Vaticana* (BAV) e a *National Library of Iran* (NLAI) não foram encontradas as documentações relativas aos modelos de dados e os metadados utilizados.

5 CONSIDERAÇÕES FINAIS

A pesquisa desenvolvida demonstra a importância do método *crosswalk* para promover a interoperabilidade entre sistemas de metadados heterogêneos. Com os resultados, pontua-se que a harmonização dos dados provenientes de diversas bibliotecas nacionais que adotam práticas de dados abertos conectados.

A análise inicial revelou que o *crosswalking*, seja ele absoluto ou relativo, enfrenta desafios significativos, especialmente em situações de correspondência um-para-muitos, muitos-para-um e um-para-nenhum. Tais desafios ressaltam a necessidade de abordagens cuidadosas e criteriosas para minimizar a perda de informações e maximizar a precisão na correspondência dos elementos de metadados. Portanto, a pesquisa não apenas contribui para a compreensão e a aplicação do método *crosswalk*, mas também reforça a importância

de práticas colaborativas e padronizadas na gestão de metadados, visando uma maior integração e acesso da informação no âmbito global. Os metadados mapeados e comparados poderão estabelecer um conjunto mínimo de informações que devem ser descritas e quais as principais ontologias que são utilizadas.

Nesse viés, ressalta-se a contribuição social e inovadora da publicação e difusão dos dados de bibliotecas, visto o acesso democrático e aberto das informações sobre autores e recursos informacionais, ampliando a dinamização dos ambientes que se adequem às tecnologias semânticas, com melhorias e facilidades para qualquer aplicação tecnológica que queira utilizar os dados em questão. A partir da disponibilização e dos tratamentos dos dados adequadamente das instituições, estará disponível à população a identificação de recursos informacionais que até então não eram localizados.

Para validação da proposta, os dados serão aplicados em uma biblioteca pública. A escolha da biblioteca foi motivada pelos desafios que uma biblioteca pública possui em relação a recursos, pessoal e infraestrutura. Além de representar um ecossistema de informações complexas e similares a muitas outras bibliotecas brasileiras.

Os próximos passos da pesquisa incluirão a implementação prática do crosswalk entre os modelos de dados das bibliotecas estudadas, com foco na validação e refinamento das correspondências mapeadas. Será importante desenvolver e testar ferramentas de software que possam automatizar partes do processo de *crosswalking*, tornando-o mais eficiente e menos sujeito a erros humanos.

Adicionalmente, a pesquisa futura poderá explorar a criação de uma ontologia unificada ou de um conjunto de ontologias interoperáveis que possam servir como um padrão de referência para bibliotecas nacionais e outras instituições. Esse trabalho pode envolver a colaboração com organizações internacionais e a participação em iniciativas de padronização.

REFERÊNCIAS

BALLINGER, Linda. Metadata through the pages of information standards quarterly. **National Information Standards Organization**. 2015. Disponível em: https://www.niso.org/niso-io/2015/09/metadata-through-pages-information-standards-quarterly. Acesso em: 3 jul. 2024.

BERNERS-LEE, Tim. **Linked data**: design issues. 2006. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 4 dez. 2016.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The semantic web. **Scientific american**, v. 284, n. 5, p. 28–37, 2001.

BIZER, Christian; HEATH, Tom; IDEHEN, Kingsley; BERNERS-LEE, Tim. Linked data on the web. 2008. *In*: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 17., 2008, Beijing, China. **Proceedings of the 17th [...].** New York, USA: ACM, 2008. p. 1265–1266.

CHAN, Lois Mai; ZENG, Marcia Lei. Metadata interoperability and standardization–a study of methodology part I. **D-Lib magazine**, v. 12, n. 6, p. 1082–9873, 2006.

HYLAND, Bernadette; ATEMEZING, Ghislain; VILLAZÓN-TERRAZAS, Boris (ed.). **Best practices for publishing Linked Data**: W3C Working Group Note 09 January 2014. 2014. Disponível em: https://www.w3.org/TR/ld-bp/.

JESUS, Ananda Fernanda de. **Recomendações teórico-metodológicas para a publicação de dados bibliográficos abertos e conectados**. 2021. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de São Carlos (UFSCar), São Carlos, SP, 2021. Disponível em: https://repositorio.ufscar.br/handle/ufscar/14228.

LÓSCIO, Bernadette Farias; BURLE, Caroline; CALEGARI, Newton. **Data on the Web Best Practices**: W3C Recommendation 31 January 2017. 2017. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em: 19 mar. 2022.

PIERRE, Margaret St; LAPLANT, William P. **Issues in crosswalking content metadata standards**. [*S. l.*]: NISO Baltimore, Maryland, USA, 2000.

SHADBOLT, Nigel; BERNERS-LEE, Tim; HALL, Wendy. The semantic web revisited. **IEEE intelligent systems**, v. 21, n. 3, p. 96–101, 2006. Disponível em: https://ieeexplore.ieee.org/abstract/document/1637364.

VAN HOOLAND, Seth; VERBORGH, Ruben. **Linked data for libraries, archives and museums:** How to clean, link and publish your metadata. [*S.l.*]: Facet publishing, 2014.

AGRADECIMENTOS

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio e financiamento da pesquisa.