XXI ENANCIB
Encontro Nacional de Pesquisa em Ciência da Informação
50 anos de Ciência da Informação no Brasil: diversidade, saberes e transformação social
Rio de Janeiro • 25 a 29 de outubro de 2021

XXI Encontro Nacional de Pesquisa em Ciência da Informação – XXI ENANCIB

GT-8 – Informação e Tecnologia

DADOS DE BIBLIOTECA E PROVENIÊNCIA: ANÁLISE DOS PADRÕES DE METADADOS

LIBRARY DATA AND PROVENANCE: ANALYSIS OF METADATA STANDARDS

Felipe Augusto Arakaki - Universidade de Brasília (UnB)
Plácida Leopoldina Ventura Amorim da Costa Santos - Universidade Estadual Paulista (Unesp)

Modalidade: Trabalho Completo

Resumo: Considerando-se a tendência de abertura dos dados de bibliotecas, foi realizada uma pesquisa que buscou analisar como são tratadas as questões de proveniência nos seguintes padrões de metadados: *Bibliographic Framework Initiative* (BIBFRAME), *Format for Bibliographic Data* (MARC21), *Dublin Core*, *Schema.org* e *Preservation Metadata Maintenance Activity* (PREMIS). Essa pesquisa qualitativa, exploratória, de cunho teórico foi conduzida através do método *Crosswalk* de análise de dados. O mapeamento entre a ontologia PROV (PROV-O) e os padrões de metadados utilizados em bibliotecas foram os resultados obtidos. Entre os padrões analisados, o *Schema.org* apresentou o maior número de classes/subclasses correspondidas com a PROV-O e, portanto, esse padrão demandaria menor adaptação para uma possível utilização conjunta com essa ontologia. Foi identificada uma compatibilidade entre a PROV-O e os padrões utilizados pelas bibliotecas, o que possibilita a implementação dessa ontologia para a descrição de registros com informações sobre a proveniência dos recursos informacionais.
Palavras-Chave: Metadados; Proveniência; Crosswalk; Dados de biblioteca.

Abstract: *Considering the trend for the opening of library data, a survey was carried out to analyze how issues of provenance are treated in the following metadata standards: Bibliographic Framework Initiative (BIBFRAME), Format for Bibliographic Data (MARC21), Dublin Core, Schema.org, and Preservation Metadata Maintenance Activity (PREMIS). This qualitative, exploratory, theoretical search was conducted using the Crosswalk data analysis method. The mapping between the PROV ontology (PROV-O) and the metadata standards used in libraries were the results obtained. Among the standards analyzed, Schema.org presented the larger number of classes/subclasses matched with PROV-O and, therefore, this standard would require less adaptation for a possible combined use with this ontology. Compatibility between PROV-O and the standards used by libraries was identified, which allows the implementation of this ontology to describe records with information about the provenance of informational resources.*
Keywords: *Metadata; Provenance; Crosswalk; Library data.*

1 INTRODUÇÃO

O *World Wide Web Consortium* (W3C) publicou, em 2013, um conjunto de documentos, chamado família PROV, que tratam da identificação da proveniência. A família PROV é composta por quatro recomendações, além de oito notas que auxiliam no mapeamento das informações sobre o modelo.

De acordo com a literatura da Ciência da Informação, proveniência foi discutida no contexto da Arquivologia, sendo fundamentada como um princípio para organização e estruturação dos fundos arquivísticos. Na Biblioteconomia, a questão da proveniência ainda é pouco identificada, mas observa-se que a disseminação e a estruturação de ambientes em dados de bibliotecas no meio digital amparam-se para que essa discussão seja iniciada nas reflexões sobre a publicação de dados de bibliotecas na Web.

De acordo com W3C (2011, não paginado, tradução nossa), dados de biblioteca "[...] referem-se a qualquer tipo de informação digital produzida ou curada por bibliotecas que descreve recursos ou ajuda a sua descoberta."

Assim, ao considerar o movimento de abertura de dados, questiona-se quais metadados podem ser utilizados para a identificação da proveniência em dados de bibliotecas. Dessa forma, o objetivo deste trabalho é analisar como são tratadas as questões de proveniência nos seguintes padrões de metadados utilizados em bibliotecas: *Bibliographic Framework Initiative* (BIBFRAME), *Format for Bibliographic Data* (MARC21), *Dublin Core, Schema.org* e *Preservation Metadata Maintenance Activity* (PREMIS).

O estudo em questão corrobora o campo científico no desenvolvimento das temáticas como metadados e a proveniência de dados, sendo que a discussão proposta oferece subsídios para a fundamentação teórica, além de refletir sobre a construção de registros e a representação de recursos informacionais em bibliotecas. No campo profissional, acredita-se que a identificação clara da proveniência dos dados a partir dos metadados administrativos é fundamental para promover a confiança e credibilidade das informações disponibilizadas por esses ambientes digitais.

Na área social, o estudo proporciona discussões sobre a aplicação da proveniência em padrões de metadados utilizados em bibliotecas. Dessa forma, acredita-se que a concretização da proposta facilitará a identificação das informações e possibilitará a localização de diversos outros recursos relacionados ao que se está visualizando. Logo, o usuário poderá otimizar a navegação em sistemas de informação utilizados pelas bibliotecas, e ainda, poderá escolher um determinado recurso informacional, suas derivações e recursos

relacionados.

Dentre os trabalhos correlatos que tratam da proveniência, destacam-se Albuquerque e Souto (2013), Millar (2015), Rautenberg, Marx, Ermilov e Auer (2016), Arakaki, Alves e Santos (2019), Freund, Sembay e Macedo (2019), Tognoli e Guimarães (2019) e Arakaki e Santos (2021).

Este texto é estruturado em uma seção para revisão de literatura sobre a proveniência e a ontologia PROV (PROV-O), além de abordar os principais padrões de metadados utilizados em bibliotecas, como MARC21, *Dublin Core*, BIBFRAME*, Schema.org* e PREMIS. Em seguida, são apresentados os procedimentos metodológicos, seguidos da seção de análise de dados, na qual são identificadas quais classes e propriedades podem ser utilizadas para identificação da proveniência dos dados em cada padrão de metadados. Por fim, são tecidas as considerações finais.

2 BACKGROUND: DA PROVENIÊNCIA AOS PADRÕES DE METADADOS

Os estudos de Albuquerque e Souto (2013) e Tognoli e Guimarães (2019) apresentam uma perspectiva histórica do termo proveniência no contexto da Arquivologia. Nesse contexto, identifica-se que a proveniência é tratada como princípio para organização do fundo arquivístico. Isso garante a organicidade dos acervos arquivísticos e a autenticidade e confiabilidade dos dados.

Com o movimento de publicação de dados em acesso aberto, a questão da proveniência transcende os domínios da Arquivologia. Dessa forma, o conceito de proveniência torna-se fundamental para representar informações sobre um determinado recurso informacional em ambientes digitais. Nesse contexto, o W3C estruturou um conjunto de documentos, batizado de família PROV, para representar adequadamente a proveniência de recursos informacionais com o intuito de garantir a autenticidade e confiabilidade dos dados.

A família de documentos PROV está estruturada em três aspectos: entidade, atividade e agente. Uma "entidade" corresponde a coisas físicas, digitais ou conceituais. A atividade corresponde aos processos, ou seja, registra as ações e etapas para construção e alteração do recurso. O agente está relacionado ao criador de um recurso, ou seja, uma pessoa ou organização (ARAKAKI; SANTOS, 2021, e124).

A estrutura definida entre *Entity* (entidade), *Activity* (atividade) e *Agent* (agente) serve de base para todos os outros documentos da PROV. A família de documentos PROV é composta por um modelo de dados - "*PROV Data Model (PROV-DM)*", uma ontologia - "*PROV Ontology (PROV-O)*", formas de anotação "*Provenance Notation (PROV-N)*" e limitações - "*Constraints of the PROV Data Model (PROV-CONSTRAINTS)*". Esse conjunto de recomendações possibilita a estruturação de dados de proveniência. Entretanto, destaca-se que o foco deste trabalho foi a PROV-O, pois apresenta classes e propriedades que podem ser mapeadas com padrões de metadados utilizados em bibliotecas.

A PROV-O possui as três classes: *prov:Entity* (Entidade), *prov:Activity* (Atividade) e *prov:Agent* (Agente) e nove propriedades principais: *prov:wasGeneratedBy* (Foi gerado por), *prov:wasDerivedFrom* (Foi derivado de), *prov:wasAttributedTo* (Foi atribuído a), *prov:startedAtTime* (Iniciado em), *prov:used* (Usado), *prov:wasInformedBy* (Foi informado por), *prov:endedAtTime* (Finalizado em) *prov:wasAssociatedWith* (Foi associado com) e *prov:actedOnBehalfOf* (Agiu em nome de) (LEBO, SAHOO, MCGUINNESS, 2013).

A PROV-O possui ainda classes e propriedades expandidas e qualificadas que identificam as características de um recurso, agente ou atividade. Essas classes e propriedades estendidas auxiliam na explicitação das classes e propriedades, o que possibilita a sua identificação e diferenciação.

Essas características devem ser inseridas na descrição de um recurso informacional. Dentre os padrões de metadados utilizados pelas bibliotecas pode-se destacar: *Dublin Core*, *Format for Bibliographic Data* (MARC21), *Bibliographic Framework Initiative* (BIBFRAME), *Schema.org* e *Preservation Metadata Maintenance Activity* (PREMIS).

2.1 *Dublin Core*

O padrão de metadados *Dublin Core* surgiu em 1995 a partir da necessidade de localizar recursos informacionais na *Web*, e é utilizado em repositórios e bibliotecas digitais. Seu histórico foi relatado por autores como Arakaki, Alves e Santos, P. (2018).

Ao longo dos anos, o *Dublin Core* passou por diversas modificações. De acordo com Eckert, Garijo e Panzer (2011), logo na década de 90, foram realizadas discussões para definir um vocabulário e diretrizes de metadados de proveniência. Esses componentes abordavam metadados administrativos para o registro, atualizações e alterações, além de promoverem o intercâmbio de registros em lote (HANSEN; ANDRESEN, 2003). Entretanto, o projeto não teve continuidade, pois carecia de um modelo para relacionar as informações de proveniência com os elementos já estabelecidos.

Em 2007, Powell et al. (2007) propuseram uma padronização para a elaboração da estrutura de representação de sistemas de informação com o *Dublin Core Abstract Model* (*DCAM* - Modelo Abstrato do *Dublin Core*). Como processo de ampliação e compatibilidade de aplicação do *Dublin Core* no contexto da proveniência, Eckert, Garijo e Panzer (2011) apresentaram uma arquitetura para incluir informações de proveniência incorporadas ao *DCAM*. A proposta consistiu em relacionar as informações de proveniência com os elementos das entidades existentes do *DCAM*. Fruto das discussões e trabalhos de Eckert, Garijo e Panzer (2011) e Garijo e Eckert (2013) apresentaram um documento com mapeamento das classes/subclasses e propriedades/subpropriedades da PROV para os *DC terms*, o que possibilitou a extensão do *DC* para suportar informações de proveniência.

2.2 MARC21

O MARC21 é um formato de intercâmbio de dados criado na década de 60 com o intuito de compartilhar fichas catalográficas. Alguns autores como Gonzales (2014) e Assumpção e Santos (2015) discutiram esse formato, abordando seu histórico e evolução.

O MARC21 é composto por uma estrutura de registro, indicação de conteúdo, e conteúdo dos elementos que compõem o registro. A estrutura do registro é uma implementação do padrão internacional da ISO 2709, que trata do formato para intercâmbio de informações e do protocolo de intercâmbio de informação bibliográfica ANSI/NISO Z39.2. A indicação de conteúdo é estabelecida por códigos e convenções definidas explicitamente para identificar e caracterizar os elementos de dados e suportar a manipulação desses dados. O conteúdo dos elementos de dados é definido por padrões externos, como *International Standard Bibliographic Description* (ISBD), *Anglo-American Cataloguing Rule*s (AACR), entre outros (LIBRARY OF CONGRESS, 2006).

A estrutura é estabelecida pela ISO 2709 e possui Líder, Diretório e Campos variáveis.

O Líder fornece informações para o processamento do registro. O diretório corresponde por uma série de entradas que contêm a posição inicial e o tamanho de cada etiqueta (TAG) dentro do registro bibliográfico. Já os campos variáveis são dados em um registro bibliográfico e estão organizados em campos variáveis, cada um identificado por uma etiqueta de três (3) caracteres numéricos, registrados na entrada do diretório referente a cada campo (FERREIRA, M., 2002, p. iv).

O MARC21 possui cinco formatos com usos e finalidades distintas, eles são: MARC21, para dados bibliográficos; MARC21 para dados de autoridade, MARC21 para itens, MARC21 para dados de classificação e MARC21 para informação de comunidade. Entretanto, este trabalho focou no Formato MARC21 para Dados Bibliográficos.

Apesar da configuração do MARC21 ter sido criada na década de 90, a questão da proveniência no MARC21 teve destaque no ano de 2012, a partir de solicitações da *Deutsche Nationalbibliothek* (Biblioteca Nacional da Alemanha), da *Library of Congress* (EUA) e da *Online Computer Library Center* (OCLC). Essas bibliotecas realizaram uma revisão e solicitaram a criação de alguns campos para representar informações sobre a proveniência em registros MARC21. Então, o campo "883 - *Machine-generated Metadata Provenance*" foi criado em 2012.

Além do campo 883, Reinhold Heuvelmann (2016) destaca alguns elementos que o MARC21 possui para garantir a descrição da proveniência dos dados. Os seguintes campos foram destacados: Líder 008 (Campo de Tamanho Fixo), 040 (Fonte da catalogação), 042 (Código de autenticação), 588 (Fonte da descrição nota), 884 (Descrição de informações de conversão) e 881 (Informações de transliteração/romanização).

2.3 BIBFRAME

Considerando as tecnologias semânticas, a *Library of Congress* dos Estados Unidos iniciou a construção de um novo modelo de dados para o domínio bibliográfico. Essa proposta foi denominada *Bibliographic Framework Initiative* (BIBFRAME) e tem como missão substituir o MARC21. De acordo com a *Library of Congress* (2012), o novo modelo apresenta características da proposta do *Functional Requirements for Bibliographic Records* (FRBR) e ainda os princípios do *Linked Data*. No Brasil, o BIBFRAME foi discutido por Ramalho (2016), Arakaki et al (2017), Silva et al (2017), Espíndola e Pereira (2018), dentre outros.

O BIBFRAME foi lançado em 2012 e, após diversos estudos e testes, a *Library of*

Congress (EUA) publicou o BIBFRAME 2.0 em abril de 2016. O BIBFRAME 2.0 consiste de três classes principais: *Work* (Obra), *Instance* (Instância) e *Item* (Item). A Obra foi definida como o nível mais alto de abstração e manteve seu relacionamento com a Obra e Expressão do FRBR. A Instância foi caracterizada por possuir uma ou mais formas de realização de uma Obra. Uma Instância reflete informações como editor, local, data de publicação e formato e foi equiparada à entidade Manifestação do FRBR. Um Item é definido como uma cópia real (física ou eletrônica) de uma instância. Ele possui informações, tais como a sua localização (física ou virtual), marca de prateleira e código de barras, e foi equiparado à entidade Item do FRBR. (LIBRARY OF CONGRESS, 2016; ARAKAKI et al., 2017).

O Modelo BIBFRAME 2.0 aborda, ainda, algumas outras classes, como *Agents* (Agentes), *Subjects* (Assuntos) e *Events* (Eventos). A classe Agentes pode ser definida como pessoas, organizações e jurisdições e está associada a uma Obra ou Instância. A classe Agentes pode ter diversas funções, como autor, editor, artista, fotógrafo, compositor, ilustrador, etc. A classe Assuntos foi caracterizada pelas informações de conteúdo, ou seja, "sobre o quê" de uma Obra e pode ter um ou mais conceitos, incluindo temas, lugares, expressões temporais, eventos, obras, instâncias, itens, agentes, etc. A classe Eventos foi definida como ocorrências que podem estar relacionadas ao conteúdo de uma Obra (LIBRARY OF CONGRESS, 2016).

A proveniência foi pouco discutida no BIBFRAME e foram identificadas apenas indicações de trabalhos futuros, como o de Kovari, Folsom e Younes (2017).

2.4 *Schema.org*

Paralelo ao *Dublin Core* e BIBFRAME, o *Schema.org* surgiu de uma iniciativa de motores de busca, como Google, Yahoo, Bing e Yandex, para desenvolver uma estrutura que fosse capaz de melhorar a busca de informações na *Web*. O *Schema.org* é uma iniciativa comunitária e colaborativa com a missão de criar, manter e promover esquemas de dados estruturados para Internet (SCHEMA.ORG, 2011). O *Schema.org* foi abordado no Brasil por Ouchi e Simionato (2018), Roa-Martínez, Vidotti e Pastor-Sánchez (2018) e Tomoyose, Santos e Arakaki, A. (2019).

Segundo Pomerantz (2015), o *Schema.org* é baseado em microdados, que é uma especificação para a incorporação de metadados dentro de uma página da *Web*. Atualmente, o *Schema.org* está na versão 3.5 e pode ser utilizado com diversas codificações, como *RDFa*,

Microdata e *JSON-LD*.

Apesar de ter sido criado para diversos domínios, em especial para a *Web*, o *Schema.org* foi usado na estruturação em *Linked Data* do *WorldCat*, catálogo que busca reunir diversas bibliotecas em um catálogo universal gerenciado pela OCLC.

2.5 PREMIS 3.0

O *PREservation Metadata: Implementation Strategies* (PREMIS) é uma iniciativa de 2003 da *Online Computer Library Center* (OCLC) e *Research Libraries Group* (RLG) e, posteriormente, passou a ser administrado pela *Library of Congress* (EUA). Autores como Arakaki et al (2018), Arakaki, Alves, R. e Santos, P. (2019) e Castro e Alves, R. (2021) apresentaram pesquisas sobre o PREMIS.

O PREMIS possui um conjunto básico de elementos de metadados de preservação e está baseado no modelo *Open Archival Information System* (OAIS). O PREMIS define cinco entidades: *Environment* (Tecnologia que suporta um objeto digital), *Object* (Objeto), *Event* (Evento), *Agent* (Agente), e *Rights Statement* (Declaração de direitos). A entidade Suporte pode ser descrita como Entidades Intelectuais que são capturadas e preservadas no repositório como Representações, Arquivos e/ou *Bitstreams.* Suporte são ainda Tecnologias (*software* ou *hardware*) de um recurso informacional digital (por exemplo, renderização ou execução). A entidade Objeto pode ser definida como uma unidade discreta de informação sujeita a preservação digital e é usada como parte do processo de preservação. Evento é uma ação que envolve e afeta pelo menos um objeto ou agente associado ou conhecido. A entidade Agente pode ser pessoa, organização ou programa/sistema de *software* associada a eventos na vida de um objeto ou com direitos associados a um objeto. Por fim, a entidade Declaração de direitos é uma afirmação de um ou mais direitos ou permissões pertencentes a um objeto e/ou agente. (PREMIS…, 2015).

No contexto da *Web* Semântica, a *Library of Congress* (EUA) desenvolveu o PREMIS 3.0, que consiste na formalização semântica do dicionário de dados PREMIS 2.2 por meio de uma ontologia em OWL. Alguns estudos, como Li e Sugimoto (2014, 2017, 2018) e Arakaki, Alves, R. e Santos, P. (2019), realizaram um mapeamento entre o PROV e o PREMIS com o intuito de criar um modelo de proveniência no contexto da preservação digital.

3 METODOLOGIA

Esta é uma pesquisa qualitativa, exploratória, de cunho teórico que utiliza o método *Crosswalk*, proposto pela *National Information Standards Organization* (NISO) em 1999, para análise dos dados.

O método *Crosswalk* é utilizado como processo para viabilizar a interoperabilidade entre sistemas que utilizam padrões de metadados heterogêneos. De acordo com a *National Information Standards Organization* (1999, não paginado, tradução livre), "*Crosswalks* fornece capacidade de tornar o conteúdo de elementos definidos em um padrão de metadados disponível para as comunidades que utilizam padrões de metadados relacionados."

Esse método estabelece quatro etapas, a saber: harmonização, mapa semântico, mapeamento elemento a elemento, e hierarquia, objeto e visão lógica. Cada etapa foi estruturada como subetapa para facilitar o *crosswalk*, conforme apresentado no Quadro 1.

Quadro 1 - Apresentação do método *Crosswalk*

Etapa	Subetapa	Observação
1º etapa: Harmonização, extração da terminologia comum, propriedades, organização e processos utilizados pelos padrões de metadados e criação de um quadro genérico para que se possa desenvolver novos ou rever padrões de metadados já existentes.	Subetapa A: Terminologia	Utilização de terminologias diferentes dos padrões dificultam o mapeamento entre eles.
		É essencial chegar a um acordo sobre a terminologia dos padrões, além de estabelecer uma definição formal para cada termo.
	Subetapa B: Propriedades - as semelhanças das propriedades dos padrões são extraídas e os conceitos generalizados.	Identificadores únicos para cada metadado, por exemplo, TAG, etiqueta, identificador.
		Qual a definição semântica de cada metadado?
		O metadado é obrigatório, opcional ou obrigatório em certas condições?
		Um metadado pode ocorrer várias vezes?
		Organização de um metadado em relação ao outro, por exemplo, as relações hierárquicas.
		Quais são as restrições impostas pelos valores do elemento (texto livre, escala numérica ou data)?
		O suporte opcional para elementos de metadados são definidos localmente?
		As propriedades comuns podem ser expressas e utilizadas de forma similar dentro de cada padrão? Esta etapa simplifica o desenvolvimento do *Crosswalk*.
	Subetapa C: Organização	Para facilitar, cada padrão deve ser organizado de forma similar, de modo que determinada seção de um padrão possa ser encontrada em uma seção de outro padrão.
	Subetapa D: Processo	Há ocasiões nas quais a escolha do processo selecionado é arbitrária e não um processo análogo a outro padrão relacionado.
2º etapa: Mapa semântico.	O mapeamento semântico é a especificação de cada elemento do padrão com o elemento semanticamente equivalente para o outro padrão. De acordo com St.Pierre e LaPlant (1999), esse é o processo mais importante da harmonização e desenvolvimento do *Crosswalk*, pois determina o mapeamento semântico entre os padrões de metadados de origem e destino.	
3º etapa: Mapeamento elemento a elemento - Identificar os	**Uma para muitos:** ocorrência de vários elementos de origem a uma única ocorrência no elemento alvo. Um elemento que se está verificando será correspondente a diversos elementos do outro padrão de metadados.	
	Muitos para um: muitos elementos de um padrão de metadados para apenas um metadado no padrão de destino. Deve-se aproximar todos os elementos do primeiro metadado e indicar a um único	

metadados opcionais e obrigatórios. Nesta fase considerar as propriedades de cada metadado.	elemento do outro padrão. Se a resolução é mapear todos os valores do elemento de origem para um único valor no elemento alvo, regras explícitas são obrigadas a especificar como os valores serão anexados juntos. Caso seja apenas mapear um valor de elemento de origem para o destino, com a possível consequência de perda de informações, a resolução deve indicar os critérios para a seleção de elementos.
	Elementos extras na fonte: Outro caso importante que requer resolução é a manipulação de um elemento de origem que não é mapeado para qualquer elemento apropriado no padrão alvo. Uma vez que muitos padrões fornecem a capacidade de capturar informações adicionais, a resolução deve especificar exatamente como o valor do elemento deve ser adicionado.
	Elementos obrigatórios /não resolvidos em alvo: Em alguns casos, pode haver elementos obrigatórios no alvo que não têm mapeamento correspondente no padrão de metadados de origem. Porque o alvo requer um valor para os elementos obrigatórios, o *Crosswalk* deve fornecer uma resolução para os seus valores.
4º etapa: Hierarquia, objeto e visão lógica	**Hierarquia:** A maioria dos padrões de metadados organizam seus metadados hierarquicamente. Em alguns casos, a profundidade da hierarquia pode ser fixada, ao passo que em outros essa profundidade é ilimitada.
	Objeto: Item versus coleção. Item é um único documento, ou seja, os metadados associados a um documento. Coleção é um conjunto de itens, ou seja, os metadados referem-se a mais de um item.
	Visão Lógica: Permite ver um conjunto específico de metadados do padrão organizado de maneira específica.
	Conversão de conteúdo: Padrões de metadados restringem o conteúdo de cada metadado para um determinado tipo de dado, intervalo de valores ou vocabulário controlado. Muitas vezes, as conversões são baseadas não só nas propriedades que definem a fonte e os metadados alvo, mas também no conteúdo dos elementos de metadados de origem.
	Combinações de conversão: Quando as propriedades de conversão são consideradas de forma independente, as conversões de metadados podem parecer simples para especificar e processar. Na prática, vários problemas de conversão refletem em uma combinação, o que dificulta a especificação de conversão e processo. Deve-se considerar as transformações necessárias para converter um metadado alvo, onde várias propriedades são diferentes do metadado de origem.

Fonte: Baseado em National Information Standards Organization (1999).

Para realização do *crosswalk*, Chan e Zeng (2006) esclarecem que 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 perda de informações. Para minimizar esse problema, as autoras sugerem a realização do "*crosswalking* relativo", usado para corresponder os elementos de um esquema de fonte com pelo menos um elemento de um esquema de destino.

De acordo com Chan e Zeng (2006), 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 para apenas um metadado no outro padrão. A 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. A 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.

Nesse contexto, optou-se em realizar o *crosswalk* da PROV-O para os padrões aplicados em bibliotecas, pois o intuito desta pesquisa foi verificar a compatibilidade da proveniência a partir da PROV-O nos padrões utilizados em bibliotecas e, dessa forma, não apresenta a correspondência inversa, ou seja, dos padrões aplicados em bibliotecas para a PROV-O.

Para a estruturação dos resultados, foi realizado primeiro o *crosswalk* individual para cada padrão de metadados, em razão da complexidade dos mapeamentos. Posteriormente, foi criado um quadro com o *crosswalk* da PROV-O para todos os padrões analisados. Após esse mapeamento, foi possível estabelecer tanto as especificidades de cada correspondência, quanto um panorama do mapeamento.

4 RESULTADOS E DISCUSSÕES

Com o levantamento e a análise dos metadados contidos nos padrões de metadados, foi possível identificar os elementos que apresentam características e podem ser usados para identificação da proveniência nos registros bibliográficos.

Primeiro, foi realizado um mapeamento da PROV-O com os campos e subcampos do MARC21. Os campos foram equiparados e mapeados para as classes da PROV, pois entende-se que eles possuem similaridades. O mesmo ocorreu com os subcampos do MARC21, que foram mapeados com as propriedades da PROV-O.

O levantamento possibilitou identificar correspondência de dez (10) classes/subclasses definidas na PROV-O com o MARC21 e vinte (20) classes/subclasses que não tiveram correspondência com nenhum dos campos do MARC21. Vinte e três (23) propriedades/subpropriedades apresentaram alguma identificação com subcampos do MARC21 e 34 propriedades não apresentaram nenhuma correspondência com esses

subcampos. Apesar do MARC21 ter um campo específico para questões de proveniência (campo 883), nem todas as informações do campo puderam ser mapeadas para PROV.

A correspondência do *Dublin Core* para a PROV foi realizada com base nos estudos de Garijo e Eckert (2013), que fizeram adaptações para conciliar os dois vocabulários. Nesse contexto, algumas classes, subclasses, propriedades e subpropriedades do *Dublin Core* tornaram-se subclasses ou subpropriedades da PROV, validando assim a compatibilidade de alguns elementos entre os dois vocabulários.

No mapeamento da PROV para o BIBFRAME, observou-se que ambos possuem a mesma terminologia para muitas classe e propriedades. Isso facilitou o mapeamento, pois não houve necessidade de adaptações entre conceitos de classes e propriedades. O mapeamento revelou que dez (10) classes e oito (8) propriedades da PROV foram mapeadas para o BIBFRAME.

Para realizar o mapeamento da PROV para o *Schema.org*, foram consideradas as classes e propriedades das duas ontologias. Entretanto, destaca-se que, em alguns momentos, não foram identificadas classes equivalentes em ambas ontologias, mas em alguns casos foi possível mapear equivalências entre classes e propriedades. Para o mapeamento, foram consideradas principalmente as classes *Thing* e *CreativeWork* e a tipologia *Book* do *Schema.org*. No *crosswalk* foi possível mapear quatorze (14) classes e oito (8) propriedades.

Em relação ao PREMIS, foi considerado para o mapeamento o PREMIS 3.0, que teve como base a PROV-O. O mapeamento identificou que oito (8) classes e seis (6) propriedades foram utilizadas de fato.

Ao considerar o *crosswalk* geral, da PROV para os outros padrões de metadados analisados, constatou-se que as classes *prov:Agent* e *prov:Location* foram identificadas em todos os padrões de metadados analisados. As classes *prov:Person* e *prov:Organization* foram identificadas em quatro dos cinco padrões analisados. Já as classes, *prov:Activity*, *prov:Start* e *prov:PrimarySource* foram mapeadas em três dos cinco padrões de metadados. Foi observado ainda que sete (7) classes foram mapeadas entre dois dos cinco padrões (*prov:Entity*; *prov:Collection*; *prov:Bundle*; *prov:End*; *prov:Derivation*; *prov:Quotation*; *prov:Generation*).

As classes *prov:SoftwareAgent*, *prov:Usage*, *prov:Revision*, *prov:Communication*, *prov:Plan* e *prov:Role* foram identificadas em apenas um dos cinco padrões de metadados,

analisados.

Foi constatado ainda que dez (10) classes não foram mapeadas em nenhum dos cinco padrões (*prov:EmptyCollection*; *prov:Influence*; *prov:EntityInfluence*; *prov:ActivityInfluence*; *prov:Invalidation*; *prov:AgentInfluence*; *prov:Attribution*; *prov:Association*; *prov:Delegation*; *prov:InstantaneousEvent*). Esse resultado mostra a importância de estudos sobre a questão da proveniência e da utilização da PROV para dados de bibliotecas.

5 CONSIDERAÇÕES FINAIS

Este trabalho buscou analisar como são tratadas as questões de proveniência de dados nos padrões de metadados BIBFRAME, MARC21, *Dublin Core*, *Schema.org* e PREMIS utilizados em bibliotecas. Foi realizada, então, a correspondência (*crosswalk*) entre a PROV-O com cada padrão de metadados com o intuito de identificar qual a aderência da PROV-O no contexto de bibliotecas.

Entre os padrões analisados, o *Schema.org* apresentou o maior número de classes/subclasses correspondidas com a PROV-O e, portanto, esse padrão demandaria menor adaptação para uma possível utilização conjunta com essa ontologia. Em seguida, os formatos MARC21 e BIBFRAME apresentaram dez (10) classes/subclasses correspondidas enquanto o PREMIS apresentou oito classes/subclasses correspondidas. Na proposta de Garijo e Eckert (2013), o *Dublin Core* passou por uma harmonização com a PROV-O, possibilitando a efetiva utilização conjunta dos dois sistemas.

Em relação às propriedades mapeadas, o formato MARC21 foi o que apresentou o maior número de elementos mapeados (23), seguido do BIBFRAME e do *Schema.org* com oito (8) propriedades mapeadas e do PREMIS (6 propriedades).

O mapeamento revelou que duas classes/subclasses da PROV estão presentes em todos os padrões de metadados analisados. Duas classes/subclasses foram identificadas em quatro padrões, três classes/subclasses foram mapeadas em três padrões, sete classes/subclasses foram mapeadas em dois padrões, seis classes/subclasses foram mapeadas em um padrão, e dez classes/subclasses não foram mapeadas em nenhum padrão.

Esses resultados corroboram a compatibilidade da PROV-O com os padrões utilizados pelas bibliotecas, o que possibilita sua implementação. Entretanto, destaca-se que, para tanto, será necessário adequar todos os padrões de metadados analisados.

REFERÊNCIAS

ALBUQUERQUE, A. C.; SOUTO, D. V. B. Acerca do princípio da proveniência: apontamentos conceituais. **Ágora,** v. 23, n. 46, p. 14-44, 2013.

ARAKAKI, F. A. **Metadados administrativos e a proveniência dos dados**: modelo baseado na família PROV. 2019. 139 f. Tese (Doutorado em Ciência da Informação – Faculdade de Filosofia e Ciências) – Universidade Estadual Paulista "Júlio Mesquita Filho", Marília, 2019.

ARAKAKI, F. A.; ALVES, R. C. V.; SANTOS, P. L. V. A. C. Preservação digital e proveniência: interseções entre premis e o prov. ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 2019, Santa Catarina. **Anais**… Ancib, 2019.

ARAKAKI, F. A. et al. Bibframe: tendência para a representação bibliográfica na web. **Revista Brasileira de Biblioteconomia e Documentação**, v. 13, p. 2231-2249, 2017a.

ARAKAKI, F. A. et al. Web Semântica e preservação digital: o padrão de metadados PREMIS na proposta do Linked Data. **Informação & Tecnologia**, Marília/João Pessoa, v.5, n.1, jan./jun. 2018.

ARAKAKI, F. A.; SANTOS, P. L. V. A. C. Proveniência e contexto digital: contribuições da ciência da informação. **Palabra Clave** (La Plata), [S.L.], v. 10, n. 2, p. 124-124, 2021.

ASSUMPÇÃO, F. S.; SANTOS, P. L. V. A. C. Representação no domínio bibliográfico: um olhar sobre os Formatos MARC 21. **Perspectivas em Ciência da Informação**, v. 20, n. 1, p. 54–74, 30 mar. 2015.

CASTRO, F. F.; ALVES, R. C. V. Cloud services e o padrão PREMIS: rumos para a preservação digital. **Revista Digital de Biblioteconomia & Ciência da Informação**, v. 19, 2021.

CHAN, L. M.; ZENG, M. L. Metadata interoperability and standardization–a study of methodology part I. **D-Lib magazine**, v. 12, n. 6, p. 3, 2006.

ECKERT, K.; GARIJO, D.; PANZER, M. Extending DCAM for Metadata Provenance. In: INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATIONS, 2011, The Hague. **Anais...** The Hague: [s.n.], 2011. p. 12–25.

ESPíNDOLA, P. L.; PEREIRA, A. M. A influência do bibliographic framework para a visibilidade dos dados. ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 2018, Londrina. **Anais**… Ancib, 2018.

FERREIRA, Margarida M. **MARC 21**. Marília: UNESP, 2002.

FREUND, G. P.; SEMBAY, M. J.; MACEDO, D. D. J. Proveniência de dados e segurança da informação: relações interdisciplinares no domínio da ciência da informação. **Revista Ibero-Americana de Ciência da Informaçã**o, v. 12 No 3, n. 3, p. 807-825, 2019.

GARIJO, D.; ECKERT, K. **Dublin Core to PROV Mapping**. W3C, 2013. Disponível em: https://www.w3.org/TR/prov-dc/. Acesso em: 24 ago. 2021.

GONZALES, B. M. Linking Libraries to the Web: Linked Data and the Future of the Bibliographic Record. **Information Technology and Libraries**, v. 33, n. 4, 18 dez. 2014.

HANSEN, J.; ANDRESEN, L. AC - **Administrative Components**. [S.l.]: DCMI, 2003. Disponível em: dublincore.org/groups/admin/AdminComp_final_June_2003.doc. Acesso em: 24 ago. 2021.

HEUVELMANN, Reinhold. **Provenance in MARC 21**. 2016. Disponível em: https://pt.slideshare.net/sollbruchstelle/provenance-in-marc-21. Acesso em: 24 ago. 2021.

KOVARI, J.; FOLSOM, S.; YOUNES, R. Towards a BIBFRAME Implementation: The Bibliotek-o Framework. In: INTERNATIONAL CONFERENCE ON DUBLIN CORE AND METADATA APPLICATIONS, 2 dez. 2017, Whashington. **Anais.**.. Washington: DCMI, 2 dez. 2017. p. 52–61.

LEBO, T.; SAHOO, S.; MCGUINNESS, D. **PROV-O**: The PROV Ontology. Disponível em: https://www.w3.org/TR/2013/REC-prov-o-20130430/. Acesso em: 24 ago. 2021.

LI, C.; SUGIMOTO, S. Provenance Description of Metadata Application Profiles for Long-Term Maintenance of Metadata Schemas. **Journal of Documentation**, v. 74, n. 1, p. 36–61, 8 jan. 2018.

LI, C.; SUGIMOTO, S. **Provenance description of metadata using PROV with PREMIS for long-term use of metadata**. 2014, [S.l: s.n.], 2014. p. 147–156.

LI, C.; SUGIMOTO, S. Provenance description of metadata vocabularies for the long-term maintenance of metadata. **Journal of Data and Information Science**, v. 2, n. 2, p. 41–55, 2017.

MILLAR, L. A. A morte dos fundos e a ressurreição da proveniência: o contexto arquivístico no espaço e no tempo. **Informação Arquivística**, v. 4, n. 1, 2015.

NATIONAL INFORMATION STANDARDS ORGANIZATION. Issues in crosswalking content metadata standards. **Information standards quarterly**, v. 11, n. 1, p. 01–16, 1999.

OUCHI, M. T.; SIMIONATO, A. C. Descrição de conjuntos de dados na web com schema.org. **Informação & Tecnologia**, v. 5, n. 1, p. 128-140, 2018.

POWELL, A. et al. **DCMI**: DCMI Abstract Model, 2007. Disponível em: http://dublincore.org/documents/2007/06/04/abstract-model/. Acesso em: 24 ago. 2021.

PREMIS Data Dictionary for Preservation Metadata, Version 3.0. . [S.l.]: Library of Congress, 2015. Disponível em: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf. Acesso em: 24 ago. 2021.

RAMALHO, R. A. S. BIBFRAME: modelo de dados interligados para bibliotecas. **Informação & Informação**, Londrina, v. 21, n. 2, p. 292-306, maio/ago. 2016.

RAUTENBERG, S.; MARX, E.; ERMILOV, I.; AUER, S. Linked data workflow project ontology: uma ontologia de domínio para publicação e preservação de dados conectados. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 9, n. 2, 2016.

ROA-MARTíNEZ, S. M.; VIDOTTI, S. A. B. G.; PASTOR-SáNCHEZ, J. A. Marcação semântica enriquecida para programas de pós-graduação na américa latina. **Perspectivas em Ciência da Informação**, v. 23, n. 3, p. 67-88, 2018.

SILVA, L. C.; SEGUNDO, J. E. S.; ZAFALON, Z. R.; SANTOS, P. L. V. A. C. O código RDA e a iniciativa bibframe: tendências da representação da informação no domínio bibliográfico. **Em Questão**, v. 23, n. 3, p. 130-156, 2017.

TOGNOLI, N. B.; GUIMARÃES, J. A. C. Provenance as a knowledge organization principle. **Knowledge organization**, v. 46, n. 7, 558-68, 2019.

TOMOYOSE, K.; SANTOS, A. A.; ARAKAKI, A. C. S. Schema.org para recuperação da informação em redes sociais. In: BARROS, T. H. B.; TOGNOLI, N. B. (Org.). **Organização do conhecimento responsável**: promovendo sociedades democráticas e inclusivas. Belém,PA: Ed.da UFPA, 2019, v. 5, p. 176-182.