

CROSSWALK ENTRE PADRÕES DE METADADOS DUBLIN CORE E A FAMÍLIA PROV

CROSSWALK BETWEEN DUBLIN CORE METADATA STANDARDS AND THE PROV FAMILY

CROSSWALK ENTRE LOS ESTÁNDARES DE METADATOS DE DUBLIN CORE Y LA FAMILIA PROV

Felipe Ivo da Silva – Universidade Federal de São Carlos (UFSCar) –

felipe_ivodasilva@hotmail.com

Ana Alice Rodrigues Pereira Baptista – Universidade do Minho (UMinho) –

analice@dsi.uminho.pt

Felipe Augusto Arakaki – Universidade de Brasília (UnB) – felipe.arakaki@unb.br

Modalidade: Resumo Expandido Relatos de Pesquisa

Resumo: Este estudo investiga a integração entre os padrões de metadados Dublin Core (DC) e a família PROV, conforme as diretrizes da W3C, para garantir a rastreabilidade e confiabilidade de informações em repositórios digitais. Utilizando o método Crosswalk, foi realizada uma análise de correspondência entre as classes de ambos os padrões. Os resultados mostram uma equivalência de 50% entre as classes de metadados, destacando a compatibilidade entre DC e PROV. Essa correspondência reforça a possibilidade de alinhar ambos os modelos, promovendo a gestão eficiente da proveniência de dados e facilitando a adoção do modelo PROV.

Palavras-Chave: Metadados. Proveniência. Família PROV. Dublin Core.

Abstract: This study investigates the integration between the Dublin Core (DC) metadata standards and the PROV family, according to W3C guidelines, to ensure the traceability and reliability of information in digital repositories. Using the Crosswalk method, a correspondence analysis between the classes of both standards was performed. The results show a 50% equivalence between the metadata classes, highlighting the compatibility between DC and PROV. This correspondence reinforces the possibility of aligning both models, promoting the efficient management of data provenance and facilitating the adoption of the PROV model.

Keywords: Metadata. Provenance. PROV family. Dublin Core.

Resumen: Este estudio investiga la integración entre los estándares de metadatos Dublin Core (DC) y la familia PROV, según las directrices del W3C, para garantizar la trazabilidad y confiabilidad de la información en repositorios digitales. Utilizando el método Crosswalk se realizó un análisis de correspondencia entre las clases de ambos patrones. Los resultados muestran una equivalencia del 50% entre clases de metadatos, destacando la compatibilidad entre DC y PROV. Esta correspondencia refuerza la posibilidad de alinear ambos modelos, promoviendo una gestión eficiente de la procedencia de los datos y facilitando la adopción del modelo PROV.

Palabras clave: Metadatos. Procedencia. Familia PROV. Dublin Core.

1 INTRODUÇÃO

A crescente digitalização de objetos físicos e a disponibilidade de informações na Web levantam preocupações sobre a confiabilidade e credibilidade dos dados encontrados nos ambientes informacionais. Nesse contexto, a proveniência dos dados se torna essencial para rastrear a origem e o histórico das informações em repositórios digitais. Em 2013, o World Wide Web Consortium (W3C) lançou a família de documentos PROV, que fornece diretrizes para abordar a identificação de procedência em ambientes digitais, promovendo transparência e confiança nas informações online.

Os metadados desempenham um papel crucial na organização de recursos informacionais na Web. Entre os padrões mais amplamente adotados, o Dublin Core destaca-se por oferecer um vocabulário abrangente para descrever recursos digitais. Ele aprimora a recuperação de informações essenciais sobre recursos, tornando sua organização mais eficiente e estruturada, o que, por sua vez, facilita melhor compreensão e utilização dessas informações.

A conexão entre metadados Dublin Core e proveniência de dados, conforme descrito pelas diretrizes PROV, é crucial para garantir a credibilidade de recursos de informação baseados na web. O PROV fornece uma estrutura para descrever e compreender a proveniência em ambientes digitais e, quando combinado com padrões de metadados como Dublin Core, pode aprimorar a rastreabilidade e a confiabilidade das informações. Essa integração visa melhorar a confiança no gerenciamento de dados, especialmente em repositórios digitais.

Este estudo se concentra na coleta de metadados de repositórios que empregam o software DSpace de código aberto. Embora a comunidade do DSpace recomende o Dublin Core, ele não é obrigatório, permitindo que as instituições estabeleçam padrões alternativos ou personalizem metadados para atender às suas necessidades. O objetivo é ilustrar como a combinação de metadados com a identificação de metadados de proveniência pode melhorar a qualidade e a confiabilidade das informações disponibilizadas na Web.

2 PROCEDIMENTOS METODOLÓGICOS

A pesquisa em questão é caracterizada como teórico-aplicada e de natureza qualitativa, sendo classificada como exploratória em termos de método. Foi realizada uma pesquisa bibliográfica, abrangendo tanto fontes nacionais quanto internacionais, com o objetivo de reunir informações relevantes e atualizadas sobre o tema.

A revisão da literatura foi realizada em diversas bases de dados, como Scopus, Web of Science (WoS), Biblioteca Digital Brasileira de Teses e Dissertações (BDTD), Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI), Portal de Periódicos da Capes, Scientific Electronic Library Online (SciELO) e ScienceDirect. Os critérios de seleção incluíram artigos acadêmicos, teses e dissertações, capítulos de livros e livros que abordassem o tema central da pesquisa. Os termos, em português e inglês, utilizados para consulta nas bases foram: metadados, proveniência, metadados de proveniência, Família PROV e Dublin Core.

Para a análise dos dados, foi utilizado o método Crosswalk, desenvolvido pela National Information Standards Organization (NISO) em 1999. Esse método facilita a interoperabilidade semântica, permitindo que diferentes sistemas convertam dados de um padrão de metadados para outro, possibilitando a pesquisa simultânea em bases de dados heterogêneas como se fossem uma única base (Chan; Zeng, 2006; Baca, 2016).

Embora a W3C tenha realizado um mapeamento de correspondência entre os padrões de metadados Dublin Core e a família de metadados de proveniência PROV, tal abordagem não abrangeu todas as classes do Dublin Core, limitando a abrangência da análise. Em contraste, a presente pesquisa se diferencia ao incluir, em sua maioria, as classes do Dublin Core no mapeamento Crosswalk, fortalecendo assim a análise dos dados e a compreensão da interoperabilidade semântica. Além disso, não foram localizadas evidências do uso completo do método de Crosswalking pela W3C.

Dentro desse contexto, a pesquisa propõe realizar a transposição das classes de metadados da família PROV para o padrão Dublin Core, a fim de verificar a compatibilidade da proveniência de informações conforme os princípios do PROV dentro do padrão de metadados Dublin Core.

A escolha pelo Crosswalk é justificada pelo objetivo de examinar a interoperabilidade entre os padrões de metadados, verificando como a proveniência, conforme definida pela família PROV, pode ser mapeada e integrada ao padrão Dublin Core, contribuindo para a gestão e compreensão da proveniência em ambientes digitais.

3 APRESENTAÇÕES DOS DADOS

Com base no objetivo delineado neste estudo, são apresentados a seguir os dados obtidos da coleta de metadados Dublin Core, considerados metadados de proveniência ou

que podem conter informações relacionadas à proveniência, de acordo com as diretrizes da W3C, dentro da plataforma DSpace.

De acordo com a W3C, diversos termos do Dublin Core podem ser utilizados para descrever aspectos de proveniência de um recurso, como quando ele foi modificado no passado, quem foi responsável pelas modificações e de que maneira elas ocorreram. Os demais termos do DCMI (Iniciativa de Metadados Dublin Core) fornecem informações sobre o que foi modificado.

Em resumo, o W3C destaca os termos do Dublin Core que podem ser considerados como informações de proveniência, conforme mencionado anteriormente. É importante observar que a maioria desses termos está incluída na categoria de "Metadados Descritivos". Embora a definição dessa categoria seja amplamente semelhante entre os autores, as diferenças residem nas subcategorias usadas para classificá-los (Arakaki, 2019). A definição de metadados administrativos tem sido descrita da seguinte maneira:

[...] criados para fins de gerenciamento, tomada de decisão e manutenção de registros. Eles fornecem informações sobre os requisitos técnicos, de preservação e armazenamento de recursos de informações, especialmente objetos digitais. Metadados administrativos auxiliam no monitoramento, acesso, reprodução, digitalização e backup de recursos digitais. (JOURNEY; TAYLOR; WISSER, 2018, p. 191-192).

O padrão Dublin Core é composto por vocabulários que incluem classes e propriedades. No presente estudo, foram consideradas apenas as classes, com base nas sugestões fornecidas pela Família PROV. A correspondência entre cada elemento foi determinada a partir da análise das definições e das orientações de uso recomendadas pelas organizações responsáveis por ambos os padrões. Esse levantamento permitiu identificar a relação entre as classes dos termos do Dublin Core e as da PROV, utilizando o método Crosswalk.

Quadro 1 – Crosswalk DC term para PROV term

Dublin Core	PROV	Grau de correspondência	Observações
Classes			
Agent	Agent	Um-para-um	<i>Crosswalking</i> absoluto. Ambos dct:Agente prov:Agentreferem-se ao mesmo conceito: um recurso que tem o poder de agir (que então tem responsabilidade por uma atividade, entidade ou outro agente).

IV Seminário Informação, Inovação e Sociedade

São Carlos, SP • 22 a 25 de outubro de 2024

AgentClass	-	Um-para-nenhum	
BibliographicResource	Entity	Um-para-muitos	<i>Crosswalking</i> relativo. Um recurso bibliográfico refere-se a livros, artigos etc., que são entidades concretas do PROV.
FileFormat	-	Um-para-nenhum	
Frequency	-	Um-para-nenhum	
Jurisdiction	-	Um-para-nenhum	
LicenseDocument	Entity	Um-para-muitos	<i>Crosswalking</i> relativo. Documento que concede permissão para fazer algo em relação a um recurso. Assim, ele é mapeado como um tipo de prov:Entity.
LinguisticSystem	Plan	Um-para-muitos	<i>Crosswalking</i> relativo. A dct:LinguisticSystem é um sistema de símbolos, sons, gestos, etc. usado na comunicação. Portanto, o sistema linguístico define o plano a seguir para aprender um determinado idioma.
Location	Location	Um-para-um	<i>Crosswalking</i> absoluto. Ambos dct:Location e prov:Location definem locais como "regiões espaciais ou locais nomeados".
LocationPeriodOrJurisdiction	-	Um-para-nenhum	
MediaType	-	Um-para-nenhum	
MediaTypeOrExtent	-	Um-para-nenhum	
MethodOfAccrual	Plan	Um-para-muitos	<i>Crosswalking</i> relativo. dct:MethodOfAccrual define o método pelo qual os itens são adicionados a uma coleção (ou seja, o método prov:Plan seguido na atividade de inserção).
MethodOfInstruction	Plan	Um-para-muitos	<i>Crosswalking</i> relativo. "Processo que serve para gerar conhecimento, atitude e habilidades". Como dct:MethodOfInstruction define o método associado a uma atividade, ele é mapeado como uma subclasse de prov:Plan.
PeriodOfTime	-	Um-para-nenhum	
PhysicalMedium	-	Um-para-nenhum	
PhysicalResource	Entity	Um-para-muitos	<i>Crosswalking</i> relativo. Uma coisa material, que é um tipo concreto de prov:Entity.
Policy	Plan	Um-para-muitos	<i>Crosswalking</i> relativo. dct:Policy é definido como "um plano ou curso de ação de uma autoridade, destinado a influenciar e determinar decisões, ações e outros assuntos." Esta é uma especialização de prov:Plan.

ProvenanceStatement	Bundle	Um-para-um	Crosswalking absoluto. A dct:ProvenanceStatement is defined as "A statement of any changes in ownership and custody of a resource since its creation", which is a container for any provenance related assertion.
RightsStatement	Entity	Um-para-muitos	Crosswalking relativo. Declaração sobre os direitos intelectuais de um recurso (por exemplo, um Documento). Assim, ele é mapeado como um tipo de prov:Entity.
SizeOrDuration	-	Um-para-nenhum	
Standard	-	Um-para-nenhum	

Fonte: Elaborado pelos autores

De acordo com os dados coletados a partir da realização do Crosswalk do padrão Dublin Core para o modelo sugerido da Família PROV, foi observado, na classe, que dos 22 elementos DC apenas metade possuem equivalência, sendo que apenas 3 elementos tidos como Crosswalking absoluto e 8 Crosswalking relativo, já o Grau de Equivalência apontado é: 3 Um-para-um, 8 Um-para-muitos e 11 Um-para-nenhum.

4 CONCLUSÃO

Com base nos resultados obtidos, observou-se que há uma equivalência significativa entre as classes de metadados, evidenciando uma correspondência de 50% entre as classes do Dublin Core (DC) e as da Família PROV. Esse nível de correspondência sugere que, embora existam semelhanças consideráveis entre os dois modelos, ainda há áreas onde as diferenças estruturais ou semânticas podem dificultar a correspondência direta.

O Crosswalking dos padrões DC para a PROV reforça a possibilidade de alinhamento e compatibilidade entre os dois modelos, conforme os objetivos da W3C. Ao preencher a lacuna entre as comunidades DC e PROV, essa correspondência permite uma integração mais fluida entre os modelos, fornecendo informações valiosas sobre as características distintas de ambos. Isso não apenas fortalece o entendimento entre as duas abordagens de metadados, mas também facilita a adoção do modelo PROV em contextos em que o Dublin Core já é amplamente utilizado.

Além disso, a integração desses padrões contribui para um melhor rastreamento da proveniência de dados em ambientes digitais, uma vez que permite mapear e documentar o ciclo de vida dos recursos de forma mais detalhada. Ao combinar a capacidade descritiva do

DC com o enfoque na proveniência da PROV, essa correspondência pode auxiliar na construção de repositórios mais confiáveis e robustos, onde a rastreabilidade e a integridade das informações sejam garantidas de maneira mais eficiente.

REFERÊNCIAS

ARAKAKI, Felipe Augusto. **Metadados administrativos e a proveniência dos dados: modelo baseado na família prov**. 2019. 139 f. Tese (Doutorado) - Curso de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista, Marília-Sp, 2019. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/180490/arakaki_fa_dr_mar.pdf?sequence=5&isAllowed=y . Acesso em: 15 agosto 2024

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. Disponível em: <https://www.dlib.org/dlib/june06/chan/06chan.html#Zeng-Xiao>. Acesso em 15 agosto 2024

DCMI - **Dublin Core Metadata Initiative**. Disponível em: <https://www.dublincore.org/>. Acessado em 21 de maio de 2024.

DCMI. **Metadata Terms**. <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>. Acesso em: 15 de maio de 2024.

Dublin Core to PROV Mapping. <https://www.w3.org/TR/prov-dc/#provenance-in-dublin-core>. Acessado em 8 de maio de 2024.

JOUDREY, Daniel N.; TAYLOR, Arlene G.; WISSER, Katherine M. **The organization of information**. Fourth edition ed. Santa Barbara, California: Libraries Unlimited, 2018. (Library and information science text series).

Metadata and Bitstream Format Registries - **DSpace 7.x Documentation** - LYRISIS Wiki. [https://wiki.lyrasis.org/display/DSDOC7x/Metadata+and+Bitstream+Format+Registries#MetadataandBitstreamFormatRegistries-LocalMetadataRegistry\(local\)](https://wiki.lyrasis.org/display/DSDOC7x/Metadata+and+Bitstream+Format+Registries#MetadataandBitstreamFormatRegistries-LocalMetadataRegistry(local)). Acesso em 2 de junho de 2024.

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