XX ENANCIB
21 a 25 Outubro/2019 – Florianópolis
A Ciência da Informação e a era da Ciência de Dados

ISSN 2177-3688

GT-8 – Informação e Tecnologia

PRESERVAÇÃO DIGITAL E PROVENIÊNCIA: INTERSEÇÕES ENTRE PREMIS E O PROV

DIGITAL PRESERVATION AND PROVENANCE: INTERSECTIONS BETWEEN PREMIS AND PROV

Felipe Augusto Arakaki – UnB
Rachel Cristina Vesu Alves – Unesp
Plácida Leopoldina Ventura Amorim da Costa Santos - Unesp

Modalidade: Trabalho Completo

Resumo: Os metadados desempenham diversas funções nos ambientes informacionais digitais, dentre elas destacam-se a preservação digital e a proveniência dos dados. O problema norteador do estudo está no estabelecimento de metadados para a preservação digital e proveniência para seu uso e reuso em ambientes informacionais digitais. A hipótese é que os metadados de preservação digital são essenciais para garantirem o acesso a longo prazo ao patrimônio cultural e que os metadados de proveniência possibilitam a identificação da origem dos dados compartilhados e maior credibilidade das informações reutilizadas. Por meio de uma metodologia exploratória e de levantamento bibliográfico, o objetivo do trabalho é abordar os principais aspectos teóricos que caracterizam os metadados de preservação digital e de proveniência e a relação entre os padrões PREMIS e PROV. Os resultados deste trabalho demonstram o uso conjunto do PROV-O e do PREMIS, pois o PROV fornece a descrição generalizada de procedência e o PREMIS a descrição de metadados de preservação. Apesar da estrutura PROV apresentar um modelo mais geral, o PROV permite que outros padrões possam utilizar e complementar seus metadados para melhor identificação da proveniência em diversos contextos. Assim, considera-se que o uso conjunto do PROV e do PREMIS, possibilita um registro mais detalhado da preservação digital e proveniência dos dados.
Palavras-Chave: Metadados; Preservação digital; Proveniência; PREMIS; PROV.

Abstract: Metadata has several functions in digital informational environments, among them digital preservation and data provenance. The guiding problem of the study is the establishment of metadata for digital preservation and provenance for its use and reuse in digital information environments. The hypothesis is that digital preservation metadata is essential to ensure long-term access to data of cultural heritage and provenance metadata make it possible to identify the origin of the shared data and the credibility of the reused information. Through an exploratory methodology and bibliographic survey, the objective of this work is to address the main theoretical aspects that characterize the digital preservation and provenance metadata and the relationship between PREMIS and PROV standards. The results of this work demonstrate that the joint use of PROV-O and PREMIS, since PROV provides the generalized description of origin and PREMIS the description of preservation metadata. Although the PROV framework presents a more general model, it allows other standards to use and complement their metadata to better identify provenance in various contexts. Thus, it is considered that the joint use of PROV and PREMIS allows a more detailed record of digital preservation and data provenance.
Keywords: Metadata; Digital Preservation; Provenance; PREMIS; PROV.

1 INTRODUÇÃO

Os metadados são elementos fundamentais para estruturação dos dados na era digital, pois possibilitam, dentre outros fatores, a realização dos processos de busca, recuperação, acesso e interoperabilidade em ambientes informacionais. Além das funções que tradicionalmente já desempenham como, por exemplo, descrição para acesso e recuperação, os metadados são responsáveis por atenderem a funções emergentes nos ambientes informacionais digitais. Dentre as funções destacam-se: o registro da autoria e propriedade intelectual; as formas de acesso e restrições de uso; valoração do conteúdo e visibilidade da informação; atualização e conservação dos recursos; preservação e proveniência (CHOWDHURY; CHOWDHURY, 2007; HAYNES, 2004; MÉNDEZ RODRÍGUEZ, 2002). Das funções emergentes citadas, destacam-se neste trabalho a preservação e proveniência, papéis fundamentais na era digital e nos ambientes informacionais.

Como contraponto ao desenvolvimento tecnológico, tem-se o constante problema da obsolescência de *hardware* e *software*. Neste cenário, a preservação digital do patrimônio cultural e seu acesso para gerações futuras são requisitos imprescindíveis de serem estabelecidos em diversos ambientes informacionais. Outro aspecto igualmente importante refere-se a expansão e popularização da publicação de dados na Web, necessitando de sistemas cada vez mais interoperáveis e da identificação da proveniência e registro das ações sobre os dados reaproveitados (ARAKAKI, 2019).

Neste contexto, o problema norteador deste estudo está no estabelecimento de metadados para proveniência e preservação digital e seu uso e reuso em ambientes informacionais digitais. A hipótese deste trabalho é que os metadados de preservação digital e de proveniência são elementos essenciais para garantirem que o patrimônio cultural produzido seja acessado ao longo do tempo e que a identificação clara da proveniência dos dados compartilhados possibilita maior credibilidade das informações reutilizadas.

Deste modo, o objetivo deste trabalho consistiu em abordar os principais aspectos teóricos que caracterizam os metadados de preservação digital, os metadados de proveniência e as relações de ambos com os metadados administrativos. Além disso, objetiva-se também abordar a relação entre os padrões PREMIS e PROV, para a preservação

digital e o acesso aos recursos informacionais a longo prazo e para o registro da proveniência para uso e reuso dos dados com consistência.

Este estudo caracteriza-se como qualitativo e a metodologia utilizada consistiu em uma combinação da abordagem exploratória da literatura disponível sobre o tema com pesquisa bibliográfica (GIL, 2010). O levantamento bibliográfico foi realizado em fontes bibliográficas (primárias e secundárias) da literatura científica impressa e digital, levando em consideração os seguintes temas: tipologia de metadados, metadados administrativos, metadados de preservação, metadados de proveniência, PREMIS, PROV. As bases de dados consideradas para o levantamento bibliográfico foram: Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI); *Library and Information Science Abstracts* (LISA); L*ibrary, Information Science and Technology Abstracts* (LISTA); Biblioteca Digital Brasileira de Teses e Dissertações (BDTD); Oasisbr: portal brasileiro de publicações científicas em acesso aberto do IBICT; P@rthenon; Portal de Periódicos da Capes; Scopus; Web of Science. Além dos Anais do *Dublin Core Metadata Initiative International Conference on Dublin Core and Metadata Applications*; *Google Scholar* e documentos do site do *World Wide Web Consortium* (W3C). A tipologia dos trabalhos considerados foram: Artigos, livros, teses, dissertações, trabalhos em eventos e documentos do W3C, nos idiomas em português, inglês e espanhol.

Após o levantamento e a identificação do corpus do trabalho, foi realizado uma leitura do resumo e quando necessário, uma leitura prévia do texto para que pudesse aplicar os critérios para seleção do material como: pertinência da temática do artigo para o escopo da pesquisa e atualidade dos documentos, ou seja, quando dois ou mais trabalhos tinham temáticas similares, foi considerado o trabalho mais atual. Uma vez identificado o material necessário, foi realizado leitura e fichamento dos textos, proporcionando a base teórica para a discussão dos diferentes pontos de vista identificados na literatura sobre o tema.

Dessa forma, a análise exploratória da literatura permitiu a construção do conhecimento teórico sobre os metadados administrativos e, a partir disso, foi realizada a análise e estudo sobre a proveniência em diversos contextos. Também permitiu o estudo dos metadados de preservação digital abordados neste trabalho. Ressalta-se que os principais aspectos sobre a proveniência apresentados neste trabalho são resultantes da tese de doutorado recém defendida sobre metadados administrativos, proveniência dos

dados e família PROV de um dos autores. Já os aspectos de preservação digital, são resultantes de pesquisas e estudos sobre o tema derivados da prática em sala de aula com ministração da disciplina de preservação digital.

Nesse contexto, na seção 2 deste trabalho é apresentada uma discussão da tipologia dos metadados de preservação digital e de proveniência. Na seção 3 é apresentada uma discussão da preservação digital e a importância do PREMIS para descrição de objetos para preservação digital. Logo em seguida, na seção 4 é discutida a questão da proveniência e um panorama dos padrões de metadados para proveniência com o foco no PROV. Por fim, foi apresentado uma discussão das relações entre os padrões da preservação digital (PREMIS) e da proveniência (PROV).

2 METADADOS DE PRESERVAÇÃO DIGITAL E METADADOS DE PROVENIÊNCIA: CONTEXTUALIZAÇÕES TIPOLÓGICAS

Existe uma variedade de categorizações de metadados na literatura. Para os autores Joudrey; Taylor e Wisser (2018) e a NISO (2004), os metadados podem ser categorizado em três grandes grupos: metadados descritivos, metadados estruturais e metadados administrativos. Entretanto, há também uma outra categorização mais comum adotada por outros autores que determinam cinco tipos básicos de metadados: administrativos, descritivos, técnicos, de uso e de preservação (GILLILAND, 1999, 2008, 2016; MÉNDEZ RODRIGUEZ, 2002; ZENG, QIN, 2008, 2016).

Basicamente os cinco tipos de metadados podem ser definidos da seguinte forma, de acordo com suas funções (GILLILAND, 1999, 2008, 2016; MÉNDEZ RODRIGUEZ, 2002; ZENG, QIN, 2008, 2016):

a) Metadados Administrativos: “[...] usados no gerenciamento e administração de coleções e recursos informacionais”;
b) Metadados Descritivos: “[...] usados para descrever e identificar informações sobre recursos;
c) Metadados de Preservação: “metadados relacionados à gestão de preservação de coleções e informacionais”;
d) Metadados Técnicos: “metadados relacionados com as funções do sistema e o comportamento dos metadados”;
e) Metadados de Uso: “metadados relacionados com o nível e tipo de uso de coleções e recursos informacionais” (GILLILAND, 2016).

É importante destacar que essas tipologias não são excludentes, um metadado pode ser classificado em mais de uma categoria, ou seja, um metadado de uma tipologia pode

pertencer ou exercer a função de outra, dependendo do contexto onde está sendo utilizado.

A variação dos metadados nas tipologias citadas dependem da função que o metadado exerce no momento de seu uso. Como exemplos destacam-se: um metadado descritivo que pode exercer a função de metadado administrativo, como a data de criação do recurso, que pode ser utilizada para o levantamento da criação de itens no sistema; um metadado de preservação que pode exercer a função de um metadado administrativo como, por exemplo, requisitos legais para acesso a longo prazo; um metadado de preservação que pode exercer a função de metadado técnico, quando descreve especificações sobre requisitos de *software*; ou ainda um metadado de proveniência que pode atender a função de uso e reuso dos dados, entre outros.

Com base na categorização de Gilliland (2016) pode-se considerar que os metadados de preservação, embora representem características próprias para preservação digital, também podem exercer funções nas categorias de metadados administrativos e, em alguns casos, categorias de metadados técnicos.

De acordo com Arakaki (2019) existem duas vertentes na literatura para a categorização dos metadados de preservação, uma que considera os metadados de preservação como subcategorias dos metadados administrativos (RILEY, 2004; POMERANTZ, 2015; GARTNER, 2016; RILEY, 2017; JOUDREY; TAYLOR E WISSER, 2018); e outra que considera como sendo uma categoria independente (GILLILAND, 1999, 2008, 2016; MÉNDEZ RODRÍGUEZ, 2002; SENSO E ROSA PIÑERO, 2003; HAYNES, 2004, 2017; ALVES, 2010; ALVES, SANTOS, 2013; ZENG, QIN, 2008; 2016).

Os metadados de preservação podem ser considerados como "[...] uma forma de metadados administrativos que documentam os processos de preservação executados em recursos em fluxos de trabalho tanto convencionais quanto de digitalização.", mas por atender a uma função específica, são considerados como sendo uma tipologia separada (GILLILAND, 1999, 2008, 2016; ZENG, QIN, 2008, 2016). Complementando essa definição a autora Gilliland (2008, 2016) aponta que os metadados de preservação estão relacionados com o

> [...] gerenciamento de preservação de coleções e recursos informacionais (exemplos incluem documentação das condições físicas de recursos, ações tomadas para preservar versões físicas e digitais de recursos (ex., atualização de dados e migração) e alterações durante a digitalização ou preservação).

Além de uma categoria independente e uma subcategoria de metadados administrativos, observa-se que os metadados de preservação também possuem relações com os metadados técnicos. Os metadados técnicos estão "[...] relacionados com o funcionamento dos sistemas ou o comportamento dos metadados (exemplos incluem informações sobre requisitos de *hardware* e *software*, digitalização técnica (como formatos, taxas de compressão e rotinas de escalonamento) e dados de autenticação e segurança (por exemplo, chaves de criptografia e senhas)." (GILLILAND, 2008, 2016). Com base nesta definição, observa-se que algumas dessas informações técnicas são dados que precisam ser preservados ao longo do tempo como, por exemplo, requisitos de *hardware* e *software*, migração e formatos de arquivo, entre outros.

Isso mostra que as categorias de metadados não são excludentes e que alguns dos metadados de preservação também podem pertencer a categoria de metadados técnicos.

É importante considerar que os recursos informacionais estarão sujeitos a diferentes usos, alterações e atualizações ao longo do seu ciclo de vida em um ambiente informacional digital, por isso, devem estar relacionados a um processo de preservação contínua. Portanto, o uso de metadados administrativos, descritivos, técnicos, de preservação e de uso, são importantes para registrar informações diversas. Além disso, torna-se importante no contexto do uso e reuso dos dados, o registro, preservação e migração dos próprios metadados, para que eles não fiquem obsoletos e desconectados dos recursos informacionais que descrevem (GILLILAND, 1999, 2008, 2016; ZENG, QIN, 2008, 2016; ARAKAKI, 2019).

Em relação a questão de uso e reuso dos dados, destaca-se neste trabalho a importância do uso dos metadados de proveniência. Constituem-se como um tipo específico de metadados, mas também, podem estar relacionados com as categorias anteriormente citadas e, portanto, exercer funções diferentes conforme seu uso e pertencer a mais de uma categoria de tipos de metadados.

Conforme analisado não há um consenso na literatura se os metadados de proveniência estão contemplados ou não na categoria de metadados administrativos. Para alguns autores (BACA, 2016; CHOWDHURY; CHOWDHURY, 2007; LIU, 2007; POMERANTZ, 2015; TAYLOR; JOUDREY, 2009; ZENG; QIN, 2008) os metadados administrativos podem englobar como subcategorias os metadados de preservação, metadados estruturais, metadados técnicos e metadados de uso e apenas Pomerantz (2015) classifica os

metadados de proveniência como subcategoria dos metadados administrativos (ARAKAKI, 2019).

Apesar dos metadados de proveniência não configurar com tipologia ou subcategoria dos metadados administrativos, diversos estudos já estão se preocupando na representação e identificação da proveniência das informações disponibilizadas no ambiente digital. De acordo com Glushko (2013), um recurso necessita da garantia de permanência ou persistência dos dados ao longo do tempo, isso garante a efetividade (tempo de vida ou de validade do recurso), autenticidade (certificação que o recurso possui em relação ao seu original) e proveniência dos dados (custódia do recurso). O autor ainda destaca duas formas de persistência, a primeira está relacionada aos identificadores persistentes e a segunda está relacionada com a persistência dos recursos que envolve a preservação do mesmo. Para Moreau e Missier (2013) a proveniência são informações referentes às entidades, as atividades e as pessoas envolvidas na produção de um dado ou coisa, que pode ser usado para formar avaliações sobre a sua qualidade, segurança ou confiabilidade (ARAKAKI, 2019).

3 PRESERVAÇÃO DIGITAL

De acordo com os estudos de Grácio (2012) a preservação digital pode ser estabelecida com base em três aspectos principais: o aspecto organizacional, o aspecto legal e o aspecto técnico. Os metadados são estabelecidos nos aspectos técnicos da preservação digital, mas devem ser considerados de modo amplo em todos os outros aspectos citados no momento da determinação da política de preservação digital.

O desenvolvimento constante das tecnologias ocasionam o inevitável problema de obsolescência de *hardware, software*, formatos de arquivo, além da degradação das mídias digitais. Para que o patrimônio cultural produzido e disponibilizado seja acessado ao longo do tempo, é preciso desenvolver programas de preservação digital em grande escala (ZENG; QIN, 2016). Assim, as diversas áreas do conhecimento que disponibilizam seus recursos informacionais em ambientes digitais, passaram a enfrentar o desafio de estabelecer a preservação digital de seus recursos e dados. Neste contexto, os metadados de preservação desempenham um papel fundamental em todo o processo de preservação digital.

Outro fator a ser estabelecido em uma política de preservação digital refere-se às estratégias de preservação a serem utilizadas (migração, emulação, encapsulamento etc)

para preservar o aspecto físico, de conteúdo e a autenticidade dos recursos informacionais, garantindo seu acesso ao longo do tempo. A escolha das estratégias de preservação deverão ser refletidas no conjunto de metadados utilizados na preservação digital.

As autoras Zeng e Qin (2016, p.474, tradução nossa), ressaltam que os metadados de preservação digital são "[...] a infra-estrutura de informação que suporta todo o processo associado à preservação digital", registram dados e informações que irão manter os recursos digitais acessíveis a longo prazo. As autoras consideram ainda que os metadados de preservação "[...] abordam vários processos-chave: a proveniência de um objeto digital arquivado (a história custodial do objeto), autenticidade, atividade de preservação, ambiente técnico e gerenciamento de direitos [...]" (ZENG; QIN, 2016, p.474, tradução nossa). Deste modo, os metadados de preservação são necessários para registrar as informações sobre o processo de digitalização e preservação dos recursos analógicos e os processos de preservação dos recursos nascidos digitais, além de todas as ações tomadas no processo de preservação a longo prazo (ZENG; QIN, 2016).

Os marcos para o estabelecimento dos aspectos técnicos na preservação digital na atualidade são: o *PREservation Metadata: Implementation Strategies* (PREMIS) e o *Reference Model for na Open Archival Information System* (OAIS). O PREMIS é uma iniciativa de 2003 da *Online Computer Library Center* (OCLC) e pelo *Research Libraries Group* (RLG) e posteriormente passou a ser administrado pela *Library of Congress* (EUA), que define um conjunto de elementos de metadados de preservação baseados no Modelo de Referência OAIS (ARAKAKI et al., 2018; ZENG; QIN, 2016).

O Modelo de Referência OAIS fornece uma base conceitual para os sistemas de preservação, com a taxonomia para os recursos informacionais em pacotes arquivados e uma estrutura para seus metadados associados (ZENG; QIN, 2016). De acordo com as autoras Zeng e Qin (2016, p. 476, tradução nossa) "O OAIS categoriza as informações necessárias para preservação como: Conteúdos de Informação, Informações de Representação, Informações de Descrição da Preservação (divididas em referência, contexto, proveniência e informações fixas) e Informações de Pacotes." O modelo OAIS não especifica nenhum método de implementação, mas serve de base para a análise dos metadados de preservação necessários em um sistema de preservação digital (ZENG; QIN, 2016).

O PREMIS é a implementação do modelo OAIS e define os metadados de preservação para um processo de preservação digital. O PREMIS é o resultado da tradução do "*A Metadata Framework to Support the Preservation of Digital Objects*", originado pelo grupo de trabalho que o desenvolveu, e propõe um conjunto de unidades semânticas implementáveis pelo dicionário de dados denominado "PREMIS *Data Dictionary for Preservation Metadata*" (ZENG; QIN, 2016). O PREMIS define os metadados de preservação como sendo "[...] as informações que um repositório usa para dar suporte ao processo de preservação digital" (ZENG; QIN, 2016, p. 476, tradução nossa).

O Dicionário de Dados PREMIS, define cinco entidades: *Environment* (Suporte), *Object* (Objeto), *Event* (Evento), *Agent* (Agente) e *Rights Statement* (Declaração de direitos). Suportes são Entidades Intelectuais capturadas e preservadas no repositório como Representações, Arquivos e / ou Bitstreams. São também tecnologias (*software* ou *hardware*) de um Objeto Digital de alguma forma (por exemplo, renderização ou execução). O Objeto (ou Objeto Digital) pode ser definido como uma unidade discreta de informações sujeita a preservação digital e usado como parte do processo de preservação. A entidade Evento é uma ação que envolve ou afeta pelo menos um objeto ou agente associado ou conhecido. O Agente pode ser uma pessoa, organização ou programa / sistema de software associado a eventos na vida de um objeto ou com direitos associados a um objeto. E por fim, a Declaração de direitos é afirmação de um ou mais direitos ou permissões pertencentes a um objeto e / ou agente (PREMIS…, 2015; ARAKAKI et al., 2018).

Cada entidade será representada por um conjunto de metadados de tipos diferentes. Assim, o Dicionário de Dados PREMIS também ressalta a relação dos metadados de preservação com outros tipos de metadados, em especial, os metadados técnicos:

> Metadados técnicos descrevem as características físicas e não intelectuais dos objetos digitais. Metadados técnicos detalhados e específicos de formato são claramente necessários para implementar a maioria das estratégias de preservação, mas o grupo não tinha tempo nem conhecimento para lidar com metadados técnicos específicos de formato para vários tipos de arquivos digitais. Portanto, restringiu os metadados técnicos incluídos no Dicionário de Dados às unidades semânticas que eles acreditavam aplicar a objetos em todos os formatos. O desenvolvimento adicional de metadados técnicos é deixado para os especialistas formatarem. (PREMIS EDITORIAL COMMITTEE, 2015, p. 32, tradução nossa).

O PREMIS também destaca a importância do registro da proveniência digital e a inclusão de outras categorias de metadados para o registro de informações importantes para a preservação como, por exemplo, metadados descritivos, metadados sobre agentes, metadados técnicos, metadados sobre mídias e *hardware*, metadados sobre regras de negócios e informações sobre ações de preservação (ZENG; QIN, 2016).

4 PROVENIÊNCIA

A proveniência é utilizada em diversos contextos como nas Artes, na Museologia, na Arquivologia, na Computação entre outras áreas. O conceito está relacionado principalmente na identificação do responsável pela criação, guarda e gerenciamento de um recurso informacional para garantir a autenticidade das informações prestadas. Assim, a proveniência pode descrever pessoas, instituições, entidades e atividades envolvidas na produção e entrega de um recurso informacional. De acordo com Gil e Miles (2013),

> A proveniência pode ser usada para muitos propósitos, como entender como os dados foram coletados para que possam ser usados de forma significativa, determinar propriedade e direitos sobre um objeto, fazer julgamentos sobre informações para determinar se confiar nele, verificar se o processo e as etapas usadas para obter um resultado está em conformidade com determinados requisitos e reproduzindo como algo foi gerado.

Segundo Gil e Miles (2013) a proveniência pode representar três (3) diferentes perspectivas e tipos de informação. A primeira perspectiva está relacionada ao agente, ou criador, podendo ser uma pessoa ou organização, garantindo assim a identificação de quem criou o recurso informacional. A segunda perspectiva está atrelada ao próprio recurso informacional, identificando por exemplo, sua origem. A terceira perspectiva está relacionada aos processos, registrando as ações, etapas e alterações para construção e conservação do recurso informacional.

No contexto da preservação digital, a proveniência é vital para registrar informações dos responsáveis pela criação, custódia, alteração, curadoria e administração do recurso digital. Haynes (2018, p. 134, tradução nossa) complementa que “No contexto dos materiais digitais, fornecer informações de proveniência pode ajudar a demonstrar que um registro não foi adulterado e que a evidência que ele apresenta é, portanto, confiável.” De acordo com o dicionário da *Library of Congress* conjunto ao Premis Editorial Committee (2018, p. 211, tradução livre),

> Proveniência digital: documentação de processos no ciclo de vida de um objeto digital. Proveniência Digital tipicamente descreve Agentes responsáveis pela custódia e administração de Objetos Digitais, eventos-chave que ocorrem ao longo do ciclo de vida do objeto digital e outras informações associadas à criação, gerenciamento e preservação do objeto digital.

Paralelo às questões de preservação digital, a proveniência é fundamental para identificação, autenticidade e confiabilidade das informações compartilhadas no ambiente digital. De acordo com Haynes (2018, p. 134, tradução livre) "Quando se trata de estabelecer a autenticidade de um item, sua história torna-se importante, sua proveniência: as circunstâncias de sua criação, quem a possuiu e as condições sob as quais sua propriedade foi transferida." Segundo Moreau e Groth (2013, p. 4, tradução nossa),

> No contexto da Web, proveniência é um registro que pode ser criado, trocado e processado por computadores. [...] O registro processável por computador contém descrições dos eventos ocorridos, levando para um recurso ou uma coisa, como existe em algum contexto. Muitos fatores podem contribuir para tal estado de assuntos, incluindo as pessoas envolvidas, as organizações em que atuam em nome dos processos que estão sendo executados e outros dados, recursos ou coisas que fazem parte dele.

As informações necessárias para garantir a proveniência e autenticidades dos registros podem incluir quem e quando um determinado recurso informacional foi acessado, quais mudanças foram realizadas, entre outras informações. Conforme destacado por Haynes (2018, p. 134, tradução nossa), "O gerenciamento de registros e a boa governança dependem da capacidade de demonstrar a autenticidade de um registro, e fornecer a documentação sobre seu histórico e a maneira como ele foi gerenciado." Essas informações permitem identificar mudanças e alterações no registro, garantindo assim, a verificação da autenticidade das informações prestadas.

O registro das atualizações de um recurso é fundamental para garantir a autenticidade das informações prestadas. O controle de versão, quem fez a atualização e quando são informações que devem persistir no registro informacional do recurso. Dessa forma, os metadados são fundamentais pois,

> [...] podem fornecer um registro da proveniência de um documento e evidências de que ele foi mantido para estabelecer padrões e seguir procedimentos definidos. Isso é vital para documentos que foram escaneados e digitalizados e onde o original foi destruído, bem como os documentos nascidos digitais. (HAYNES, 2018, p. 134-135, tradução nossa).

Segundo Hayes (2018), o registro de informações a partir dos metadados, auxilia na veracidade e integridade dos recursos informacionais digitais. No mesmo contexto que eram estabelecidas as autenticidades de recursos informacionais com assinaturas, selos ou marcas d'água em papel como contratos e testamentos, os metadados auxiliam na identificação das informações de pessoas envolvidas no processo de criação e alteração do recurso, as condições de uso e alterações realizadas ao longo dos anos.

Diante desse contexto, algumas iniciativas têm explorado a questão de como representar a proveniência, como exemplo, o *Open Provenance Model* (OPM), *Provenance, Authoring and Versioning* (PAV) e o PROV (*Provenance*).

O OPM é um modelo conceitual de proveniência que define quais informações são necessárias em um sistema de proveniência (CRUZ; CAMPOS; MATTOSO, 2009). As discussões da construção do OPM tiveram início em 2006, no primeiro *International Provenance and Annotation Workshop* (IPAW), mas só foi lançado para comunidade em 2007. A proposta do OPM foi definir um modelo de dados que seja aberto do ponto de vista da interoperabilidade, mas também com relação à comunidade de seus colaboradores, revisores e usuários (MOREAU et al., 2011).

O PAV é um padrão focado na questão garantia da proveniência e identificação de pessoas e organizações e suas funções. Isto é, focado em quem criou, contribuiu e faz a curadoria dos dados e não faz uma abordagem específica dos processos conforme destacado por Ciccarese et al. (2013).

Entretanto, com o grupo de trabalho de proveniência da W3C com pesquisadores que estudavam o OPM, surgiu o PROV. De acordo com Moreau e Groth (2013) e Bivar et al. (2013, p. 2, tradução nossa) "O uso de proveniência, independente do modelo utilizado, fornece um fundamento essencial para avaliar a autenticidade de dados, permitindo confiabilidade e reprodutibilidade." Para Bivar et al. (2013) com o surgimento do PROV, há uma probabilidade de migração dos sistemas que utilizam OPM para PROV, pois o PROV é apoiado por uma instituição de peso como o W3C. Nesse contexto, foi abordado o PROV para discussões e construção do modelo desta tese.

De acordo com Haynes (2018, p. 135) "O PROV é um padrão para metadados de proveniência, que é hospitaleiro para fornecer metadados de outros esquemas. É baseado em um modelo de Agente, Entidade e Atividade [...]" e apresenta um modelo geral para representar informações de proveniência. Destaca-se que a proposta do padrão PROV não

é abranger todas as especificidades de vários domínios, mas fornecer um conjunto de metadados para garantir um mínimo de informações de proveniência aplicável a todos domínios.

O grupo do W3C, responsável pelas discussões sobre a proveniência, publicaram diversos documentos que são conhecidos como família PROV. A família de documentos PROV conta com quatro recomendações, *The PROV Data Model* (PROV-DM), *The PROV Ontology* (PROV-O), *The Provenance Notation* (PROV-N) e *Constraints of the PROV Data Model* (PROV-CONSTRAINTS). Ainda foram publicadas oito (8) notas que auxiliam no mapeamento e nas informações sobre o modelo PROV.

Para iniciar os estudos sobre PROV, recomenda-se a leitura do PROV-PRIMER que oferece uma introdução ao modelo de dados de proveniência. O PROV-O define uma ontologia OWL2 para o modelo de dados de proveniência, destinado à comunidade Linked Data e Web Semântica. O PROV-DM define um modelo de dados conceitual para a proveniência, incluindo diagramas UML. O PROV-N define uma notação legível para o modelo de proveniência. Isso é usado para fornecer exemplos dentro do modelo conceitual, bem como usado na definição de PROV-CONSTRAINTS. Já o PROV-CONSTRAINTS, define um conjunto de restrições no modelo de dados PROV que especifica uma noção de proveniência válida. A partir desses documentos, é possível ter uma base para uma descrição geral e inclusão da proveniência em registros dos recursos informacionais.

Dessa forma, o PROV estabelece três (3) classes principais que norteiam toda a sua estrutura. Elas são: *Entity* (Entidades) que pode ser considerada qualquer coisa física, digital ou conceitual; *Activity* (Atividades) são definidas como ações e processos dinâmicos e são como entidades (*Entity*) que passam a existir quando seus atributos mudam para se tornarem novas entidades. Atividades podem gerar novas entidades, por exemplo, escrever um documento traz o documento à existência, enquanto a revisão do documento traz uma nova versão à existência (GIL; MILES, 2013); e *Agent* (agente), que de acordo com Yolanda Gil e Miles (2013, não paginado, tradução nossa), “Um agente pode ser uma pessoa, um software, um objeto inanimado, uma organização ou outras entidades que podem ser responsabilizadas.”

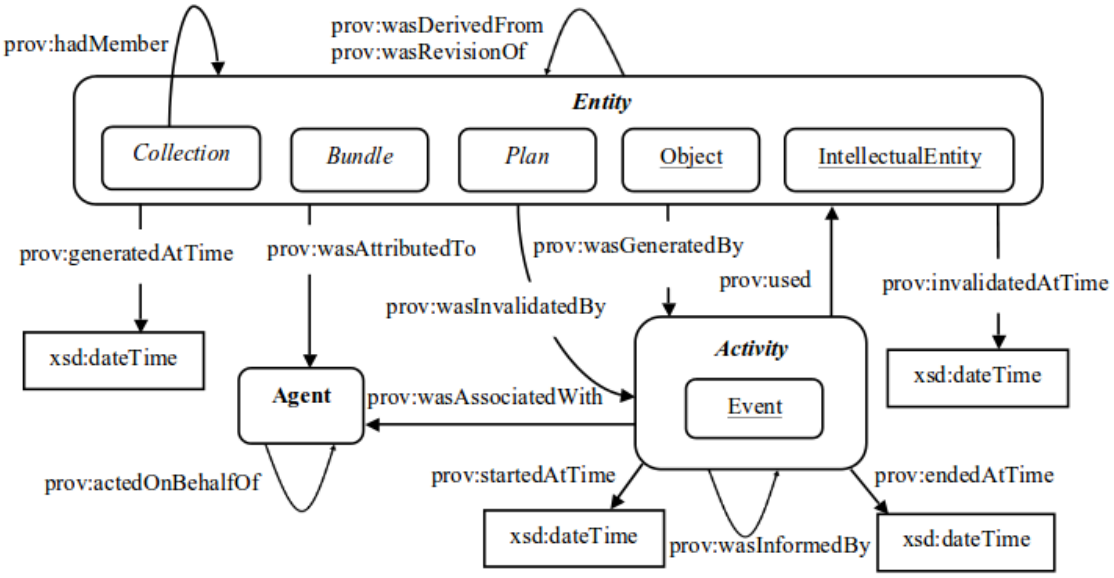
Outros conceitos fundamentais no PROV estão relacionados à função, derivação, revisão, planos e tempo. A **função** específica o relacionamento entre uma entidade com uma atividade e especificam como os agentes estão envolvidos em uma atividade. A

derivação e a **revisão** estão relacionadas às características do conteúdo de uma entidade. O PROV permite a descrição de alguns tipos de derivação, por exemplo, um documento pode passar por várias revisões ao longo do tempo e o resultado de cada revisão é uma nova entidade e permite relacionar essas entidades fazendo uma descrição de que uma foi uma revisão de outra. O PROV refere-se às **atividades** que seguem procedimentos predefinidos, como receitas, tutoriais, instruções ou fluxos de trabalho, como **planos**, e permite a descrição de que um plano foi seguido, por agentes, na execução de uma atividade. O registro do **tempo** é fundamental para proveniência, nesse sentido, o PROV é capaz de registrar o cronograma de eventos da entidade e/ou da atividade. (GIL; MILES, 2013). Entretanto, Arakaki (2019) destaca que em alguns casos, outras entidades alternativas e específicas podem ser requeridas, pois há diversas maneiras de descrever algo em um registro de proveniência, assim como, a necessidade de especificação de alguma entidade.

5 RESULTADOS E DISCUSSÕES

Estudos de Li e Sugimoto (2014, 2017, 2018) buscaram relacionar e destacar a proveniência para preservação digital e da importância de registrar as alterações e evolução dos padrões de metadados. O modelo proposto por Li e Sugimoto (2014) relaciona as principais classes do PROV-O com o PREMIS 2.2, com o intuito de criar um modelo de proveniência no contexto da preservação digital, conforme apresentado na figura 1.

Figura 1 - Relação entre o PREMIS e PROV.



Fonte: Li e Sugimoto (2014)

Entre as principais classes mapeadas do PROV para o PREMIS, estão *Agent* (Agente), *Activity* (Atividades) e *Entity* (Entidade). As classes *premis:IntellectualEntity* e *premis:Object* são mapeados como subclasses de *prov:Entity*, pois, *Entity* é considerada uma classe mais ampla podendo ser algo físico, digital ou conceitual. A classe *premis:Event* é mapeado como uma subclasse de *prov:Activity*. Enquanto a classe *premis:Agent* e *prov:Agent* são equivalentes.

Ao analisar o mapeamento proposto, observa-se que no PREMIS, o Agente não está diretamente conectado ao Objeto. Assim no PREMIS, o Agente influencia o Objeto a partir de um Evento. Já no PROV, é possível relacionamento entre Agente, Entidade e Atividade diretamente. Outro ponto a se considerar é que PROV define relacionamentos entre Atividades e relacionamentos entre Agentes, enquanto o PREMIS não inclui esses relacionamentos. (FEDORA… 2015).

Destaca-se que o PROV-O e PREMIS devem ser usados em conjunto, pois o PROV é projetado para a descrição generalizada de procedência, enquanto o PREMIS foi projetado para a descrição de metadados de preservação. Assim, um não substitui o outro, e os termos do PREMIS podem enriquecer o poder expressivo do PROV e vice-versa.

Diante desse cenário, a *Library of Congress* publicou uma nova versão do PREMIS com adaptações para Web Semântica em 2015. A versão PREMIS 3.0 está baseada na *Web Ontology Language* (OWL), que consiste em uma linguagem para definir e instanciar Ontologias na Web. (MÉNDEZ; GREENBERG, 2012).

> A ontologia PREMIS OWL é uma formalização semântica do dicionário de dados PREMIS 2.2 e define um modelo conceitual para a informação de preservação de um arquivo digital. A ontologia PREMIS OWL permitiu à comunidade interessada expressar metadados de preservação no RDF, usando o modelo conceitual do PREMIS-DD e, como tal, pode ser usado para publicar a informação de preservação como Linked (Open) Data). (DI IORIO; CARON, 2016, p. 2, tradução nossa)

Nessa atualização, alguns conceitos do PROV foram incorporados ao PREMIS 3.0 com oito (8) classes e seis (6) propriedades utilizadas de fato (ARAKAKI, 2019). Isso reforça a importância da representação da proveniência para os registros com o foco na preservação digital. Coppens et al. (2015) fazem um panorama do PREMIS 3.0 e destacam a importância dos metadados de proveniência, em especial da iniciativa do PROV para identificação da procedência da informações no ambiente web.

Já o PROV, publicou uma recomendação específica para Web Semântica denominada de PROV-*Ontology* ou PROV-O. "A Ontologia PROV (PROV-O) define a codificação da Linguagem de Ontologia da *Web* OWL2 do Modelo de Dados PROV (PROV-DM). (LEBO; SAHOO; MCGUINNESS, 2013, não paginado, tradução nossa). Dessa forma, o PROV-O estabelece "[...] um conjunto de classes, propriedades e restrições que podem ser usadas para representar e intercambiar informações de proveniência geradas em diferentes sistemas e sob diferentes contextos." (LEBO; SAHOO; MCGUINNESS, 2013, não paginado, tradução nossa).

Apesar de apresentar um modelo geral, a estrutura do PROV permite que outros padrões possam utilizar e complementar seus metadados para melhor identificação da proveniência em diversos contextos.

5 CONSIDERAÇÕES FINAIS

O uso dos metadados para descrição de recursos informacionais são indispensáveis para recuperação em um sistema de informação. Além dos metadados descritivos, destaca-se que outros tipos de metadados também são fundamentais para uma representação mais completa, atendendo assim, a diversos propósitos como a preservação digital e proveniência dos recursos informacionais.

Conforme abordado neste artigo, os metadados de preservação garantem a identificação do recurso informacional após suas alterações e migrações, auxiliando na preservação do recurso e de seus metadados a longo prazo. Os metadados de preservação possibilitam a infra-estrutura necessária para a realização de todo o processo de preservação digital, independentemente do tipo de ambiente digital. Entretanto, para proporcionar o acesso a longo prazo aos recursos, é importante o uso de outros tipos de padrões de metadados além dos metadados de preservação como, por exemplo, metadados descritivos, técnicos e de proveniência.

Em paralelo, ressalta-se a importância dos metadados de proveniência para garantir a procedência dos envolvidos na criação e alteração de um recurso informacional, além do registros dessas alterações realizadas, auxiliando assim, na representação, integridade e confiabilidade das informações disponibilizadas.

Observou-se ainda, nas discussões apresentadas, que há uma tendência na adequação dos padrões PREMIS e PROV para atender aos conceitos da Web Semântica. Ou

seja, há uma preocupação em ampliar o uso dos padrões em diversos contextos. Foi descrito ainda, a relação entre o PREMIS e PROV, apresentando a importância desses padrões para preservação digital e proveniência e comprovando que esses padrões podem ser utilizados simultaneamente para representação, evidenciando assim, a complementaridade desses padrões na preservação digital e proveniência dos dados.

REFERÊNCIAS

ALVES, Rachel Cristina Vesu; SANTOS, Plácida Leopoldina Ventura Amorim da Costa. **Metadados no domínio bibliográfico**. Rio de Janeiro: Intertexto, 2013.

ARAKAKI, Felipe Augusto. **Metadados administrativos e a proveniência dos dados**: modelo baseado na família PROV. 2019. 139 f. Tese (Doutorado) - Doutorado em Ciência da Informação, Universidade Estadual Paulista "Júlio Mesquita Filho", Marília, 2019.

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

BIVAR, Bárbara et al. Uma Comparação entre os Modelos de Proveniência OPM e PROV. **Proceedings of BRESCi**, 2013.

CHOWDHURY, G. G.; CHOWDHURY, Sudatta. **Organizing information**. London: Facet, 2007.

CICCARESE, Paolo et al. PAV Ontology: Provenance, Authoring and Versioning. **Journal of Biomedical Semantics**, [Sl.], v. 4, n. 1, p. 37, 2013.

COPPENS, San et al. PREMIS OWL: a semantic long-term preservation model. **International Journal on Digital Libraries**, [Sl.], v. 15, n. 2–4, p. 87–101, 2015.

CRUZ, Sérgio Manuel Serra da; CAMPOS, Maria Luiza M.; MATTOSO, Marta. **Towards a taxonomy of provenance in scientific workflow management systems**. 2009, [S.l.]: IEEE, 2009. p. 259–266.

DI IORIO, A.; CARON, B. **PREMIS 3.0 Ontology**: Improving Semantic Interoperability ofPreservation Metadata. EUA: Library of Congress, 2016.

FENDORA. **Repository Home.** Duraspace 2015. Disponível em: https://wiki.duraspace.org/display/FF/Fedora+Repository+Home#app-switcher. Acesso em: 08 ago. 2019.

CHOWDHURY, G. G.; CHOWDHURY, Sudatta. **Organizing information**. London: Facet, 2007.

GARTNER, Richard. **Metadata**. New York, NY: Springer Berlin Heidelberg, 2016.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. [S.l: s.n.], 2010.

GIL, Yolanda; MILES, Simon. **PROV Model Primer**. 2013. Disponível em: https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/. Acesso em: 08 ago. 2019.

GILLILAND, Anne J. Setting the Stage. *In:* BACA, Murtha (org.). **Introd. Metadat**a. 1. ed. Los Angeles: Getty Research Institute, 1999.

GILLILAND, Anne J. Setting the Stage. *In:* BACA, Murtha (org.). **Introd. Metadata**. 2. ed. Los Angeles: Getty Research Institute, 2008.

GILLILAND, Anne J. Setting the Stage. *In:* BACA, Murtha (org.). **Introd. Metadata**. 3. ed. Los Angeles: Getty Research Institute, 2016.

GLUSHKO, Robert J. **The Discipline of Organizing**. 1 ed. ed. Massachusetts, EUA: The MIT Press, 2013. Disponível em: http://site.ebrary.com/id/10841924. Acesso em: 08 ago. 2019.

GRÁCIO, José Carlos Abbud. **Preservação digital na gestão da informação:** um modelo processual para as instituições de ensino superior. São Paulo: Cultura Acadêmica, 2012.

HAYNES, David. **Metadata for Information Management and Retrieval**: Understanding metadata and its use. [S.l.]: Facet Publishing, 2018.

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).

LEBO, Timothy; SAHOO, Satya; MCGUINNESS, Deborah. **PROV-O: The PROV Ontology**. Disponível em: <https://www.w3.org/TR/2013/REC-prov-o-20130430/>. Acesso em: 30 dez. 2018.

LI, Chunqiu; SUGIMOTO, Shigeo. Provenance description of metadata using PROV with PREMIS for long-term use of metadata. DCMI, 2014. **Anais**… p. 147–156.

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

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

LIBRARY OF CONGRESS; PREMIS EDITORIAL COMMITTEE. **PREMIS Data Dictionary for Preservation Metadata, Version 3.0 (Library of Congress)**. webpage. Disponível em: http://www.loc.gov/standards/premis/v3/. Acesso em: 08 ago. 2019.

LIU, Jia. **Metadata and its applications in the digital library:** approaches and practices. [S.l.]: Libraries Unlimited Westport, CT, 2007.

MENDEZ, E. GREENBERG, J. Linked data for open vocabularies and HIVE's global framework. **El profesional de la información**, [S.l.], , v. 21, n. 3, maio/jun. 2012.

MÉNDEZ RODRÍGUEZ, Eva Ma. **Metadatos y recuperación de información**. Gijón, Asturias: Ediciones Trea, 2002. (Biblioteconomía y administración cultural, 66).

MOREAU, Luc et al. The Open Provenance Model Core Specification (v1.1). **Future Generation Computer Systems**, [S.l.], v. 27, n. 6, p. 743–756, jun. 2011.

MOREAU, Luc; GROTH, Paul. Provenance: an introduction to prov. Synthesis Lectures on the Semantic Web: Theory and Technology. **Morgan&Claypool Publishers**, [S.l.], v. 3, n. 4, p. 1–129, 2013.

NISO. **Understanding Metadata.** Bethesda: NISO Press, 2004. Disponívem em: <http://www.niso.org/standards/resources/UnderstandingMetadata.pdf>. Acesso em: 25 jan. 2006.

POMERANTZ, Jeffrey. **Metadata**. Cambridge, Massachusetts; London, England: The MIT Press, 2015. (The MIT Press essential knowledge series).

PREMIS EDITORIAL COMMITTEE. **PREMIS Data Dictionary for Preservation Metadata**, Version 3.0. p. 283, 2015. Disponível em: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf. Acesso em: 08 ago. 2019.

RILEY, Jenn. **Understanding Metadata**. NISO Press: National Information Standards Organization (U.S.), 2004.

SENSO, José A.; ROSA PIÑERO, Antonio De la. El concepto de metadato: algo más que descripción de recursos electrónicos. **Ciência da Informação**, [S.l.], v. 32, n. 2, 2003.

TAYLOR, Arlene G.; JOUDREY, Daniel N. **The organization of information**. 3. ed. Westport, Conn: Libraries Unlimited, 2009. (Library and information science text series).

ZENG, Marcia Lei; QIN, Jian. **Metadata**. New York: Neal-Schuman Publishers, 2008.

ZENG, Marcia Lei; QIN, Jian. **Metadata**. 2. ed. London: facet publishing, 2016.