



Palabra Clave (La Plata)
ISSN: 1853-9912
palabraclave@fahce.unlp.edu.ar
Universidad Nacional de La Plata
Argentina

Proveniência e contexto digital: contribuições da ciência da informação

Arakaki, Felipe Augusto; Santos, Plácida Leopoldina Ventura Amorim da Costa

Proveniência e contexto digital: contribuições da ciência da informação

Palabra Clave (La Plata), vol. 10, núm. 2, 2021

Universidad Nacional de La Plata, Argentina

Disponible en: <https://www.redalyc.org/articulo.oa?id=350566284004>

DOI: <https://doi.org/10.24215/18539912e124>



Esta obra está bajo una Licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional.

Proveniência e contexto digital: contribuições da ciência da informação

Provenance and digital context: contributions from information science

Felipe Augusto Arakaki
Universidade de Brasília. Faculdade de Ciência da
Informação, Brasil
fe.arakaki@gmail.com

 <https://orcid.org/0000-0002-3983-2563>

DOI: <https://doi.org/10.24215/18539912e124>
Redalyc: <https://www.redalyc.org/articulo.oa?id=350566284004>

Plácida Leopoldina Ventura Amorim da Costa Santos
Universidade Estadual Paulista (UNESP). Faculdade de
Filosofia e Ciências, Brasil
placidasantos@gmail.com

 <https://orcid.org/0000-0003-4478-5623>

Recepción: 21 Mayo 2020
Aprobación: 08 Agosto 2020
Publicación: 01 Abril 2021

RESUMO:

O objetivo deste artigo é discutir o conceito da proveniência e evidenciar sua importância no ambiente digital, focado na perspectiva da Ciência da Informação. Como procedimentos metodológicos, caracterizou-se por ser uma pesquisa qualitativa e exploratória, a partir de uma revisão de literatura sobre o contexto da proveniência em diferentes domínios. O artigo apresenta uma discussão do termo proveniência em diversos contextos, como na Arquivologia, Museologia, Preservação digital e Computação e destaca sua importância no âmbito digital. Conforme discutida a relevância da proveniência em diferentes contextos, destaca-se a necessidade de estudos mais aprofundados, se os instrumentos para representar a proveniência são adequados para garantir a veracidade e a inalterabilidade das informações. Dessa forma, as discussões expostas neste artigo revelam possibilidades em identificar metadados para cada característica que a proveniência possa apresentar, além da viabilidade de ampliação para outros contextos.

PALAVRAS-CHAVE: Proveniência, Ambiente digital, Contexto da proveniência.

ABSTRACT:

The objective of this article is to discuss the concept of provenance and highlight its importance in the digital environment, focusing on the perspective of Information Science. As methodological procedures, it was characterized by being a qualitative and exploratory research, from a literature review on the context of provenance in different domains. The article presents a discussion of the term provenance in several contexts, such as in Archival Science, Museology, Digital Preservation and Computing, highlighting its importance in the digital sphere. As discussed the relevance of provenance in different contexts, the need for more in-depth studies is highlighted, analyzing whether the instruments to represent provenance are adequate to guarantee the veracity and inalterability of the information. Thus, the discussions exposed in this article reveal possibilities in identifying metadata for each characteristic that the provenance may present, in addition to the feasibility of expansion to other contexts.

KEYWORDS: Provenance, Digital environment, Context of provenance.

1. INTRODUÇÃO

Com o movimento de publicação de dados abertos, tornou-se crescente a importância de estruturar os ambientes digitais para que os usuários possam localizar e recuperar informações desejadas, e ainda ampliar as possibilidades de uso e reuso de dados por agentes humanos e não-humanos.

Diante desse contexto, há uma preocupação com a identificação e verificação da proveniência das informações compartilhadas, principalmente no contexto digital. A identificação adequada garante a

confiabilidade das informações compartilhadas, pois é possível verificar quem criou, quando criou, se houve modificação, entre outras informações.

Entretanto, a disparidade terminológica de diversas áreas tem causado pontos de vista e perspectivas diferentes sobre um mesmo objeto ou conceito. World Wide Web Consortium (2011) e Furner (2019) ressaltam a importância de discutir as terminologias utilizadas em diversas áreas para representar conceitos similares. No relatório publicado pela *World Wide Web Consortium* (W3C) (2011) destaca, por exemplo, que a comunidade da biblioteca e a comunidade da *Web Semântica* têm terminologias diferentes para conceitos semelhantes. “Cada comunidade tem seu próprio vocabulário, e isso reflete diferenças em seus pontos de vista. O entendimento mútuo deve ser promovido, pois os dois grupos trazem conhecimentos importantes para a construção de uma rede de dados.” (World Wide Web Consortium, 2011, não paginado).

Nesse sentido, a proveniência perpassa por diferentes contextos. Assim, Lemieux e imProvenance Group (2016) destacam a importância de discutir a questão da proveniência em um âmbito multidisciplinar ou interdisciplinar, pois, muitas vezes, diferentes áreas utilizam as mesmas tecnologias para representar informações de proveniência.

A partir desse contexto, este artigo é fruto da pesquisa de tese de doutorado desenvolvida por Arakaki (2019) com o objetivo de aprofundar as discussões sobre o conceito da proveniência e evidenciar sua importância no ambiente digital, focado na perspectiva da ciência da informação. Este artigo apresentou, além da seção de introdução, os procedimentos metodológicos e uma seção de discussão do termo proveniência em diversos contextos. Logo em seguida, é exposta uma análise do levantamento realizado, sobre as diversas vertentes da proveniência no contexto digital. Por fim, foram apresentadas as considerações finais.

2. PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa caracterizou-se por ser uma pesquisa qualitativa e exploratória. A análise exploratória foi realizada pelo estudo sobre a proveniência em diversos contextos, por meio de uma pesquisa bibliográfica. O recorte da pesquisa abrange pesquisas publicadas internacionalmente e no Brasil, nos idiomas: português, espanhol e inglês. Os procedimentos metodológicos foram divididos em quatro etapas conforme o que se segue:

A primeira etapa está relacionada ao levantamento bibliográfico que foi realizada em bases de dados específicas em meados de 2019, como: Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI); Repositório Questões em Rede – Coleções (Coleção BENANCIB); *Library and Information Science Abstracts* (LISA); *Library, Information Science and Technology Abstracts* (LISTA); Oasisbr: portal brasileiro de publicações científicas em acesso aberto do Ibict; P@rthenon; Scopus; *Web of Science*. Além de documentos do World Wide Web Consortium (W3C).

Para realização do levantamento, foi elaborada estratégia de busca com termos para localização dos materiais selecionados, como: proveniência, prov, metadados e dados nos idiomas: inglês, português e espanhol. Quando possível, a busca foi realizada pelo título, palavras-chave e resumos na última década, ou seja, 2008-2018. A tipologia dos trabalhos pesquisados foram: artigos, livros, teses, dissertações, trabalhos em eventos e documentos do W3C.

Após o levantamento e a identificação do *corpus* do trabalho foi realizada a segunda etapa: uma leitura do resumo e, quando necessário, uma leitura prévia do texto para que pudesse aplicar os seguintes critérios para seleção do material para fundamentação teórica do texto:

- a partir de uma leitura do resumo, foi verificada a relevância da temática do artigo para o escopo da pesquisa;
- idioma dos documentos (português, inglês e espanhol); e
- atualidade e pertinência do texto para pesquisa.

A terceira etapa caracterizou-se pela leitura e fichamento dos textos completos. Essa etapa teve como propósito o desenvolvimento da base teórica para a discussão dos diferentes pontos de vista identificados na literatura sobre o tema.

Logo em seguida, a quarta etapa caracterizou-se pela análise e estabelecimento das características fundamentais extraídas da literatura sobre a proveniência, para elucidação do problema e redação da pesquisa. Para analisar os diversos pontos de vista da proveniência, considerou-se o Perspectivismo proposto por Peterson (1996), pois

a adequação de uma perspectiva é sempre relativa, mas ela não se estabelece de modo arbitrário. Na escolha de uma perspectiva deve-se considerar a sua adequação ao usuário a uma situação determinada, a um processo, a uma ontologia e a uma meta específica. É a análise de cada um destes componentes que se pretende desenvolver no estudo dos processos que atuam nas diretrizes, modelagens e estruturas de sistemas para atendimento de necessidades de sujeitos em ambientes informacionais específicos. (Santos e Vidotti, 2009, não paginado).

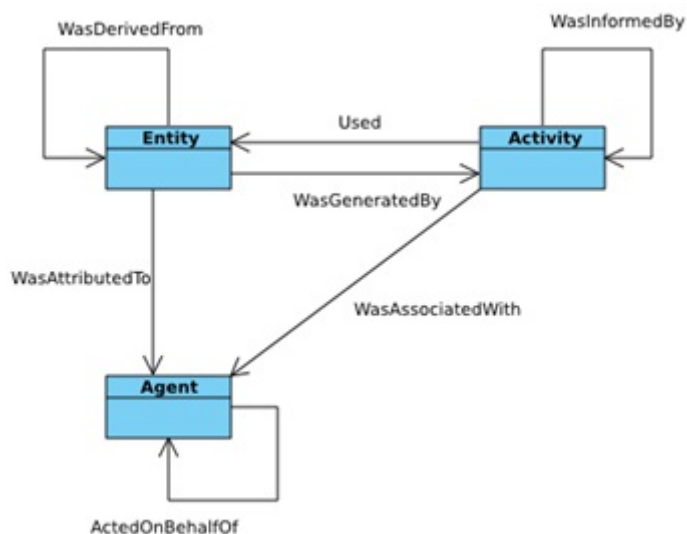
Cada uma das diversas discussões analisadas toma por base um ponto de vista da questão da proveniência e apresenta diferentes perspectivas.

3. CONCEITO DE PROVENIÊNCIA

O termo proveniência é empregado na identificação do responsável pela criação, guarda e gerenciamento de informações e recursos das mais diversas áreas. Segundo Moreau e Groth (2013, p. 4): “A proveniência é definida como um registro que descreve as pessoas, instituições, entidades e atividades envolvidas na produção, influência ou entrega de um dado ou coisa”.

Para World Wide Web Consortium (2013a) a proveniência pode representar três (3) diferentes perspectivas e tipos de informação. A primeira perspectiva está relacionada ao agente ou criador, que pode ser uma pessoa ou organização, que garante a identificação de quem criou o recurso. A segunda perspectiva está atrelada ao próprio recurso informacional, que identifica, por exemplo, sua origem. A terceira perspectiva está relacionada ao processo, registrando as ações e etapas tomadas para construção do registro informacional. Essas características foram instanciadas e são bases para o modelo criado pelo W3C, conforme apresentado na figura 1.

FIGURA 1
PROV-DM



Fonte: World Wide Web Consortium (2013b).

De acordo com Zafar et al. (2017) corresponde a informação que explica de onde veio, como foi derivado ou criado e quem esteve envolvido na criação, manipulação e processos aplicados. “A proveniência tem sido amplamente utilizada para a autenticidade de dados e processos” (Zafar et al., p. 50).

O termo ‘proveniência’ é utilizado em diversas áreas e de acordo com Arakaki, Alves e Santos (2019, p. 10): “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”. Segundo World Wide Web Consortium (2013a):

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 (não paginado).

De acordo com Pearce-Moses e Baty (2005, p. 317), a proveniência é definida como “1. A origem ou fonte de alguma coisa. - 2. Informações sobre a custódia das origens e propriedade de um item ou coleção”. No contexto da biblioteconomia e arquivologia, o *Dicionário de Biblioteconomia e Arquivologia* apresenta uma definição de proveniência mais próxima à Arquivologia, pois “indica a origem e procedência de documentos que, no decorrer das atividades de uma instituição, administração ou pessoa foram criados ou recebidos, acumulados e conservados” (Cunha, Cavalcanti, 2008, p. 301).

A proveniência na arquivologia garante a organicidade do fundo arquivístico a partir do produtor do documento e torna-se uma parte do conjunto de documentos criados por uma mesma instituição ou pessoa. Conforme Pearce-Moses e Baty (2005), a proveniência é um princípio fundamental dos arquivos, e refere-se ao indivíduo, família ou organização que criou ou recebeu os itens em uma coleção. O princípio da proveniência ou o respeito dos fundos corresponde ao fato de que registros de diferentes origens (procedência) sejam mantidos separados para preservar seu contexto. Segundo Duranti et al. (2016, p. 11): “Os relacionamentos entre registros e as organizações ou indivíduos que criaram, acumularam e/ou mantiveram e usaram na condução de atividades pessoais ou corporativas”.

Tognoli e Guimarães (2019) fizeram um levantamento histórico do princípio da proveniência focado na perspectiva da arquivologia. De acordo com os autores, “Quando o conceito de proveniência surgiu no contexto da ciência arquivística, no século XIX, ele tinha um objetivo pragmático: organizar o conjunto de registros cujas relações orgânicas com seus criadores haviam sido destruídas devido a uma classificação temática” (não paginado).

A definição do Conselho Internacional de Arquivos (2007, p. 10) sobre a proveniência está alinhada à definição de World Wide Web Consortium (2013a), e diz: “Relações entre os documentos e as organizações e/ou indivíduos que os produziram, acumularam e/ou mantiveram e usaram no curso de suas atividades pessoais ou corporativas. Proveniência é também a relação entre os documentos e as funções que geraram a necessidade dos documentos”. Nesse caso, os documentos são os recursos, as organizações ou indivíduos são os agentes e as funções podem estar relacionadas às atividades.

Em conformidade com o Arquivo Nacional (2015, p. 140), a proveniência é um “Termo que serve para indicar a entidade coletiva, pessoa ou família produtora de arquivo”. A proveniência é fundamental para organização do fundo arquivístico, assim, o princípio da proveniência garante a integridade arquivística e consistência do fundo arquivístico. Segundo o Arquivo Nacional (2015, p. 136), o princípio da proveniência é “Princípio básico da arquivologia segundo o qual o arquivo produzido por uma entidade coletiva, pessoa ou família não deve ser misturado aos de outras entidades produtoras. Também chamado princípio do respeito aos fundos”. A definição apresentada pelo Arquivo Nacional (2015) está focada, principalmente, nas pessoas ou entidades que produziram um determinado recurso e reforça uma das perspectivas destacadas por World Wide Web Consortium (2013a).

Ao analisar os trabalhos que falam sobre proveniência na arquivologia, Macedo (2018) realizou um levantamento da ocorrência do termo proveniência e observou escassez na utilização de termos e definições na arquivologia. Evidencia-se, assim, a necessidade de estudos mais aprofundados sobre a temática.

Como na arquivologia e em outras áreas, como museologia e artes, a proveniência é fundamental para garantir a procedência do item museológico ou obra de arte, também atesta a autenticidade como objeto único. Essas duas características contempla dois princípios observados por World Wide Web Consortium (2013a): ‘quem fez’ e ‘qual sua origem’.

O valor percebido de uma obra é alterado a partir de quem e quando ela foi criada e sua originalidade. Isso também é aplicado a livros e outros artefatos físicos, que podem ter um valor associado, como um manuscrito original ou uma primeira edição (Haynes, 2018). De acordo com o autor, “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” (Haynes, 2018, p. 134).

Na percepção da preservação digital, a proveniência é importante para registrar informações dos responsáveis pela criação, custódia, alteração, curadoria e administração de um recurso digital (Arakaki, 2019, Arakaki, Alves e Santos, 2019). Li e Sugimoto (2014, 2017 e 2018) apresentaram diversas discussões e relações entre a proveniência e a preservação digital. Haynes (2018, p. 134) complementa que “fornecer informações de proveniência pode ajudar a demonstrar que um registro não foi adulterado e que a evidência que ele apresenta é, portanto, confiável”. Observa-se nessa definição, a terceira perspectiva de World Wide Web Consortium (2013a), ou seja, as atividades ou ações que o recurso sofreu.

De acordo com o dicionário da *Library of Congress* conjunto ao Premis Editorial Committee (2015, p. 221):

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.

Tal definição, abrange as três perspectivas de World Wide Web Consortium (2013a), pois a documentação dos processos no ciclo de vida de um recurso corresponde às alterações sofridas. São destacados os responsáveis pelo recurso informacional e informação sobre a criação, gerenciamento e preservação do recurso (2013a).

Paralelo às questões de preservação digital, de acordo com Arakaki (2019) e Arakaki, Alves e Santos (2019), a proveniência é fundamental para identificação e confiabilidade das informações compartilhadas no ambiente digital. Haynes (2018, p. 134) explica que “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): “No contexto da Web, proveniência é um registro que pode ser criado, trocado e processado por computadores”. Ainda de acordo com os autores, “O registro processável por computador contém descrições dos eventos ocorridos, levando para um recurso ou uma coisa, como existe em algum contexto”.

Para garantir a confiabilidade e autenticidade dos dados, são necessários metadados, como quem criou o documento, última alteração, quando foi publicado, quais mudanças foram realizadas, além de informações sobre segurança de acesso do documento. Conforme destacado por Haynes (2018, p. 134), “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 e garante a verificação da autenticidade das informações prestadas. Ressalta-se que o controle de versão, quem fez a atualização e quando essas alterações foram realizadas são informações que devem persistir no registro informacional do recurso.

Ao analisar a proveniência como metadado passível de ser representado, Zafar et al. (2017, p. 50) esclarecem que “A proveniência é um metadado que descreve o histórico de dados e processos”. De acordo com o DCMI, o metadado proveniência corresponde a “Uma declaração de quaisquer alterações na propriedade e custódia do recurso desde a sua criação que são significativas por sua autenticidade, integridade e interpretação.” Ainda, destaca que “A declaração pode incluir uma descrição de quaisquer alterações feitas pelos custodiantes sucessivos no recurso” (Dublin Core Metadata Initiative, 2020, não paginado).

Entretanto, destaca-se que muitas vezes apenas um metadado não é suficiente para registrar todas as informações necessárias para descrever a proveniência, sendo necessário um conjunto de metadados para tal função. 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” (Haynes, 2018, pp. 134-135).

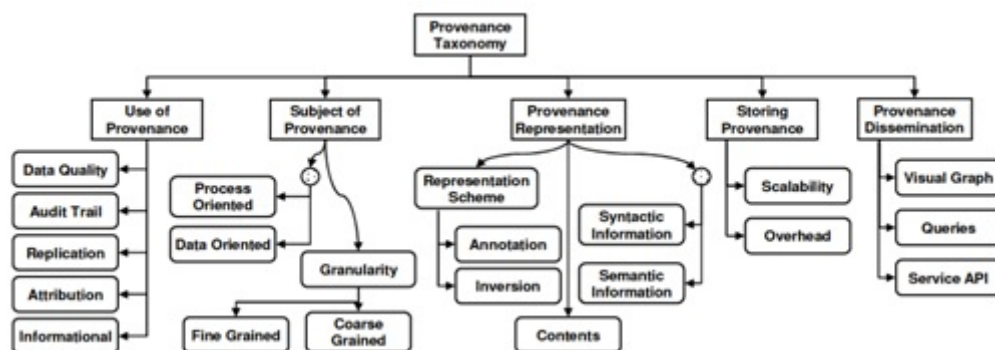
Segundo Haynes (2018), o registro de informações a partir dos metadados auxilia na veracidade e integridade dos recursos informacionais digitais. No mesmo contexto em que eram estabelecidas as autenticidades de recursos informacionais com assinaturas, selos ou marcas d’água em papel como contratos e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

No contexto da ciência da computação, diversos estudos foram encontrados. Dentre eles, destacou-se o trabalho de Marins (2008), que apresentou um modelo genérico de proveniência. Entretanto, observou-se que a partir da literatura a proveniência tem sido abordada, principalmente, no desenvolvimento de fluxos de trabalhos (*workflows*) para garantia da proveniência de dados científicos.

A proveniência de dados é tratada por Buneman, Khanna e Tan (2000, p. 316) como “a descrição das origens de um dado e o processo pelo qual ele chegou em um banco de dados”. A definição aborda dois dos três aspectos estabelecidos por World Wide Web Consortium (2013a), a origem e o processo do recurso. Destaca-se que na definição de agentes por World Wide Web Consortium (2013a) consideram-se agentes humanos e não-humanos. O reuso de dados sem a devida identificação pode resultar em um problema sério, como a não identificação da proveniência das informações prestadas.

Ao verificar nos requisitos de proveniência no contexto do reuso de dados, Simmhan, Plale e Gannon (2005) estabeleceram uma taxonomia das diversas possibilidades de informações da proveniência. Assim, foram estabelecidas cinco características principais relacionadas à proveniência: *use of provenance* (uso da proveniência), *subject of provenance* (objeto da proveniência), *provenance representation* (representação da proveniência), *storing provenance* (armazenamento da proveniência) e *provenance dissemination* (disseminação da proveniência), que representam categorias conforme apresentado na figura 2.

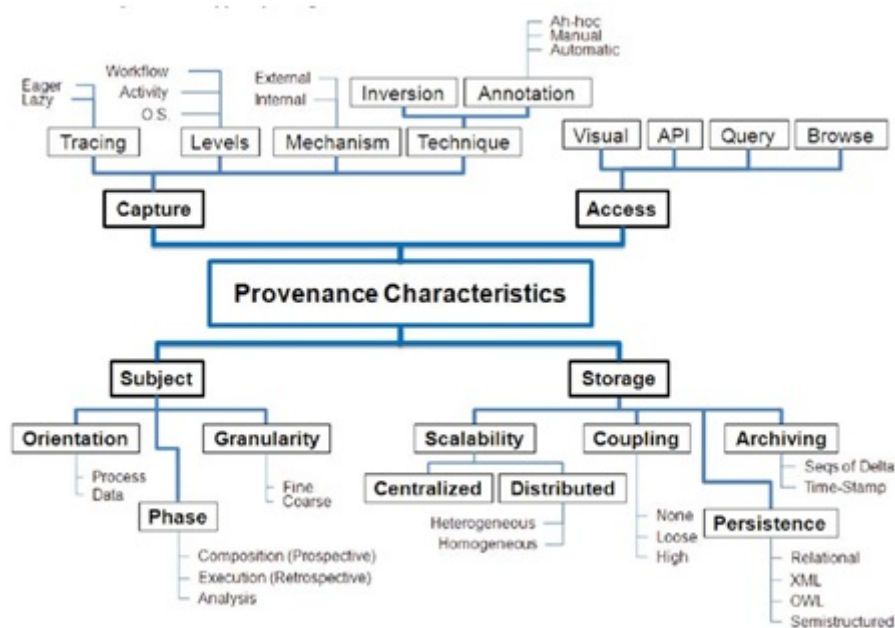
FIGURA 2
Taxonomia da proveniência



Fonte: Simmhan, Plale e Gannon (2005).

Diferente do trabalho de Simmhan, Plale e Gannon (2005) que possui uma concepção mais genérica, Cruz, Campos e Mattoso (2009) buscaram especificar a taxonomia para o ciclo de vida do fluxo de trabalhos científicos (*Scientific Workflow Management Systems - SWfMS*). Nesse contexto, a taxonomia apresentada por Cruz, Campos e Mattoso (2009) apresenta quatro categorias gerais: *capture* (captura), *access* (acesso), *subject* (objeto) e *storage* (armazenamento), conforme apresentado na figura 3.

FIGURA 3
Características do Sistema de Taxonomia de Proveniência

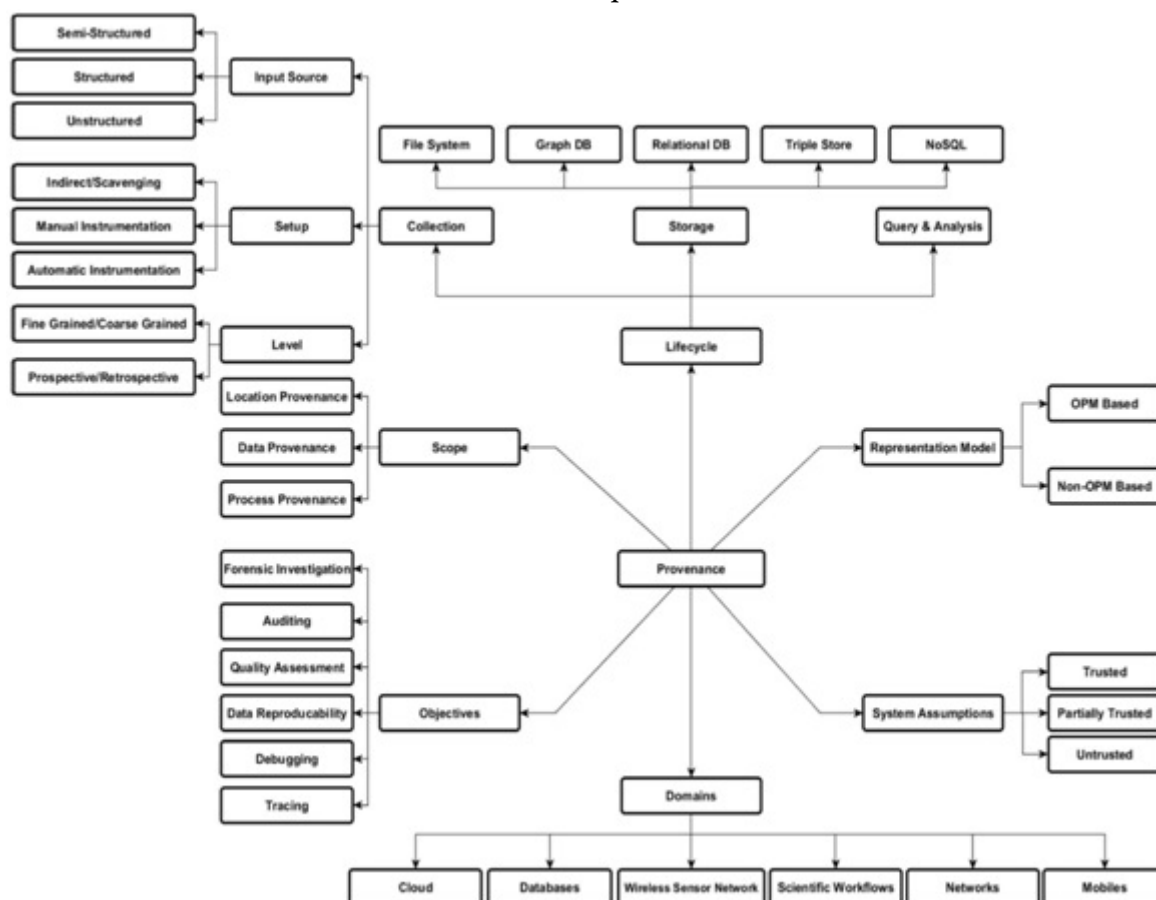


Fonte: baseado em Cruz, Campos e Mattoso (2009).

As diversas possibilidades de representação e captura da proveniência, apresentadas por Simmhan, Plale e Gannon (2005) e Cruz, Campos e Mattoso (2009), revelam as diversas dimensões e possibilidades de identificação da proveniência. Conforme observado no contexto científico, os metadados de proveniência são importantes para reprodutibilidade dos procedimentos de uma pesquisa.

Os trabalhos de Simmhan, Plale e Gannon (2005) e Cruz, Campos e Mattoso (2009), abordaram a questão da reprodutibilidade e reutilização de resultados e fluxos de trabalho experimentais e científicos. Diante deste contexto, Zafar et al. (2017) buscaram discutir como a confiança é alcançada a partir da proveniência segura. Na figura 4, é apresentado o ciclo de vida dos dados de proveniência em alguns contextos relacionados à ciência da computação.

FIGURA 4
Ciclo de vida da proveniência



Fonte: Zafar et al. (2017).

Diante da interdisciplinaridade da ciência da informação, diversos domínios apresentados por Zafar et al. (2017), também podem ser abordados pela ciência da informação, principalmente quando são discutidos banco de dados, *workflows científicos*, dispositivos móveis. Especialmente com a propagação da *e-science* e a disponibilização de conjunto de dados em repositórios digitais é de vital importância uma representação adequada, sobretudo para a equipe envolvida no levantamento dos dados, que precisa saber quando foi realizado o levantamento e se houve alteração após sua disponibilização na internet.

No contexto dos repositórios digitais, estudos como de Vidotti et al. (2016) destacaram a importância de identificar a origem dos registros coletados pelo Repositório Institucional UNESP das diversas bases de dados, para identificar as principais fontes de cada área, com o intuito de ampliar o número de fontes para a coleta de registros. Esses dados podem auxiliar uma instituição no mapeamento da produção científica a identificar quais áreas publicaram em um determinado periódico e quais bases de dados indexam esses periódicos. Isso pode auxiliar na aquisição de bases de dados, a partir das bases que mais possuem artigos da Universidade; mapear o comportamento de publicação das pesquisas da Universidade; e identificar a propagação e visibilidade da pesquisa, entre outras possibilidades.

Quanto à preocupação com a identificação da proveniência no âmbito científico destacam-se alguns eventos que reúnem pesquisadores do mundo todo, como o *International Provenance and Annotation Workshop* (IPAW) e o *Provenance Challenge*. O modelo de proveniência digital *Open Provenance Model* (OPM), publicado em 2006, é resultado do encontro *Provenance Challenge*, iniciado no primeiro workshop da IPAW (Bivar et al., 2013, p. 2).

Preocupados com a representação da proveniência no contexto digital, os trabalhos de Eckert, Garijo e Panzer (2011) e Eckert (2012) apresentaram um perfil de aplicação para metadados de proveniência e foi base para as discussões de um grupo de trabalho da W3C.

Em 2013, foi publicado um conjunto de documentos com recomendações, modelo de dados, padrões, ontologia, entre outras informações, resultado da discussão e pesquisa de diversos pesquisadores que trabalham com a questão da proveniência. Esse conjunto de documentos é conhecido 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 (Arakaki, 2019).

No início de 2017, o W3C publicou uma recomendação sobre boas práticas para publicação de dados na *Web* (*Data on the Web Best Practices W3C*). O documento apresenta 14 tópicos como recomendações para publicação de dados, entre eles, reforça a importância da proveniência dos dados, tópico “8.4 *Data Provenance*”, e categoriza quais os benefícios da inserção de informações da proveniência para práticas de publicação dos dados.

A partir da proveniência, como a origem e histórico, é possível que os consumidores de um conjunto de dados possam avaliar, a partir de um contexto interpretativo, a sua qualidade e determinar se devem confiar nesse conjunto de dados. Dessa forma, a proveniência é tratada como boa prática de publicação de dados, 5 - *Provide data provenance information*, com o fornecimento de informações sobre a origem dos dados e a indicação de quaisquer alterações realizadas. (World Wide Web Consortium, 2017).

Para testar essa boa prática, é necessário verificar se os metadados do conjunto de dados incluem informações de proveniência sobre a origem e alterações do *dataset* em um formato legível por humanos. Os benefícios dessa boa prática correspondem ao reuso, compreensão e confiança (World Wide Web Consortium, 2017).

A proveniência ganhou destaque em discussões referentes à *Web Semântica*. A falta de registro das possibilidades de ligações e de relacionamento de dados pode causar grandes problemas para o reuso dos dados. Nesse contexto, Moreau e Groth (2013) discutiram a possibilidade de inclusão das questões de proveniência no bolo de noiva, também conhecido como modelo em camadas ou pirâmide da *Web Semântica*, proposto por Berners-Lee, Hendler e Lassila (2001). Para os autores Moreau e Groth (2013), a proveniência perpassa as camadas de ontologia e lógica e está atrelada à camada *Proof* (Prova). Essa perspectiva destaca a importância da proveniência para dar confiabilidade às informações prestadas.

4. RESULTADOS E DISCUSSÕES

Diante do exposto, nos dados levantados pode-se observar semelhanças e diferenças entre as definições e modelos apresentados. Com essa dimensão teórica de conceitos sobre a proveniência foi proposta uma taxonomia que contempla o principal objetivo deste artigo, como também, a divulgação do trabalho de Arakaki (2019). Dessa forma, a taxonomia das características da proveniência, proposta na tese de Arakaki (2019), está baseada nos autores Simmhan, Plale e Gannon (2005) e Cruz, Campos e Mattoso (2009), e consiste em seis (6) categorias: Coleta, Representação, Estrutura, Armazenamento, Disseminação e acesso e Uso conforme apresentado na figura 5.

FIGURA 5
Taxonomia das características da proveniência



Fonte: adaptado de Arakaki (2019).

Cada categoria representa um aspecto de como podem ser extraídos dados de proveniência no contexto digital. Ao relacionar as características da proveniência destacadas neste artigo com o modelo PROV, proposto pelo W3C (Agente, recurso e processo), apresenta-se o quadro 1.

QUADRO 1
Características da proveniência e PROV

Características da proveniência			Modelo PROV			
			Agente	Recurso	Processo	
O que		Como	Quem	Onde	Quando	
Coleta	Rastreamento	Ocasional	x		x	
		Imediata	x		x	
	Nível	Atividade	x		x	
		Workflow	x		x	
		Sistema operacional	x		x	
	Mecanismo	Interno	x		x	
		Externo	x		x	
	Técnica	Inversão	x		x	
Anotação		x		x		
Representação	Granularidade	Fina		x		
		Grossa		x		
	Orientação	Orientado à dados		x		
		Orientado à processos		x		
	Fase	Composição		x		
		Execução		x		
Análise			x			
Estrutura	Semântica			x		
	Conteúdo			x		
Armazenamento	Escalabilidade	Centralizada		x		x
		Distribuída	Heterogêneo	x		x
			Homogêneo	x		x
	Persistência			x		
	Sobrecarga			x		
	Acoplamento	Sem acoplamento		x		x
		Acoplado		x		x
		Misto		x		x
	Arquivamento	Sequencial		x		x
Temporal		x		x		
Disseminação e acesso	Gráfico visual			x		
	Consulta			x		
	API			x		
	Navegação			x		
Uso	Qualidade			x		
	Trilha		x		x	
	Replicação		x		x	
	Atribuição		x			
	Informacional		x		x	

Fonte: elaborado pelos autores.

Observa-se que o quadro 1 incorpora, além das entidades Agente, Recurso e Processo do PROV, ‘o que’, ‘quem’, ‘como’, ‘quando’ e ‘onde’, frequentemente utilizados na literatura para identificação básica da proveniência. Isso possibilita a visualização da interferência dos possíveis metadados em cada característica da proveniência proposto por Arakaki (2019).

No quadro 1, todas as características de coleta podem relacionar metadados de agente, ou seja, quem coletou independentemente da forma. As formas de coleta foram atribuídas a metadados de processos e atividades, isto é, como foi coletada determinada informação.

Em relação às características de representação e estrutura, todas foram relacionadas ao recurso. Essas informações relacionam-se diretamente ao recurso, pois envolvem a estrutura de como elas estão organizadas acerca de um determinado recurso.

Às características de Armazenamento, Escalabilidade, Acoplamento e Arquivamento foram atribuídos agente e processo, enquanto às características Persistência e Sobrecarga foram atribuídos o recurso. No grupo de Disseminação e acesso, todas as características foram atribuídas como elementos do recurso. Já na característica Uso, qualidade foi relacionada ao recurso e atribuição foi relacionada a agente. As outras características foram atribuídas a processos e agente.

CONSIDERAÇÕES FINAIS

O artigo apresenta uma discussão do termo proveniência em diversos contextos, em especial no âmbito digital. Nesse sentido, a proveniência foi tratada na arquivologia e destacada sua importância para área, caracterizada como princípio para construção do fundo arquivístico, preservando a organicidade das informações.

Na museologia e nas artes, a proveniência é importante para caracterizar um item museológico ou obra de arte, e agrega o valor percebido ao objeto. No contexto digital, a preservação digital relaciona a proveniência como critério para confiabilidade e autenticidade de documentos. No campo da ciência da computação, os estudos estão voltados à identificação de fluxos de trabalho para coleta de dados científicos, que permite, entre outras coisas, a extração de informações para identificar a origem de algum dado.

Conforme discutida a relevância da proveniência em diferentes contextos, destaca-se a necessidade de estudos mais aprofundados para analisar se os instrumentos de representação da Ciência da Informação contemplam as características da proveniência no intuito de garantir a veracidade e a inalterabilidade das informações em ambientes digitais.

As discussões expostas neste artigo revelam possibilidades em identificar metadados para cada característica que a proveniência possa apresentar, além da viabilidade de ampliação para outros contextos, como na identificação de *fake news*, e contribuir para outras áreas a do jornalismo, por exemplo.

REFERÊNCIAS

- Arakaki, F. A. (2019). *Metadados administrativos e a proveniência dos dados: modelo baseado na família PROV*. Tese Doutorado em Ciência da Informação, Marília, Brasil. Recuperado de <https://repositorio.unesp.br/handle/11449/180490>
- Arakaki, F. A., Alves, R. C. V., e Santos, P. L. V. A. C. (2019). *Preservação digital e proveniência: interseções entre PREMIS e o PROV*. Em XX Encontro Nacional de Pesquisa em Ciência da Informação. ANCIB, Florianópolis, Brasil. Recuperado de <https://conferencias.ufsc.br/index.php/enancib/2019/paper/view/1356/647>
- Arquivo Nacional (Brasil). (2015). *Dicionário brasileiro de terminologia arquivística*. Rio de Janeiro: Arquivo Nacional. Recuperado de http://www.arquivonacional.gov.br/images/pdf/Dicion_Term_Arquiv.pdf
- Berners-Lee, T., Hendler, J. e Lassila, O. (2001). The semantic web. *Scientific American*, 284(5), 28–37.
- Bivar, B. et al. (2013). *Uma Comparação entre os Modelos de Proveniência OPM e PROV*. Em BRESCi 2013, Maceió, Brasil.
- Buneman, P., Khanna, S. e Tan, W. (2000). Data provenance: some basic issues. En S. Kapoor, S. Prasad (eds), *FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science*. FSTTCS 2000. Berlin: Springer.

- Conselho Internacional de Arquivos. (2007). *ISDF: Norma internacional para descrição de funções*. Paris: Conselho Internacional de Arquivos. Recuperado de https://www.ica.org/sites/default/files/CBPS_2007_Guidelines_ISDF_First-edition_PT.pdf
- Cruz, S. M. S., Campos, M. L. M. e Mattoso, M. (2009). Towards a Taxonomy of Provenance in Scientific Workflow Management Systems. *2009 Congress On Services - I, IEEE*. Recuperado de <http://dx.doi.org/10.1109/service-s-i.2009.18>
- Cunha, M. B. e Cavalcanti, C. R. O. (2008). *Dicionário de Biblioteconomia e Arquivologia*. Brasília: Briquet de Lemos.
- Dublin Core Metadata Initiative. (2020). *Provenance*. DCMI. Recuperado de <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/terms/provenance/>
- Duranti, L. et al. (2016). *InterPARES Trust Terminologia*. Recuperado em <http://arstweb.clayton.edu/interlex/pt/>
- Eckert, K. (2012). *Metadata Provenance in Europeana and the Semantic Web*. (Dissertação Mestrado em Informationswissenschaft). Recuperado de <https://edoc.hu-berlin.de/bitstream/handle/18452/2727/332.pdf?sequence=1&isAllowed=y>
- Eckert, K., Garijo, D. e Panzer, M. (2011). *Extending DCAM for Metadata Provenance*. Trabalho apresentado no International Conference on Dublin Core and Metadata Applications, DCMI, The Hague. Recuperado de <http://dcpapers.dublincore.org/pubs/article/view/3621>
- Furner, J. (2019). Definitions of “Metadata”: A Brief Survey of International Standards. *Journal of the Association for Information Science and Technology*, 71(6). doi: <https://10.1002/asi.24295>
- Haynes, D. (2018). *Metadata for Information Management and Retrieval: Understanding metadata and its use*. London: Facet Publishing.
- Lemieux e inPROVENANCE Group. (2016). Provenance: Past, Present and Future in Interdisciplinary and Multidisciplinary Perspective. In V. Lemieux (ed.). *Building Trust in Information* (pp. 3-45.. Bern: Springer International Publishing.
- Li, C. e Sugimoto, S. (2014). *Provenance description of metadata using PROV with PREMIS for long-term use of metadata*. Em International Conference on Dublin Core and Metadata Applications. The Hague: DCMI. Recuperado de <https://dcpapers.dublincore.org/pubs/article/view/3709>
- Li, C. e Sugimoto, S. (2017). Provenance description of metadata vocabularies for the long-term maintenance of metadata. *Journal of data and information Science*, 2(2), 41–55.
- Li, C. e Sugimoto, S. (2018). Provenance Description of Metadata Application Profiles for Long-Term Maintenance of Metadata Schemas. *Journal of Documentation*, 74(1), 36–61, Recuperado em <http://www.emeraldinsight.com/doi/10.1108/JD-03-2017-0042>
- Library of Congress e PREMIS Editorial Committee. (2015). *PREMIS Data Dictionary for Preservation Metadata*. Recuperado de <http://www.loc.gov/standards/premis/v3>
- Macedo, S. (2018). Ascensão de. ‘Proveniência’ na terminografia arquivística de língua portuguesa: prospeção e visualização de (dis)similaridades em termos e definições. *Revista ibero-americana de ciência da informação*, 11(2), 388–409. Recuperado de <http://periodicos.unb.br/ojs311/index.php/RICI/article/view/8334>
- Marins, A. L. A. (2008). *Modelos conceituais para proveniência*. (Dissertação Mestrado em Ciência da Informação), Rio de Janeiro, Brasil. Recuperado de <https://www.maxwell.vrac.puc-rio.br/colecao.php?strSecao=resultado&nrSeq=11880@1>
- Moreau, L. e Groth, P. (2013). Provenance: an introduction to PROV. *Synthesis lectures on the semantic web: theory and technology*, 3(4), 1-129. doi: <https://10.2200/s00528ed1v01y201308wbe007>
- Pearce-Moses, R. e Baty, L. A. (2005). *A glossary of archival and records terminology*. Chicago, IL: Society of American Archivists. Recuperado de <http://www.chismechick.com/wp-content/uploads/2017/08/SAA-Glossary-2005.pdf>
- Peterson, D. (1996). *Forms of Representation: An Interdisciplinary theme for Cognitive Science*. Wilshire: Intellect Books.

- Santos, P. L. V. A. C. e Vidotti, S. A. B. G. (2009). Perspectivismo e tecnologias de informação e comunicação: acréscimos à ciência da informação? *DataGramaZero: revista de ciência da informação*, 10(3).
- Simmhan, Y. L., Plale, B. e Gannon, D. (2005). A survey of data provenance in e-Science. *SIGMOD Record*, 34(3), 31-36.
- Tognoli, N. B. e Guimarães, J. A. C. (2019). Provenance as a knowledge organization principle. *Knowledge organization*, 46(7), 558-68. Recuperado de <http://www.isko.org/cyclo/provenance>
- Vidotti, S. A. B. G. et al. (2016). Coleta automática para povoamento de repositórios digitais: conversão de registros utilizando XSLT. *Tendências da pesquisa brasileira em ciência da informação*, 9(2).
- World Wide Web Consortium. (2011). *Library Linked Data Incubator Group Final Report*. W3C Incubator Group Report. Recuperado de <http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/>
- World Wide Web Consortium (2013a). *PROV Model Primer*. Recuperado de <https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/>
- World Wide Web Consortium (2013b). *PROV-DM: the prov data model*. Recuperado de <https://www.w3.org/TR/prov-dm/>
- World Wide Web Consortium (2017). *Data on the Web best practices*. Recuperado de <https://www.w3.org/TR/dwbp/>
- Zafar, F. et al. (2017). Trustworthy data: A survey, taxonomy and future trends of secure provenance schemes. *Journal of network and computer applications*, 94, 50-68. <http://dx.doi.org/10.1016/j.jnca.2017.06.003>