Representação da proveniência em repositórios de dados de pesquisa e a confiabilidade e a autenticidade da informação

Marieta Marks Löw[1]; Rafael Port da Rocha[2];

RESUMO

Apresenta pesquisa que tem como objetivo analisar a representação da proveniência em repositório de dados de pesquisa em relação aos aspectos de confiabilidade e autenticidade. Como objetivos específicos, a pesquisa analisa aspectos conceituais acerca da proveniência, da confiabilidade e da autenticidade da informação, bem como aspectos da organização da informação e representação da proveniência em repositório de dados de pesquisa. Trata-se de estudo exploratório, de abordagem qualitativa, utilizando procedimentos metodológicos de revisão de literatura. Apresenta os resultados iniciais e discute a importância da proveniência para os aspectos da evidência e contexto de produção e para autenticidade e confiabilidade do recurso informacional.

Palavras-chave: Proveniência; confiabilidade; autenticidade; dados de pesquisa; metadados.

INTRODUÇÃO

No âmbito dos estudos arquivísticos a proveniência é um aspecto essencial na representação de um conjunto de documentos, e além da representação do contexto de produção, a proveniência tem relação com a capacidade probatória do documento, ou seja, de o documento ser autêntico e confiável. A literatura arquivística sobre proveniência é ampla, e abrange diversos aspectos do termo proveniência. Por outro lado, o termo tem sido utilizado, para além dos estudos arquivísticos, em contextos tecnológicos, onde a questão da autenticidade e da confiabilidade da informação precisa ser assegurada ao longo do ciclo de vida da informação.

Repositórios de dados de pesquisa são uma tendência cada vez maior na pesquisa científica, unindo os avanços tecnológicos à percepção social da importância de transparência e compartilhamento de dados no fazer da ciência. Repositórios de dados de pesquisa são ambientes digitais onde os recursos informacionais levantados

[1] Mestre em Ciência da Informação pelo PPGCIN/UFRGS; marietalow@gmail.com
[2] Professor do Departamento de Ciência da Informação; FABICO/UFRGS; rafael.rocha@ufrgs.br

para uma determinada pesquisa são disponibilizados para eventual acesso e reuso pela comunidade científica. Estes dados são recolhidos, armazenados e preservados para acesso futuro, e devem manter as características originais de autenticidade, integridade e confiabilidade ao longo do tempo.

A pesquisa busca verificar a relação entre a representação da proveniência dos dados de pesquisa em repositórios e aspectos de autenticidade e confiabilidade dos dados, a partir do olhar arquivístico, observando-se também o contexto tecnológico dos repositórios para o acesso e o uso dos dados. Formulou-se então a seguinte **questão de pesquisa**: os metadados de proveniência em repositórios de dados de pesquisa contribuem para a confiabilidade e autenticidade dos dados de pesquisa?

Pensando a problemática da pesquisa e seus diferentes aspectos, estabeleceu-se um objetivo geral, que abarca o problema de forma ampla, e objetivos específicos, relacionados às etapas que estão sendo percorridas durante a pesquisa para atingir o objetivo geral.

Como **objetivo geral** da pesquisa, chegou-se à seguinte formulação: analisar a representação da proveniência em repositório de dados de pesquisa em relação aos aspectos de confiabilidade e autenticidade dos dados.

Os **objetivos específicos** foram estabelecidos de modo a contemplar as diferentes etapas a serem vencidas para o bom desenvolvimento da pesquisa, a partir dos aspectos conceituais de proveniência, confiabilidade e autenticidade, passando pelos aspectos da representação, acesso e uso dos dados de pesquisa.

a) Analisar aspectos conceituais acerca da proveniência, confiabilidade e autenticidade na produção dos dados de pesquisa;
b) Analisar a representação da proveniência em modelos de metadados e ontologias voltadas para dados de pesquisa em ciência;
c) Avaliar possibilidades de representação da proveniência com modelos de metadados e ontologias em repositórios de dados de pesquisa quanto aos aspectos de confiabilidade e autenticidade.

O desenvolvimento das tecnologias de informação e comunicação modificaram diferentes aspectos do fazer humano. A pesquisa científica atual está condicionada pelo novo paradigma tecnológico, tanto no aspecto da produção, como no acesso e

uso da informação científica e dos dados produzidos. Surgem preocupações acerca da origem, confiabilidade e autenticidade dos dados, com sua preservação de longo prazo, com a padronizações de procedimentos, entre outras. A custódia, curadoria e gestão, disponibilização e preservação de longo prazo dos dados gerados pela pesquisa científica são atividades técnicas bastante específicas, e que geralmente não é parte das habilidades desenvolvidas pelos cientistas para a produção científica. Entretanto, tais questões têm sido debatidas e pesquisadas entre profissionais da informação e nas áreas de Ciência da Informação e Arquivologia. Considerando tais aspectos, é relevante pesquisar a representação da proveniência em repositório de dados, e sua relação com a confiabilidade e autenticidade das informações.

METODOLOGIA

O estudo tem por objetivo investigar o uso de metadados de proveniência em repositórios de dados de pesquisa em relação aos aspectos de confiabilidade e autenticidade da informação. Trata-se de pesquisa exploratória, com abordagem qualitativa, com procedimentos metodológicos de revisão de literatura.

A pesquisa parte de uma revisão conceitual na literatura acerca da proveniência, da confiabilidade e da autenticidade da informação. Com esse objetivo busca-se identificar os principais entendimentos na literatura acerca dos termos e sua conceitualização nos diferentes campos de conhecimento da pesquisa. Como procedimento metodológico para esse objetivo realizar-se-á uma pesquisa bibliográfica, com levantamento dos conceitos à luz da teoria arquivística e no âmbito da ciência de dados. As fontes serão a literatura sobre teoria arquivística, ciência da informação e ciência de dados.

Um segundo objetivo específico é analisar a representação da proveniência em modelos de metadados e ontologias voltadas para dados de pesquisa. Com esse objetivo busca-se desenvolver uma análise de alto nível dos modelos de representação acerca dos aspectos relacionados à proveniência, autenticidade e confiabilidade. No contexto da pesquisa, este objetivo identifica viabilidades e benefícios do uso de metadados de proveniência em repositórios de dados de pesquisa. Para esse objetivo será realizada pesquisa bibliográfica com o

levantamento dos modelos de metadados e ontologias para dados de pesquisa. As fontes serão literatura relacionada à representação da proveniência, modelos de metadados e ontologias para ciência.

O último objetivo específico busca avaliar possibilidades de representação da proveniência com modelos de metadados e ontologias em repositórios de dados de pesquisa quanto aos aspectos de confiabilidade e autenticidade. Com isso, busca-se verificar as possibilidades de uso da representação da proveniência em dados de pesquisa e sua contribuição na melhoria do aspecto evidencial sobre a inferência de autenticidade e confiabilidade dos dados de pesquisa em repositório.

Como resultados gerais esperados, a pesquisa busca alcançar uma compreensão acerca das possibilidades de uso de metadados de proveniência em repositórios de dados de pesquisas e suas vantagens e desvantagens em relação aos aspectos de confiabilidade e autenticidade da informação. Outros resultados esperados são: a identificação dos principais entendimentos, na literatura, acerca dos termos e conceitos utilizados nos diferentes campos de conhecimento para expressar a proveniência em dados de pesquisa; compreender como os aspectos de autenticidade e confiabilidade estão representados nos diferentes modelos de metadados e ontologias voltados para dados de pesquisa; compreender as formas de uso dos modelos de metadados em situações reais de uso em repositórios de dados de pesquisa.

PRINCIPAIS DESCOBERTAS

A tecnologia alterou não somente o fazer científico, como a comunicação e o uso intensivo de dados. Há uma produção de grandes conjuntos de dados em pesquisa, e se torna possível, pela tecnologia, o acesso e reuso dos dados de pesquisa. Com isso, amplia-se a capacidade de colaboração, análises e descobertas a partir de um mesmo conjunto de dados. O tempo da ciência se modifica e os resultados entre a pesquisa e a publicação estão sendo reduzidos como nunca. A pandemia de COVID-19 mostrou a importância do acesso e compartilhamento de dados científicos na área da saúde, por exemplo.

O acesso aos dados de pesquisa tem sido uma tendência na produção científica atual. Seja por meio de repositórios de dados de pesquisa, ou de publicações ampliadas (*enhanced publication*). A compreensão sobre a importância do acesso aos dados de pesquisa ainda é desigual entre os diferentes campos de conhecimento: para algumas áreas, como a computação, a física e a genômica, a publicação de dados de pesquisa é prática usual no campo científico, com grandes avanços na pesquisa (JORGE, 2018). Por outro lado, em outras, ainda há questões de mercado e segredo industrial que dificultam o amplo acesso aos dados.

Essa abertura de dados é parte de um movimento mais amplo pela ciência aberta. E a discussão da abertura de dados é um tema importante, havendo um grande incentivo tanto no aspecto normativo, especialmente em pesquisas com financiamento público, como no desenvolvimento de padrões e tecnologias para gestão e acesso aos dados em ambientes *web* (JORGE, 2018).

Um marco importante para a abertura de dados de pesquisa é a Declaração de Berlim, de 2003. Para além dos resultados gerados pelas pesquisas científicas originais, a declaração considera que dados não processados, metadados, fontes originais, representações digitais de materiais pictóricos e gráficos, e materiais acadêmicos multimídia fazem parte do escopo do que deve ser de acesso livre ao conhecimento no âmbito das ciências e humanidades[3].

A Ciência da Informação e a Arquivologia são campos que estudam a gestão, acesso, uso e preservação de recursos informacionais. No campo da Arquivologia, o estudo da proveniência tem apoiado o desenvolvimento de soluções para garantia de autenticidade e preservação de documentos em ambientes digitais (DURANTI; EASTWOOD; MACNEIL, 2002). O aspecto da proveniência, como metadado acerca da origem, produção e custódia da informação, tem ganhado importância para o uso de dados de ciência, como forma de apoio na confiança e integridade do recurso informacional (HUANG, 2019).

A proveniência é um conceito bastante conhecido e utilizado nos estudos arquivísticos. Nesse campo do conhecimento o uso do termo está relacionado à entidade produtora de documentos e o processo de produção dos documentos de

[3] Uma versão oficial em português da Declaração está disponível em : https://openaccess.mpg.de/67693/BerlinDeclaration_pt.pdf. Aceso em 12/08/2021.

arquivo. O chamado Princípio da Proveniência é considerado a base teórica da Arquivologia. É a partir do estabelecimento desse princípio que os demais princípios teóricos se estabelecem (ROUSSEAU; COUTURE, 1998; SILVA *et al.*, 1999).

Ao longo do tempo, o princípio da proveniência foi distanciando-se do sentido de organização física dos documentos nas prateleiras dos arquivos. Seu uso passou a estar mais relacionado à uma ordenação intelectual dos documentos, relacionando o termo proveniência ao contexto de produção, no que ficou conhecido como a "redescoberta da proveniência" (COOK, 2013; TOGNOLI; GUIMARÃES, 2019).

Para Cook (2001), há também uma ressignificação do próprio conceito de proveniência. Ao invés de apenas relacionar um documento com a entidade de origem e uma determinada estrutura, o princípio agora relaciona o documento às funções e processos de negócio que o geraram, por meio de atividades em organizações que estão em constante evolução. Nesse mesmo sentido, Nesmith (1999), propõe uma definição mais ampla, possibilitando uma perspectiva multidisciplinar para o uso do conceito. Para o autor a proveniência consiste em processos técnicos e sociais de inscrição, transmissão, contextualização e interpretação de um documento, o que explica sua existência, características e história continua.

Para além da Arquivologia, o termo proveniência tem sido usado em áreas correlatas, sobretudo em contextos tecnológicos, por vezes importando o sentido dado originalmente pela literatura arquivística, ou com sentidos diferentes, atendendo as necessidades de outros campos de conhecimento. No âmbito da Organização do Conhecimento o termo consta como um verbete (*Provenance*) na Enciclopédia da ISKO (International Society for Knowledge Organization) (TOGNOLI; GUIMARÃES, 2018). O termo proveniência também é utilizado na área do Direito para descrever de onde algo surge e o caminho percorrido para chegar até onde está no momento. No direito a proveniência também está relacionada à evidência, informacional ou material, sua origem e o grau de confiança a ela atribuído, em processos e investigações (COHEN, 2013).

No contexto da *web* semântica, o termo proveniência se refere às fontes de informações (entidades e processos), envolvidas na produção ou entrega de artefatos (W3C INCUBATOR GROUP, 2010). A importância da proveniência, no contexto da *web*, está relacionada à possibilidade de confiança naquilo que é informado, a

possibilidade de integrar esse recurso a outros recursos e fontes de informação, bem como à poder referenciar a origem dos dados, em caso de reuso.

Assim, não somente a história pregressa do recurso informacional é importante, como também as relações entre o recurso e entidade que tornaram possível a criação desse recurso. A reprodutibilidade de um recurso informacional é uma característica importante para dados e informações, especialmente no contexto científico. Por outro lado, as características de autenticidade e a confiabilidade em recursos informacionais tornam-se ainda mais necessários em ambiente *web*, considerando o momento atual, com a chamada infodemia, o aumento de circulação de informações falsas, deturpadas ou descontextualizadas, especialmente em ambientes *web* como *sites*, redes sociais e aplicativos de mensagens (GROTH *et al.*, 2012).

Dois modelos foram identificados como possíveis para estudo, por terem relação com a proveniência e autenticidade. O modelo PROV-DM (PROV Data Model) é um modelo de dados conceitual que embasa uma série de documentos de especificações do W3C acerca da proveniência na *web*. Nele estão definidos os principais conceitos relacionados à proveniência. Outros documentos complementam a família, entre eles destacam-se para a pesquisa: o PROV-O, uma ontologia OWL2 que permite o mapeamento do modelo de dados PROV para o modelo RDF (Resource Description Framework); o PROV-DC, um mapeamento entre o PROV-O e termos do modelo Dublin Core; e o PROV-N, é uma recomendação de notação para proveniência voltada para a compreensão por humanos.

O segundo é o modelo Open Archival Information System (OAIS), desenvolvido para a preservação de longo prazo de objetos digitais, onde a proveniência é um dos aspectos a serem capturados para garantir autenticidade e confiabilidade do recurso ali registrado. O modelo prevê, como informações de preservação, os seguintes conjuntos de metadados: proveniência, contexto, referência, estabilidade e direitos de acesso

CONSIDERAÇÕES FINAIS

Este estudo, ainda em fase inicial busca contribuir com a área da Ciência da Informação, especialmente no âmbito da representação dos dados de pesquisa. Nos

últimos anos, com a criação de repositórios de dados de pesquisa e a maior automação da análise de dados, a presença de metadados de proveniência na descrição de recursos informacionais representa uma camada adicional de segurança e autenticidade para os recursos custodiados em repositórios de dados. A custódia, organização e preservação da informação são partes fundamentais do fazer arquivístico.

O conceito de proveniência é um dos principais na teoria arquivística, e está relacionada à produção e a autenticidade dos documentos. A gestão dos repositórios de dados de pesquisa, entretanto, não é considerada como parte das atribuições de arquivos. Por outro lado, a aplicação de conceitos arquivísticos em ambientes de custódia de dados pode apoiar a gestão, organização e preservação dos dados gerados durante uma pesquisa.

REFERÊNCIAS

COHEN, Fred. **Digital Forensic Evidence Examination**. 5th Ed.ed. Livermore: Fred Cohen & Associates, 2013.

COOK, Terry. Archival science and postmodernism: new formulations for old concepts. **Archival Science**, [*s. l.*], v. 1, n. 1, p. 3–24, 2001. Disponível em: https://doi.org/10.1007/BF02435636. Acesso em: 13 de abril de 2020.

COOK, Terry. Evidence, memory, identity, and community: Four shifting archival paradigms. **Archival Science**, [*s. l.*], v. 13, n. 2–3, p. 95–120, 2013. Disponível em: https://doi.org/10.1007/s10502-012-9180-7. Acesso em: 13 de abril de 2020.

DURANTI, Luciana; EASTWOOD, Terry; MACNEIL, Heather. **Preservation of the Integrity of Electronic Records**. Dordrecht: Springer Netherlands, 2002. *E-book*. Disponível em: https://doi.org/10.1007/978-94-015-9892-7_2. Acesso em: 23 de maio de 2019.

GROTH, Paul *et al.* Requirements for Provenance on the Web. **International Journal of Digital Curation**, [*s. l.*], v. 7, n. 1, p. 39–56, 2012.

HUANG, Jingwei. From Big Data to Knowledge: Issues of Provenance, Trust, and Scientific Computing Integrity. **Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018**, [*s. l.*], n. December, p. 2197–2205, 2019.

JORGE, Vanessa de Arruda. **Abertura e compartilhamento de dados para pesquisa nas situações de emergência em saúde pública: o caso do vírus Zika**. 2018. 263 f. [*s. l.*], 2018. Disponível em: http://ridi.ibict.br/bitstream/123456789/993/1/Tese_VanessaJorge.pdf. Acesso em: 14 de ago. de 2021.

NESMITH, Tom. Still Fuzzy, But More Accurate: Some Thoughts on the “Ghosts” of Archival Theory. **Archivaria**, [*s. l.*], v. 47, n. Feb, p. 136–150, 1999.

ROUSSEAU, Jean-Yves; COUTURE, Carol. **Os fundamentos da disciplina arquivística**. Lisboa: Publicações Dom Quixote, 1998.

SILVA, Armando Malheiro da *et al.* **Arquivística, Teoria e prática de uma ciência da informação.** Porto: Edições Afrontamento, 1999.

TOGNOLI, Natália Bolfarini; GUIMARÃES, José Augusto Chaves. **Provenance**. *In*: ENCYCLOPEDIA OF KNOWLEDGE ORGANIZATION. [*S. l.*]: International Society for Knowledge Organization, 2018. Disponível em: https://www.isko.org/cyclo/provenance. Acesso em 01 de ago. de 2019.

TOGNOLI, Natália Bolfarini; GUIMARÃES, José Augusto Chaves. Provenance as a knowledge organization principle. **Knowledge Organization**, [*s. l.*], v. 46, n. 7, p. 558–568, 2019.

W3C INCUBATOR GROUP. **Provenance XG Final Report**. [*S. l.*], 2010. Disponível em: https://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/. Acesso em: 01 de ago. de 2021.