Padrões de qualidade para dados e metadados endereçados a aplicações em ciência de dados

Quality standards for data and metadata addressed to data science applications

Daniela Lucas da Silva Lemos
Universidade Federal do Espírito Santo, Brasil.
E-mail: danielalucas@hotmail.com

Dalton Lopes Martins
Universidade de Brasília, Brasil.
E-mail: daltonmartins@unb.br

Danielle do Carmo
Universidade de Brasília, Brasil.
E-mail: docarmo.danielle@gmail.com

RESUMO

A presente pesquisa investiga como os modelos de organização e representação da informação e do conhecimento podem ser aplicados no campo da Ciência de Dados, com objetivo de evidenciar e discutir padrões de qualidade de dados que possam fornecer condições de produção de bases de dados curadas para aplicações. Quanto à metodologia, a pesquisa é classificada como qualitativa e de cunho exploratório e descritivo. Para coleta e análise dos dados, utilizou-se de pesquisa bibliográfica nos campos das Ciências de Dados e Ciência da Informação que sustentaram as discussões teóricas do estudo. O resultado permite compreender que o potencial de reúso dos dados para aplicações em Ciência de Dados depende de estratégias de organização e representação da informação e do conhecimento fundamentadas no âmbito teórico-metodológico da Ciência da Informação. Desde as etapas de criação de metadados até os processos de descrição, catalogação, classificação e indexação, a Ciência da Informação pode realizar contribuições significativas que impactem na qualidade de dados para uso e reúso em aplicações diversas.

Como citar: Lemos, D. L. S.; Martins, D. L.; & Carmo, D. (2022). Padrões de qualidade para dados e metadados endereçados a aplicações em ciência de dados. *Advanced Notes in Information Science, volume 2* (pp. 161-170). Tallinn, Estonia: ColNes Publishing. DOI: 10.47909/anis.978-9916-9760-3-6.116.

Palavras-chave: Bases de Dados em Rede; Metadados; Qualidade de Metadados; Ciência de Dados; Ciência da Informação.

ABSTRACT

The present research investigates how the models of organization and representation of information and knowledge can be applied in Data Science. We highlight and discuss how data quality standards can provide conditions for the production of curated databases for applications. As for the methodology, this is qualitative, exploratory, and descriptive research. Our theoretical discussions are supported by bibliographic research performed in the fields of data science and information science. The results contributed to understanding that the potential of data reuse for applications in Data Science depends on strategies of organization and representation of information and knowledge based on the theoretical-methodological scope of Information Science. From the metadata creation stages to the description, cataloguing, classification, and indexing processes, Information Science can make significant contributions that impact the data quality for use and reuse in diverse applications.

Keywords: *Networked Databases; Metadata; Metadata Quality; Data Science; Information Science.*

INTRODUÇÃO

Um importante fator que vem potencializando o surgimento e, por consequência, o crescimento de grandes bases de dados na rede está associado à prática da digitalização, a exemplo do que ocorre em instituições ligadas ao campo da cultura digital (Europeana Tech, 2020), conhecidas pelo acrônimo GLAM de "galerias, bibliotecas, arquivos e museus", em que se busca democratizar conhecimento científico e cultural na internet.

A partir desse contexto, tem-se o entendimento de que o crescimento de objetos digitais na rede torna-se impraticável do ponto de vista de sua preservação, acesso, recuperação e reúso, sem o apoio de estratégias de organização e representação da informação e do conhecimento alicerçadas nos campos da informação e da tecnologia (Lemos & Souza, 2020) que podem fundamentar boas práticas de curadoria para o estabelecimento de padrões de qualidade em dados e metadados oriundos de conjuntos de dados (datasets) disponíveis na internet.

Segundo o Digital Curation Center (DCC, 2021), um centro de especialidade em curadoria digital, curadoria trata-se de manutenção, preservação e agregação de valores aos dados garantindo uso e reúso em todo o seu ciclo de vida. Trata-se, portanto, de um conceito não novo, a exemplo de instituições de memória que sempre tiveram curadores para

zelar pelos dados de seus acervos, independente de ambiente digital. Porém, com o surgimento de elementos metodológicos contemporâneos para tratamento da informação em diversas mídias digitais, a exemplo do *Resource Description and Access* (RDA) (Joudrey; Taylor & Miller, 2005), do *Linked Open Data* (LOD) (Bizer; Heath & Berners-Lee, 2009), dos princípios FAIR, acrônimo para *Findability, Accessibility, Interoperability and Reuse* (Wilkinson *et al.,* 2016), e das linguagens para representações semânticas de características de objetos digitais em várias mídias na Web (Allemang; Hendler & Gandon, 2020), o escopo da curadoria se expandiu para diversos contextos de aplicação, exigindo novas habilidades dos profissionais da informação em lidar com organização e tratamento documental na Web.

No campo do patrimônio cultural, percebe-se a necessidade de desenvolver a curadoria digital na tentativa de incrementar e difundir acervos e coleções na Web visando serviços de reúso de dados culturais, incluindo agregações, espaço colaborativo, educação, pesquisa científica, aplicativos computacionais, entre outros (Freire; Sales & Sayão, 2020). Desse modo, percebe-se que instituições de memória vêm se adequando à crescente tendência de digitalizar seus acervos e de mantê-los salvaguardados em ambiente digital, especialmente com o uso de repositórios digitais (RDs), aumentando, por conseguinte, a produção de dados e metadados sobre os itens armazenados e difundidos na rede visando busca, recuperação e reúso pela sociedade.

Os metadados podem ser considerados a base para a curadoria digital, identificando-os como elemento central de representação e gestão ao longo de todo o ciclo de vida dos objetos em ambiente informacional. Para tal, a adoção de uma política de qualidade de dados (Siqueira *et al.,* 2021) tem sido explorada para dar condições de geração de infraestrutura informacional adequada a aplicações e pesquisas em Ciência de Dados, especialmente aprendizado de máquina (Europeana Tech, 2020).

A Ciência de Dados (CD) é um campo relativamente novo que agrega conhecimentos, métodos e técnicas interdisciplinares, oriundas especialmente da Ciência da Computação (CC), na perspectiva de que se possa melhorar a qualidade dos dados e, logo, extrair o seu valor, além de transformá-los em informação e conhecimento a partir de análise e mineração. No âmbito da presente pesquisa, torna-se válido destacar que o campo da CD dialoga de forma satisfatória com a Ciência da Informação (CI) pela natureza interdisciplinar de ambos os campos e por possuírem interesses comuns em relação a seus objetos epistemológicos e práticas metodológicas (Virkus & Garoufallou, 2020), incluindo produção, organização, gestão, disseminação, acesso, recuperação, uso e reúso (seja por pessoas ou máquinas) de dados, metadados, informação, conhecimento,

documento, dentre outros elementos informacionais, em variados contextos e conjunturas. Logo, o campo das CD abre espaços investigativos comuns aos trabalhadores e cientistas da informação interessados no emprego de tecnologias digitais para o desenvolvimento e gestão de infraestruturas informacionais que viabilizem a produção de conhecimento em rede, interesse central para os pesquisadores em CD e CI.

A partir desse ensejo, a questão de pesquisa que se forma estaria em responder: ?como os modelos de organização e representação da informação e do conhecimento podem ser aplicados na constituição de bases de dados curadas de modo a potencializar possíveis aplicações de CD? Acredita-se que os aportes teórico- metodológicos da CI podem trazer ganhos significativos ao campo da CD ao organizar e representar de forma consistente os elementos pertencentes às bases de dados disponíveis na rede, com as quais os dados serão manipulados e explorados para propósitos diversos. Em adição, acredita-se no potencial das ferramentas praticadas pelos cientistas de dados na projeção de fontes de informação adequadas para a realização de experimentos e aplicações de interesse tanto a CD quanto a outras frentes de pesquisa.

OBJETIVO

O objetivo do presente artigo é evidenciar e discutir a partir da literatura padrões de qualidade de dados oriundos da CI que possam fornecer condições de produção de bases de dados curadas para aplicações em CD. A partir dessa discussão, almeja-se visualizar oportunidades de ampliação da interdisciplinaridade dos profissionais da informação frente a novas demandas sociais.

PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa foi classificada como sendo de natureza teórica, qualitativa, e de cunho exploratório e descritivo, envolvendo literatura científica atual e pertinente ao propósito ora delineado. Como procedimento técnico para coleta e análise dos dados, utilizou-se de pesquisa bibliográfica no intuito de fundamentar conceitos e fornecer sustentabilidade teórica ao estudo, e para revisão na literatura, de modo a visualizar o panorama atual acerca do problema de pesquisa. Para tal, procedeu-se a um levantamento nos campos das CD e CI que sustentaram as discussões teóricas do estudo. As bases de dados consultadas para a pesquisa bibliográfica foram: Scielo, Library and Information Science Abstracts (LISA), Web of Science, Scopus, além de complementos de busca documental no Google Scholar.

RESULTADOS

Virkus e Garaoufallou (2020) apontam que os profissionais da informação teriam como uma de suas maiores responsabilidades no campo da CD, a criação e a modelagem de metadados. Nessa mesma linha de argumentação, Wang (2018) considera que é a própria relação entre os dados e a informação que deve ser observada para compreender as relações entre a CD e a CI. Logo, ainda seguindo Wang (2018), o controle da qualidade dos dados pode levar a melhores informações, e esse é um dos papéis mais importantes que o profissional da informação poderia ter no campo da CD.

Metadados no âmbito da CI é um assunto que tem sido tratado há décadas na disciplina de Catalogação ou Representação Descritiva, sendo considerada importante atividade na padronização e descrição de recursos de informação, promovendo interpretação uniforme e universal, em qualquer idioma e em qualquer tipo de unidade de informação; além de ser a forma mais comumente empregada para agregar semântica a informações com o propósito de facilitar a busca de recursos de informação. As atividades envolvidas na catalogação são geralmente orientadas por regras ou códigos de catalogação, linguagens documentárias e padrões de metadados.

Regras de catalogação determinam como elaborar a descrição de um recurso de informação e os pontos de acesso, tornando-se práticas essenciais na padronização, na descrição e, portanto, na agregação semântica de recursos de informação, com o propósito de viabilizar a interoperabilidade, o compartilhamento de recursos, o intercâmbio contínuo e a reutilização de metadados (Joudrey; Taylor & Miller, 2015). Exemplos de regras de catalogação incluem: *Anglo-American Cataloging Rules (AACR); RDA; Cataloging Cultural Objects (CCO)*.

Linguagem documentária é destinada à descrição do conteúdo do documento e está relacionada à Catalogação de Assunto ou Representação Temática que vislumbra aspectos intelectuais e semânticos, portanto, subjetivos, como a compreensão do assunto do documento para fins de tradução para uma linguagem de classificação, indexação ou resumo (NISO, 2005).

A indexação, elemento preponderante às ações bem sucedidas de recuperação da informação, pode descrever o assunto dos documentos seguindo uma linguagem documentária em que se busca obter um vocabulário controlado de um assunto ou domínio específico. Tal artefato de representação pode auxiliar nos processos de análise e descrição de documentos, permitindo a criação padronizada de metadados ao nomear, de forma consistente, os pontos de acesso aos documentos e a informação

neles contida; e os usuários no momento da escolha dos filtros destinados à busca, expansão do vocabulário das consultas, bem como na navegação em sistemas de informação para a Web. Exemplos de vocabulários controlados incluem: esquemas de classificação, listas de cabeçalhos de assuntos, tesauros, taxonomias e ontologias.

Ontologias, frente de pesquisa (mas não exclusiva) em CC, podem ser usadas como vocabulários controlados, no entanto numa perspectiva de tratamento semântico, o que permite um usuário descrever e interligar recursos existentes por meio de qualificadores como conceitos, instâncias, propriedades, relações e restrições mantidas entre tais recursos (Lemos & Souza, 2020). O modelo é endereçado à anotação semântica de documentos, a que Shadbolt *et al.,* (2006) esclarecem ser uma abordagem subjacente aos conceitos preconizados pela Web Semântica (Berners-Lee; Hendler & Lassila, 2001) no que tange ao fornecimento de semântica formal à organização e à representação da informação por meio de conexões lógicas entre os termos, viabilizando o processamento pelas máquinas e inferências computacionais.

Já os padrões de metadados têm a capacidade de prover um vocabulário comum para descrever uma variedade de estruturas de dados capazes de satisfazer a várias comunidades, e, geralmente, são estruturados seguindo modelos para tratamento dos dados, o que redunda em normalização, qualidade e intercâmbio de suas descrições. Exemplos de padrões de metadados incluem: Dublin Core; VRA Core; LIDO; MPEG-7.

Assim sendo, os artefatos de representação ora descritos podem ser considerados sínteses de um Sistema de Organização do Conhecimento (SOC) que busca, a priori, a organização de recursos de informação nos aspectos de: i) associação, gerando relacionamentos; ii) representação, gerando pontos de acesso e índices em processos de catalogação e indexação; iii) classificação, promovendo colocação e ordenação para os documentos; e iv) categorização, gerando esquemas de categorias.

Nesse sentido, tem-se que o emprego de SOCs para viabilizar a recuperação de informação nos mais diversos ambientes justifica-se na medida em que buscam construir bons esquemas de dados, padronizar a entrada de dados nas bases de dados, facilitar a estratégia de busca e, consequentemente, melhorar a interação do usuário com o sistema de recuperação de informação. Logo, vislumbra-se que o uso de SOCs em processos de organização da informação é uma boa prática em termos de curadoria digital.

Na CD, o uso de SOCs também traz benefícios diretos. Há diversos algoritmos e estratégias de aprendizagem de máquina supervisionada que demandam dados anotados para treinamento e teste de eficiência de seus resultados. A oferta de bases de dados de alta qualidade com dados

classificados por meio de vocabulários controlados ou mesmo ontologias torna-se um atrativo fundamental para o desenvolvimento de aplicações e pesquisas em CD.

As aplicações também podem utilizar metadados de boa qualidade oriundos de processos que refletem o uso de vocabulários controlados, regras de catalogação e padrões de metadados bem definidos, ampliando o potencial de uso. Nesse sentido, as aplicações podem se dedicar menos às etapas de pré-processamento de dados, em geral voltadas para melhorar a qualidade dos dados coletados, e mais às etapas de análise dos dados, modelagem algorítmica e identificação de padrões.

O profissional da informação, nesse contexto, além de conhecer seus instrumentos de trabalho tradicionais, necessita dominar outros, tais como vocabulários controlados, ontologias, linguagens de representação de recursos de informação para a Web, e também estar apto a lidar com as características multimídia dos novos documentos. Adicionalmente, esse profissional tem como papel preponderante o desenvolvimento de terminologias específicas para a estruturação de bases de dados com qualidade, seguindo modelos para tratamento documental visando normalização, qualidade e intercâmbio de dados para determinados cenários de aplicação de CD, incluindo fontes de informação médica, cultural, multimídia, artística, histórica, turística, educacional, e associadas a mídias sociais a exemplo do Wikidata, do Wikimedia Commons e do Wikipédia (Mora-Cantallops; Sánchez-Alonso & Garcíabarriocanal, 2019).

Tal perspectiva nos leva a compreender também um caminho inverso associado às contribuições de CD com a área de CI numa perspectiva de sua relação com os metadados, logo, possibilitando visualizar oportunidades interdisciplinares para os profissionais da informação em serviços de informações inovadores e de valor à sociedade. Podem-se citar técnicas de pré-processamento que visam melhorar a qualidade dos dados descritivos e temáticos nas bases de dados, sobretudo quando catalogados manualmente, incluindo normalização, limpeza, inclusão de valores ausentes, entre outros tratamentos (Virkus & Garaoufallou, 2020). Também nesse sentido, a CD contribui com a possibilidade da geração automática e semiautomática de metadados a partir de aplicações de processamento de linguagem natural. Tem-se também aplicações de CD que podem se valer de metadados de boa qualidade para aplicações de aprendizagem de máquina não-supervisionadas e supervisionadas fazendo bom reúso dos dados para clusterização, visualizações, inferências estatísticas, teste de hipóteses, reconhecimento de padrões em texto e multimídia, entre outros (Greenberg, 2017; Europeana Tech, 2020).

Finalmente, Greenberg (2017) reforça a compreensão do papel dos metadados na CD afirmando que se o campo não aprofundar a pesquisa

sobre os metadados em seu repertório conceitual e operacional, o campo enfrentará problemas em seu progresso. Para o autor, os metadados são indispensáveis no desenvolvimento de aplicações de CD, pois oferecem contexto para a interpretação uma vez que descrevem os dados.

CONSIDERAÇÕES FINAIS

Como se observa, a CI parece adquirir um papel fundamental na qualidade dos dados que são usados pelo campo da CD. Desde as etapas de criação de metadados até os processos de descrição, catalogação, classificação e indexação, a CI pode realizar contribuições significativas que impactem na qualidade dos dados e, portanto, na qualidade das pesquisas e aplicações desenvolvidas pela CD. Essa é uma visão fundamental para a presente pesquisa quando buscou responder *como os modelos de organização e representação da informação e do conhecimento podem ser aplicados na constituição de bases de dados curadas de modo a potencializar possíveis aplicações de CD?* Assim, o que se pode considerar como padrões de qualidade de dados e metadados em processos de curadoria em bases de dados são os propostos por esses artefatos de representação: metadados e seus padrões, vocabulários controlados e regras de catalogação.

A partir dos resultados evidenciados e discutidos no presente artigo, os objetivos de pesquisa foram atingidos, logo a questão respondida, quando se permitiu compreender que o potencial de reúso dos dados para aplicações em CD depende de estratégias de organização e representação da informação e do conhecimento fundamentadas no âmbito teórico-metodológico da CI.

CONFLITOS DE INTERESSE

Os autores declaram não haver conflito de interesses.

DECLARAÇÃO DE CONSENTIMENTO DE DADOS

Os dados gerados durante o desenvolvimento deste estudo foram incluídos no manuscrito.

REFERÊNCIAS

Allemang, D., Hendler, J., & Gandon, F. (2020). *Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL* (3a ed.). Association for Computing Machinery.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American, 284*(5), 34-43.

Bizer, C., Heath, T., & Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts* (pp. 205-227). IGI global. https://doi.org/10.4018/978-1-60960-593-3.ch008

DCC (Digital Curation Center). (2021). Retrieved from https://www.dcc.ac.uk/

Europeana Tech. (2020). *Interim Analysis of Europeana Tech AI in Relation to GLAMs Survey.* Retrieved from https://pro.europeana.eu/files/Europeana_Professional/Europeana_Network/Europeana_Network_Task_Forces/Final_reports/Final_Interim_Report_AI_in_GLAMs_TF.pdf

Freire, K. M. W., Sales, L. F., & Sayão, L. F. (2020). Curadoria digital no contexto artístico e cultural: possibilidades de reuso de dados de arte. *Encontros Bibli: Revista eletrônica De Biblioteconomia e Ciência da Informação,* (25), 01-21. https://doi.org/10.5007/1518-2924.2020.e74280

Greenberg, J. (2017). Big metadata, smart metadata, and metadata capital: Toward greater synergy between data science and metadata. *Journal of Data and Information Science, 2*(3), 19–36.

Joudrey, D. N., Taylor, A. G. and Miller, D. P. (2015). *Introduction to Cataloging and Classification* (11th Ed). Libraries Unlimited, Santa Barbara, CA.

Lemos, D. S. L., & Souza, R. R. (2020). Knowledge Organization Systems for the Representation of Multimedia Resources on the Web: A Comparative Analysis. *Knowledge Organization, 47*(4), 300-319. https://doi.org/10.5771/0943-7444-2020-4-300

Mora-Cantallops, M., Sánchez-Alonso, S., & García-Barriocanal, E. (2019). A systematic literature review on Wikidata. *Data Technologies and Applications, 53,* 250–268.

NISO (National Information Standards Organization). (2005). *Guidelines for the construction, format, and management of controlled vocabularies.* NISO Press, Baltimore.

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The semantic web revisited. *IEEE Intelligent Systems,* (21)3, 96–101. https://doi.org/10.1109/MIS.2006.62

Siqueira, J., Carmo, D. D., Martins, D. L., Silva Lemos, D. L. D., Medeiros, V. N., & Oliveira, L. F. R. D. (2021, March). Elements for Constructing a Data Quality Policy to Aggregate Digital Cultural Collections: Cases of the Digital Public Library of America and Europeana Foundation. In *International Conference on Data and Information in Online* (pp. 106-122). Springer, Cham. https://doi.org/10.1007/978-3-030-77417-2_8

Virkus, S., & Garoufallou, E. (2020). Data science and its relationship to library and information science: a content analysis. *Data Technologies and Applications, 54*(5), 643-663. https://doi.org/10.1108/DTA-07-2020-0167

Wang, L. (2018). Twinning data science with information science in schools of library and information science. *Journal of Documentation, 74,* 1243-1257. https://doi.org/10.1108/JD-02-2018-0036

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*(1), 1-9. https://doi.org/10.1038/sdata.2016.18