DOI http://dx.doi.org/10.1590/1981-5344/46064

MODELOS DE DIAGNÓSTICO DE QUALIDADE DE DADOS NO DOMÍNIO DO PATRIMÔNIO CULTURAL: UMA REVISÃO SISTEMÁTICA DE LITERATURA

Daniela Lucas da Silva Lemos
http://lattes.cnpq.br/9280443047358807 https://orcid.org/0000-0003-1565-7366
daniela.l.silva@ufes.br
Universidade Federal do Espírito Santo (UFES)
Vitória, ES, Brasil

Abeil Coelho Junior
http://lattes.cnpq.br/3539319991203481 https://orcid.org/0000-0003-1447-9537
abeilc@hotmail.com
Universidade Federal do Espírito Santo (UFES)
Vitória, ES, Brasil

Dalton Lopes Martins
http://lattes.cnpq.br/3774617443225038 https://orcid.org/0000-0002-6244-6791
daltonmartins@unb.br
Universidade de Brasília (UNB)
Brasília, DF, Brasil

RESUMO
Nos últimos anos, tem-se observado considerável adesão das instituições de patrimônio cultural ao processo de digitalização e disponibilização de seus dados de acervos na internet, proporcionando maior acessibilidade e democratização de conhecimento científico e cultural à sociedade. Diante deste fato, dados se tornam recursos importantes e valiosos para o século XXI, e considerações acerca da importância da qualidade para publicação de conjunto de dados na internet também surgiram nas últimas décadas em contextos diversos. Porém, apesar desses esforços, apenas investir na digitalização de objetos culturais não é suficiente, visto que questões sobre qualidade de dados frequentemente não são levantadas, considerando os diversos tipos de bancos de dados e sistemas de informação existentes. Esta pesquisa tem por objetivo identificar e analisar estudos sobre avaliação de qualidade de dados em acervos de patrimônio cultural, a partir de uma revisão sistemática da literatura no âmbito nacional e internacional. A partir da revisão de literatura realizada, ficou claro que há pouca evidência de um processo de garantia de qualidade de dados e de metadados testado que comprovasse sua eficácia em um ou mais repositórios digitais. Além disso, não há evidência de qualquer processo de avaliação de qualidade que fosse eficaz e transferível para outros contextos, ou seja, outros tipos de repositórios. Ressalta-se, ainda, a escassez de procedimentos que utilizam um modelo de catalogação de referência no campo do Patrimônio Cultural para fundamentar uma avaliação de qualidade de dados em bases de dados neste domínio.

Palavras-chave: Metadados. Acervo cultural. Revisão sistemática de literatura. Modelo de diagnóstico de qualidade de dados.

DATA QUALITY DIAGNOSIS MODELS IN THE DOMAIN OF CULTURAL HERITAGE: A LITERATURE REVIEW

ABSTRACT
In recent years, there has been a considerable trend among cultural heritage institutions to digitize and make their collection data available on the internet, providing greater accessibility and democratization of scientific and cultural knowledge to society. As a result, data has become an important and valuable resource for the 21st century, and considerations about the importance of data quality for publishing data sets on the internet have also emerged in various contexts over the last few decades. However, despite these efforts, investing solely in the digitalization of cultural objects is not enough, as data quality issues are often not raised, considering the various types of databases and information systems that exist. This research aims to identify and analyze studies on data quality evaluation in cultural heritage collections, based on a systematic review of national and international literature. Based on the literature review conducted, it became clear that there is little evidence of a tested data quality and metadata assurance process that proves its effectiveness in one or more digital repositories. Additionally, there is no evidence of any quality evaluation process that is effective and transferable to other contexts, i.e., other types of repositories. It should also be emphasized that there is a shortage of procedures that use a reference cataloging model in the field of Cultural Heritage to support a data quality evaluation in databases in this domain.

Keywords: Metadata. Cultural collection. Systematic literature review. Data quality diagnosis model.

DOI http://dx.doi.org/10.1590/1981-5344/46064

Recebido em: 03/05/2023.
Aceito em: 13/06/2023.

1 INTRODUÇÃO

Nos últimos anos, tem-se observado considerável adesão das instituições de patrimônio cultural ao processo de digitalização e disponibilização de seus dados de acervos na internet, proporcionando maior acessibilidade e democratização de conhecimento científico e cultural à sociedade. E com o advento das novas tecnologias da informação, esse aumento se tornou ainda mais considerável na produção e intercâmbio de registros em diversas áreas do conhecimento. Diante deste cenário, dados se tornam cada vez mais recursos importantes e valiosos para o século XXI, e considerações acerca da importância da qualidade para publicação de conjunto de dados (do inglês, *datasets*) abertos na internet também surgiram nas últimas décadas em contextos diversos (Bizer; Heath; Berners-Lee, 2009; Guizzardi, 2020; Macedo; Lemos, 2021; Siqueira *et al*., 2021; Wilkinson *et al*., 2016).

Apesar de essas iniciativas, investir somente na digitalização de objetos culturais não é suficiente (Martins *et al*., 2022), visto que questões relativas à qualidade de dados frequentemente não são levantadas, considerando os diversos tipos de bancos de dados e sistemas de informação ora envolvidos em processos de organização, modelagem e representação da informação. De acordo com Chapman (2005), os dados produzidos e compartilhados por esses sistemas são negligenciados quando se discute a respeito da qualidade de dados e, consequentemente, as informações obtidas a partir desses dados ficam sujeitas a desconfiança quanto à veracidade, ou podem ainda levar a tomada de decisão com base em dados enganosos.

O que se almeja para a obtenção da qualidade de dados em datasets, tanto no âmbito de pesquisas científicas quanto em práticas profissionais, é a criação e a modelagem apropriada de metadados pelo profissional da informação em processos envolvendo análise, contextualização, cálculo, síntese e descrição dos dados de modo a transformá-los em informação (Šlibar; Oreški; Ređep, 2021; Wang, 2018). Geralmente, tais processos são orientados por soluções inteligentes de organização e tratamento da informação advindas de campos científicos interdisciplinares como a Ciência da Informação (CI), a Ciência da Computação (CC) e a Ciência de Dados (CD) (Martins *et al*., 2022; Virkus; Garoufallou, 2020).

DOI http://dx.doi.org/10.1590/1981-5344/46064

Existem muitos aspectos sobre qualidade de dados, incluindo modelagem e gerenciamento, controle e garantia de qualidade, análise, armazenamento e acesso (Chapman, 2005). A abordagem usada para lidar com cada um desses aspectos dependerá da aplicação e do nível da qualidade de dados exigida para a utilização desses dados (USAID…, 2009). Assim, um dos principais desafios é determinar qual nível de qualidade de dados é aceitável para o fim almejado.

A qualidade não está necessariamente relacionada a dados isentos de erros, sendo, portanto, apenas uma das dimensões. Deve-se, logo, considerar outras dimensões, conforme exposto por Eckerson (2002), como: (i) precisão, pondera sobre a representação da realidade por meio dos dados ou se estes são provenientes de uma fonte confiável; (ii) integridade, define se os relacionamentos desses dados com a estrutura são mantidos de forma sólida; e (iii) consistência, define se os entendimentos dos elementos são definidos de forma consistente, que, por fim, corroborará com a medida da adequação dos dados a um propósito específico. No entanto, em geral, a qualidade dos dados pode ser vista como um subconjunto da qualidade da informação. Isso ocorre porque a qualidade dos dados se concentra na precisão e na integridade dos dados, enquanto a qualidade da informação também leva em consideração o significado dos dados e como eles são usados (Chapman, 2005). Complementando este conceito de qualidade de dados, pode-se afirmar que dados com qualidade são adequados para serem usados quando estiverem livres de defeitos, acessíveis, precisos, oportunos, completos, consistentes com outras fontes, relevantes, abrangentes, fornecer um nível adequado de detalhes, além de ser fácil de ler e interpretar (Ballou *et al*., 1998). E, para além disso, a qualidade é baseada no contexto, onde muitas das vezes os dados que podem ser considerados adequados para um cenário podem não ser apropriados para outro (Chapman, 2005).

Os custodiadores e proprietários de dados pertencentes ao campo da cultura digital, conhecidos pelo acrônimo GLAM (do inglês, *Galleries, Libraries, Archives, Museums*) de "galerias, bibliotecas, arquivos e museus", são os principais responsáveis pela qualidade de seus dados, com o uso de boas práticas de catalogação descritiva e de assunto (International Federation of Library Associations and Institutions, 2016). Com o uso de padrões de

documentação que orientam a estrutura de dados, valores de dados e conteúdo de dados, as instituições contam com um conjunto de ferramentas que pode levá-las a uma boa prática de catalogação, documentação consistente, e, por consequência maior acesso aos documentos pelo usuário final (Lemos; Coelho Júnior, 2023). No entanto, aqueles que fornecem os dados e aqueles que usam os dados também têm responsabilidades. Os coletores de dados e catalogadores têm o dever de rotular os dados corretamente e documentar metodologias de captação; os custodiadores têm o papel de fazer a manutenção e o controle de qualidade dos seus registros; e os usuários em reportar eventuais erros encontrados (Chapman, 2005). Nesse contexto, formula-se a questão de pesquisa que busca responder: quais práticas têm sido adotadas para a avaliação da qualidade de dados em acervos de instituições do patrimônio cultural em âmbito nacional e internacional?

Logo, o objetivo deste artigo é identificar e analisar estudos sobre a avaliação de qualidade de dados em acervos de patrimônio cultural, a partir de uma revisão sistemática de literatura. Tal interesse se sobressaiu na medida em que iniciativas de correção se mostram mais custosas que a digitalização de um acervo, sendo fundamental a prevenção e o controle da qualidade (Batini; Scannapieca, 2006; Chapman, 2005; Eckerson, 2002; English, 1999).

A hipótese é que há poucos estudos sobre a avaliação da qualidade de dados em acervos de patrimônio cultural. Nesse sentido, o resultado aqui apresentado a partir desta revisão pode se apresentar como importante instrumento de pesquisa ao possibilitar orientar e apoiar investigações mais aprofundadas em bases de dados culturais visando estratégias de avaliação diagnóstica para o controle da qualidade de dados e de metadados em acervos culturais, fazendo uso, sobretudo, de recursos de automação eficientes.

2 METODOLOGIA

Para a fundamentação teórica e metodológica da pesquisa, usou-se de levantamento bibliográfico em bases de dados de documentos científicos, sendo a principal delas o portal de periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) para recuperar artigos

DOI http://dx.doi.org/10.1590/1981-5344/46064

recentes, relevantes e com fator de impacto considerável aos propósitos de uma Revisão Sistemática de Literatura (RSL).

A fim de aperfeiçoar os resultados obtidos no portal de periódicos da Capes, foram consultadas diretamente cinco bases de dados de pesquisa acadêmica online, a saber: i) Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI) ii) Repositório Institucional da Universidade Estadual Paulista em Franca (UNESP); iii) Biblioteca Digital Brasileira de Teses e Dissertações (BDTD); iv) *Scientific Electronic Library Online* (SciELO) e v) Google Acadêmico.

A primeira base de dados científica foi escolhida por disseminar literatura de diversas áreas da CI. A segunda foi escolhida por dar acesso a documentos científicos, acadêmicos, artísticos, técnicos produzidos por pesquisadores e estudantes do Programa de Pós-graduação em Ciência da Informação da UNESP, Programa bem conceituado no Brasil. A terceira por dar acesso a pesquisas de mestrado e doutorado produzidas no Brasil. A quarta por ser repositório multidisciplinar para depósito, preservação e disseminação de dados de pesquisa e, por fim, mas não menos importante, o Google Acadêmico por permitir encontrar documentos que não são indexados nas bases mais institucionais e relacionados ao tema pesquisado.

Para a recuperação dos documentos nessas bases de dados científicas foi empregada a técnica de busca por palavras-chave que refletem o universo do assunto. As pesquisas consideraram em sua totalidade apenas artigos completos publicados em conferências e periódicos que em português (PT) incluíam os termos "Patrimônio Cultural", "Qualidade de Dados" e "Acervos Digitais" em qualquer parte de seu conteúdo; em inglês (EN) incluíam os termos "*Cultural Heritage*", "*Data Quality*" e "*Digital Collections*" em qualquer parte de seu conteúdo; Em espanhol (ES) incluíam os termos "*Patrimonio cultural*", "*Calidad de datos*" e "*Colecciones digitales*" em qualquer parte de seu conteúdo. O recorte temporal da presente pesquisa considerou a última década, isto é, se deu entre 2012 e 2023. Como resultado preliminar, 486 documentos foram recuperados (Quadro 1) nas bases de dados de pesquisa acadêmica online, que, após a retirada das duplicatas, foi reduzido para 438.

Quadro 1 – quantidade de resultados por serviço de busca

Fonte		Idioma da busca	Qtde. de resultados
Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI)		EN	*0*
		ES	*0*
		PT	*0*
Biblioteca Digital Brasileira de Teses e Dissertações (BDTD)		EN	*2*
		ES	*1*
		PT	*2*
Google Acadêmico		EN	*402*
		ES	*4*
		PT	*6*
Repositório Institucional da UNESP		EN	*1*
		ES	*0*
		PT	*4*
Scientific Electronic Library Online (Scielo)		EN	*0*
		ES	*0*
		PT	*0*
Portal de Periódicos da CAPES	*Academic Conferences International Limited*	EN, ES, PT	*1*
	American Library Association		*2*
	American Society for Information Science		*1*
	Association for Computers and the Humanities		*3*
	Association of Canadian Archivists		*1*
	Blackwell Publishing Ltd		*2*
	Brazilian Journal of Information Science		*1*
	Consejo Superior de Investigaciones Científicas		*1*
	Copernicus Publications		*2*
	El Profesional de la Informacion		*1*
	Emerald		*10*
	Encontros Bibli		*1*
	EPI SCP		*1*
	EUM – Edizioni Università di Macerata		*2*
	Information Today Inc		*1*
	International Association of School Librarianship		*1*
	Multimedia tools and applications		*1*
	MDPI		*9*
	Nomos		*1*
	Public Library of Science		*2*
	SAGE		*1*
	Springer		*4*
	Taylor & Francis		*1*
	The Global Biodiversity Information Facility		*1*
	Universidad Complutense de Madrid		*1*
	Universidad de Antioquía		*1*
	Universidade Estadual de Campinas		*1*
	Universidade do Vale do Itajai		*1*
	University of Florence		*2*
	University of Latvia		*1*
	Wiley		*6*

Fonte: elaborado pelos autores (2023).

Todos os itens foram anotados e verificados manualmente para determinar sua relevância; uma série de outros critérios foi especificada para selecionar os estudos apropriados para inclusão na revisão, expostos a seguir. Para serem incluídos, os artigos precisariam discorrer sobre a aplicação prática de avaliação de qualidade de dados em acervo de instituições culturais;

DOI http://dx.doi.org/10.1590/1981-5344/46064

estarem publicados em periódicos ou anais de conferências; estarem disponíveis para acesso; e possuir resumo. Como resultado dessa seleção, um total de 17 artigos que atenderam a estes critérios foi selecionado para a revisão (Tabela 1).

Tabela 1 – critérios de inclusão dos artigos

Fase	Critério	Estudos restantes
Busca	Quantidade total de resultados obtidos	486
Filtro de idioma e duplicatas	Artigos completos em inglês, espanhol ou português sem duplicatas	438
Filtro por tipo de estudo	Estar acessível e possuir resumo Não ser revisão de literatura, pesquisa de opinião ou capítulos de livro	379
Filtro de estudos no escopo de interesse	Relatar sobre avaliação de qualidade de dados	48
	Relatar sobre qualidade de dados em acervo culturais	22
	Fazer avaliação de qualidade de dados em acervo de instituições culturais	17

Fonte: elaborado pelos autores (2023).

3 RESULTADOS

Um exemplo de avaliação da qualidade de dados em acervos culturais pode ser observado no Instituto de Serviços de Museus e Bibliotecas que está entre as maiores agregações de patrimônio cultural dos Estados Unidos da América (EUA). Em suas coleções e conteúdos digitais (Fenlon; Efron; Organisciak, 2012) foi feita uma avaliação estatística em 92 mil documentos captados por meio do protocolo *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH) (Lagoze *et al*., 2002) de dados no padrão *Dublin Core*[1]. Foi gerado uma matriz comparativa entre elementos de metadados, entre provedores e entre elementos discricionais. A matriz gerada oferece uma visão abrangente das características linguísticas de uma coleção, sendo útil para a focalização de potenciais áreas problemáticas nos metadados ou áreas prontas para posterior aumento ou exploração por serviços de agregação. Esse tipo de visualização permite capturar características importantes de um conjunto massivo de dados, ao mesmo tempo em que permite uma exploração mais próxima, dinâmica e multidirecional.

[1] Para mais informações acesse: https://www.dublincore.org/specifications/dublin-core/.

A Biblioteca Digital da Universidade de Houston (Westbrook *et al.*, 2012) teve suas coleções auditadas por meio de amostragem com objetivo de avaliar a completude dos campos discricionais. Os valores de dados foram avaliados com base em um guia de metadados e boas práticas *Dublin Core* criado pela própria instituição. Para a avaliação de qualidade, foi selecionada uma amostra aleatória de 20 objetos por coleção, chegando ao fim o total de 600 objetos a serem avaliados. Nesta avaliação manual, foi verificada a existência de *links* quebrados, uso irregular de caixa alta, datas inválidas, ausência de cabeçalhos de assunto, dentre outros problemas. Após avaliação de qualidade desses itens, foi feito um plano de ação para correção e ajuste do fluxo de produção de metadados para corrigir e evitar mais ocorrências dos problemas encontrados.

Nas universidades de Pisa, Roma e de Turin na Itália (Bellini; Nesi, 2013), foi descrita uma metodologia de avaliação de qualidade de dados baseada na linguagem de programação *JavaScript*[2]. O *script* desenvolvido avalia as dimensões: completude, precisão e consistência do acervo cultural. A metodologia lista 19 regras de avaliação para os 11 elementos discricionais *Dublin Core*. Nas regras elencadas pelos autores, foi considerada a aplicabilidade por algoritmo computacional. Para cada elemento discricional, foi atribuído um peso de relevância para a composição do perfil de qualidade de dados dos acervos. Após a definição das regras, a coleta dos registros de metadados foi realizada por meio do protocolo OAI-PMH. Como resultado, o estudo colaborou para a definição de um perfil de qualidade de metadados para repositórios culturais, elencou regras de avaliação considerando a viabilidade de avalição por algoritmos, desenvolveu modelo de medição de qualidade de dados e integrou todos esses componentes em um serviço online, apoiando instituições no domínio cultural na avaliação de qualidade de seus repositórios.

As instituições *Organic Agriculture & Agroecology* e *Federação Natural Europe* (Palavitsinis, 2013), as quais recebem dados de diversos museus e galerias, aplicaram métodos de avaliação de qualidade de dados em seus repositórios, cujos métodos podem ser implantados ao longo do ciclo de vida de um repositório de forma a garantir que os metadados gerados a partir de

[2] Para mais informações acesse: https://developer.mozilla.org/pt-BR/docs/Web/JavaScript.

DOI http://dx.doi.org/10.1590/1981-5344/46064

provedores de conteúdo sejam de alta qualidade. O processo de avaliação atua principalmente na completude de elementos de metadados, incluindo também adequação, correção, objetividade, precisão e consistência dos registros de metadados. Dessa forma, a proposta é de avaliar a maturidade do repositório em seus estágios de desenvolvimento. A metodologia descreve um fluxo de trabalho que vai desde a construção do repositório com planejamento dos metadados utilizados, fases de desenvolvimento, atores e seus papéis. Conta ainda com atuação de especialistas em metadados, especialista de domínio, usuários internos que alimentam os bancos de dados e usuário externos que consomem os dados. Por fim, a metodologia foi implantada em um período de 26 meses, com a realização do total de oito experimentos, e pode-se observar que a qualidade dos registros de metadados nos repositórios melhorou em média 16% quanto a completude de metadados.

No Instituto Nacional Francês de Pesquisa Agrícola (Tsiflidou; Manouselis, 2013), o arquivo de publicações com temas relacionados à agroecologia teve uma avaliação da qualidade de dados realizada por meio de três ferramentas de avaliação. A primeira sendo o *Google Refine* (atual *Open Refine*), seguido da *MINT Statistics* e da *AK-Metadata Analytics Tools*. O acervo trabalhado foi obtido por meio do protocolo OAI-PMH em *eXtensible Markup Language* (XML). Foram avaliados mais de dois mil documentos. Os resultados demonstraram que as ferramentas que melhor se adequaram às necessidades de avaliação foram as *AK-Metadata Analytics Tools* e *MINT statistics*. A avaliação realizada teve cunho estatístico, com frequência de termos nos campos discricionais, quantidade de campos com valores preenchidos e quantidade de informações presentes nos campos. Como resultado, o estudo demonstrou características das ferramentas e técnicas que podem ser utilizadas para a avaliação da qualidade dos metadados em termos de análise estatística, sendo apresentada como a melhor solução a ferramenta *AK-Metadata Analytics Tool*.

O Repositório de Arquivos Online do Texas (Francisco-Revilla *et al*. 2014) foi utilizado como estudo de caso para a utilização de uma ferramenta chamada *Visualizing Archival Data system* (VADA) desenvolvida em *Javascript* usando bibliotecas como *HighChartsJS* e *JavaScript InfoVis ToolKit*. A

ferramenta faz a avaliação de inconsistências nas coleções em *Encoded Archival Description* (EAD) (Society Of American Archivists, 2022) em XML. Ao todo foram listados e avaliados 10 tipos de problemas, desde: avaliação das marcações XML, indicando a ausência de elementos obrigatórios e erros de marcação; avaliação do padrão de conteúdo de dados com relacionamento inconsistente nas marcações; dados duplicados, entre outros. Concluiu-se que, utilizando-se desta ferramenta, foi possível identificar de forma visual, quais coleções precisavam de intervenção.

O consórcio de bibliotecas acadêmicas no estado de Missouri, EUA (Moulaison, 2015), serviu de estudo de caso para avaliar a melhoria da qualidade de dados após a adoção oficial da norma *Resource Description and Access* (RDA). O estudo comparou a qualidade de dados entre seis meses e um ano após a adoção da norma, destacado as tendências no uso dos elementos discricionais. A avaliação realizada considerou a completude das descrições com o acompanhamento da quantidade de atributos nos campos de registros e frequência de valores nos campos avaliados. Desta forma, demonstrou-se por meio desse estudo o aumento da padronização dos dados pela adoção do RDA, todavia, ainda foi observada a falta de consistência na forma como os dados são preenchidos.

Na coleção de recursos digitais da biblioteca digital Europeana (Gaona García; Fermoso García; Sánchez Alonso, 2017), foi realizado estudo de cobertura do agregador. Os pesquisadores utilizaram *web crawler* para raspagem dos dados da Europeana. Com esse método de captação de dados, foram selecionados mais de 44 mil recursos digitais com total de 11 campos discricionais. Após a captação dos dados do portal, foi utilizado como referência o tesauro *Art & Architecture Thesaurus* – Getty (AAT) (Getty, 2017) da *Getty Research Institute* e o modelo de dados *Europeana Data Model* (EDM) (Charles; Clayphan; Isaac, 2017). Assim, foi avaliada a completude dos registros frente aos padrões de metadados utilizado pela EDM, além de avaliar a cobertura baseada nos termos apresentados na seção *styles and periods* do AAT. Foram avaliados separadamente os elementos recomendados, opcionais e obrigatórios do modelo de dados de referência. Os resultados do estudo mostraram que a maioria dos elementos de metadados classificados como obrigatórios tem uma alta qualidade de completude. Porém, foi encontrado

DOI http://dx.doi.org/10.1590/1981-5344/46064

deficiências nos metadados da Europeana quanto à precisão, com casos de redundância, ausência, ambiguidade e inconsistência de seus metadados.

O Portal de Dados Aberto do Governo da Suíça (Stephan; Beat; Angelina, 2018) tem como principal desafio, questões relacionadas à qualidade dos metadados, a falta de suporte de licenças padrão, interoperabilidade dos conjuntos de dados, suporte multilíngue e representação de dados geográficos. Para elencar as oportunidades de melhoria e dificuldades quanto ao uso dos dados do portal, foram realizadas entrevistas com especialistas e representantes dos principais grupos das partes interessadas a utilizarem os dados. Como resultado do estudo, a maioria dos entrevistados compreende a existência do portal e os padrões de metadados de forma muito positiva. Entretanto, destacaram áreas que requerem melhorias, levantando 13 (treze) ações de melhorias: adoção de vocabulários controlados e publicação como dados abertos ligados; criação de manual de convenções; foco na qualidade dos dados em vez da quantidade; dentre outros.

As instituições Biblioteca nacional da França, Biblioteca Nacional de Espanha, Biblioteca Nacional Britânica e Biblioteca Virtual Miguel de Cervantes tiveram seus dados utilizados em estudo de caso (Romero, 2019). As coleções dessas instituições são disponibilizadas em *Resource Description Framework* (RDF) com preceitos *Linked Open Data* (LOD), e, a partir disso, 35 critérios foram avaliados abrangendo o total de 11 dimensões, a saber: Exatidão, confiabilidade, consistência, relevância, completude, tempestividade, facilidade de compreensão, interoperabilidade, acessibilidade, licença e conexão com outros recursos. Para cada critério foram elencadas regras de avaliação com uso de expressões regulares[3], que são padrões utilizados para selecionar combinações de caracteres em uma *string*; validadores de formatação de RDF, que fazem a avaliação de escrita do documento RDF; e consultas estruturadas em *Protocol and RDF Query Language* (SPARQL) para determinar a adequação dos acervos aos critérios elencados. Como resultado, foi criada uma tabela comparativa das instituições avaliadas e suas respectivas pontuações nas dimensões e regras elencadas, de forma a permitir a seleção de repositório que melhor atenda às necessidades específicas.

[3] Para mais informações acesse: https://pt.wikipedia.org/wiki/Expressão_regular.

A visualização de metadados por meio de análise estatística foi utilizada em Harper (2016) na Biblioteca Pública Digital da América (DPLA). A partir do processamento de texto, são avaliados quais campos e metadados são mais utilizados no acervo investigado. Com o estudo realizado, foi possível visualizar o uso e o padrão de preenchimento dos campos discricionais em diferentes tipos de recursos, permitindo que gerentes de coleções digitais, agregadores como DPLA e especialistas em metadados comparem rapidamente grupos de metadados em várias facetas. Foi utilizado o *Python* com a biblioteca *Pandas* para realizar a análise na coleção em questão. Discutiu-se sobre o amplo potencial de aprendizado de máquina e ciência de dados em Bibliotecas, instituições acadêmicas e patrimônio cultural. A partir da análise estatística foi realizada ainda a avaliação da capacidade dos metadados em serem utilizados em projeto de ciência de dados. Como resultado da avaliação, baseado no levantamento da frequência de palavras e de campos discricionais, foram observadas algumas possiblidades de otimizações de mecanismos de busca que podem ser aplicadas com objetivo de aumentar as referências do Google para a Biblioteca. Além de direcionar ajustes dos índices no mecanismo de pesquisa interno da DPLA.

A avaliação de acervos das bibliotecas, Biblioteca Virtual Miguel de Cervantes, Biblioteca Nacional da França, Biblioteca Nacional Britânica e Biblioteca Nacional de Espanha (Candela *et al*., 2021) tiveram a qualidade dos dados avaliados com o uso de *Shape Expressions* (*ShEx*). *ShEx* permite a validação de RDF através da declaração de restrições no modelo RDF, permitindo definir restrições de nó para determinar o conjunto de valores permitidos de um nó, incluindo suas cardinalidades e tipos de dados associados. Com isso, foram avaliadas 11 dimensões com total de 35 critérios. Para a avaliação desses critérios, foram utilizados os esquemas de dados baseados em cada classe do repositório LOD, contendo uma lista de itens descritos a seguir: i) a consulta SPARQL, bem como o *endpoint* SPARQL que reúne os itens a serem testados; ii) uma etiqueta descrevendo o esquema e os dados utilizados; e iii) o esquema *ShEx* usado para avaliar os dados. Dessa forma, os esquemas *ShEx* criados podem ser testados online e reutilizados por outras instituições como ponto de partida para avaliar seus repositórios de LOD. Como resultado, o estudo mostrou que *ShEx* pode ser útil para avaliar

DOI http://dx.doi.org/10.1590/1981-5344/46064

dados de LOD publicados por bibliotecas. Além disso, o estudo pontuou que o *ShEx* pode ser usado como documentação, pois fornece uma representação legível por humanos que ajuda bibliotecários e pesquisadores a entender o modelo de dados.

A Biblioteca Digital Italiana (Lorenzini; Rospocher; Tonelli, 2021) fez o uso de processamento de linguagem natural para classificar textos discricionais da Biblioteca a fim de indicar quais são de alta ou de baixa qualidade. O conjunto de dados utilizados possui cerca de quatro milhões de registros, incluindo imagens, conteúdos audiovisuais e recursos textuais com descrição apresentada pelo *dc:description* do *Dublin Core*. Esses registros podem ser acessados por meio do manipulador OAI-PMH ou via terminal com consulta SPARQL. A partir da classificação manual por especialistas que indicavam descrição boa ou ruim, foram utilizadas cerca de 100 mil descrições para treino do modelo de aprendizado supervisionado. Como diretrizes de catalogação do campo *description* foi utilizado um padrão fornecido pelo Instituto Central de Catálogo e Documentação, órgão ligado ao Ministério da Cultura Italiano. Após a classificação das descrições pelos especialistas, foram executados os scripts de classificação em *Python* para avaliação dos demais recursos. Os resultados do estudo mostram que o uso de aprendizado de máquina traz bons resultados na tarefa de classificar descrições com os rótulos de alta ou baixa qualidade. Apresenta ainda a quantidade de dados de treinamento necessários para a classificação automática ao invés do manual.

O Instituto Brasileiro de Museus (Ibram) (MARTINS; MARTINS, 2021) serviu de fonte de dados para apresentação de um modelo de requisitos para a avaliação da qualidade de dados. A classificação da qualidade recebeu uma escala com cinco níveis, sendo zero o nível mais baixo e quatro o nível com maior qualidade. Nesta métrica são avaliadas as dimensões: (i) Metadados, (ii) Regras de catalogação, (iii) Linguagem documentária e (iv) Mídias e licenças. Para o requisito metadados, foi considerado o alinhamento do metadado frente ao recomendado pelo Inventário Nacional dos Bens Culturais Musealizados (INBCM) (Brasil, 2021). Quanto ao requisito regras de catalogação, foram utilizadas regras e convenções internas nas instituições museais para a avaliação. A dimensão linguagem documentária foi considerada com grau de relevância a utilização do *Thesaurus* para Acervos

Museológicos e o Tesauro de Objetos do Patrimônio Cultural nos Museus Brasileiros. Por fim, a dimensão tipo de mídia e licença considerou como fator de relevância a disponibilidade da licença de uso sobre a mídia, seja ela imagem, vídeo, áudio, texto ou objeto 3D. Além do desenvolvimento do modelo de avaliação, os autores aplicaram o diagnóstico numa dada porção de dados de acervos digitais oriundos de três museus ligados ao Ibram. Os resultados do modelo apontaram que os três museus avaliados apresentaram níveis de requisitos de qualidade praticamente idênticos para os metadados e para as regras de catalogação, recebendo o nível um de qualidade. Para as linguagens documentárias receberam o nível dois. Para os tipos de mídia e licenças o nível dois foi o evidenciado, com exceção de um dos classificados em nível superior aos demais, recebendo o nível três. Apontam ainda que a realização de investimento em projetos de digitalização de objetos culturais por si só não é suficiente para a preservação e a recuperação eficiente da informação. Finalizam dizendo que é necessário o emprego de um modelo de governança que forneça normas de gestão de dados para as instituições museais disponibilizarem seus acervos digitais na *web*.

A DPLA (Phillips; Tarver, 2021) avaliou mais de 36 milhões de registros provenientes de 43 instituições distintas, utilizando análise estatística. O estudo foi focado no campo de assunto, e incluiu a contagem de registros presentes nos repositórios, a quantidade de itens com assuntos preenchidos e a quantidade de assuntos por objeto. Além disso, os pesquisadores realizaram uma padronização dos valores para avaliar a relação entre os registros por assunto. Embora tenha havido uma redução de 1% na quantidade de registros sem conexões após a padronização dos valores, esse efeito não foi considerado significativo. Concluiu-se que praticamente qualquer coleção digital pode se beneficiar na busca e recuperação da informação ao investir na adição ou ajuste de metadados que tratem da categoria assunto em seus registros.

Na *Data Foundry* na Biblioteca Nacional da Escócia (Candela, 2023) é apresentado uma estrutura para transformar conjuntos de dados de organizações GLAM em LOD. O trabalho também discute a avaliação da qualidade dos dados, incluindo o uso de esquemas *ShEx* para definir restrições de nó e a validação de URIs externos. Os três conjuntos de dados avaliados

DOI http://dx.doi.org/10.1590/1981-5344/46064

foram o *Moving Image Archive* (MIA), o *National Bibliography of Scotland* (NBS) a e *Bibliography of Scottish Literature in Translation* (BOSLIT), e estavam disponíveis sob a licença *Creative Commons Zero* 1.0 *Universal*. Os resultados da avaliação mostraram que a estrutura pode ser útil para outras organizações interessadas em publicar conjuntos de dados como LOD seguindo as melhores práticas.

Nas coleções museológicas sob gestão Ibram (Lemos; Coelho Júnior, 2023) são apresentados os resultados de uma avaliação diagnóstica semiautomática da qualidade dos dados nas bases de dados dos museus digitais. Foi utilizado o guia de referência *Cataloguing Cultural Objects* (CCO) como instrumento metodológico central ao levantamento de regras de catalogação eficientes ao processo de avaliação diagnóstica dos dados, e cuja formalização das mesmas se deu por expressões regulares (*regex*) implementadas na linguagem de programação *Python*. Nesse sentido, a exploração dos dados foi realizada em 22 coleções de museus representando mais de 17 mil itens. Os resultados indicaram a necessidade de um tratamento mais adequado de algumas dimensões dos dados, como características físicas, descrição, localização geográfica e informações cronológicas; por outro lado, a avaliação mostrou o uso adequado de taxonomias para a dimensão classificação, que pode representar entidades associadas à classificação de temas, assuntos ou contextos de uso. Por fim, a pesquisa recomenda que sejam incorporadas práticas de catalogação maduras de instrumentos de referência para melhorar a modelagem de metadados e padrões de documentos nas bases de dados dos museus sob gestão do Ibram.

De modo a organizar a análise e os resultados acerca das metodologias de avaliação de qualidade de dados elencadas no âmbito da pesquisa e responder como a avaliação da qualidade de dados vem sendo feita em instituições do patrimônio cultural, o Quadro 2 a seguir sintetiza as iniciativas de projetos voltados à avaliação da qualidade de dados em acervos culturais.

Quadro 2 – quadro sinóptico de iniciativas de avaliação de qualidade de dados em instituições no âmbito do patrimônio cultural

Autoria	Tipo de instituição	Escopo	Método de avaliação
Fenlon, *Efron, Organisciak*	Instituto de Serviços de Museus e Bibliotecas – EUA	Geração de matriz comparativa entre dados de diferentes provedores, permitindo visão geral de um grande conjunto de dados	Modelos estatísticos.

(2012)		destacando pontos de atenção.	
Westbrook *et al.* (2012)	Biblioteca Digital da Universidade de Houston – EUA	Avaliação manual de qualidade de dados baseado em guia interno de boas práticas criado pela instituição.	Avaliação manual.
Bellini, Nesi (2013)	As universidades de: Pisa, de Roma e de Turin – Itália	Baseado em regras elencadas, são avaliados os requisitos completude, precisão e consistência do acervo cultural.	*Script* em *JavaScript*.
Palavitsinis (2013)	Empresa de Agricultura Orgânica e Agroecologia; *Federação Natural Europe*	Proposta de metodologia de avaliação de qualidade de dados e acompanhamento, desde o planejamento dos metadados armazenados ao uso pelo usuário final.	Manual por especialistas.
Tsiflidou, Manouselis, (2013)	Instituto Nacional Francês de Pesquisa Agrícola – França	Realização de *benchmark* de ferramentas de avaliação de qualidade de dados por meio de frequência de palavras nos campos discricionais.	*AK-Metadata Analytics Tools*.
Francisco-Revilla *et al*. (2014)	Repositório de Arquivos do Texas Online – EUA	Avaliação de marcações XML, incluindo ausência de elementos obrigatórios, erros de marcação e consistência de conteúdo de dados.	Ferramenta própria não especificada.
Moulaison (2015)	Consórcio de biblioteca acadêmica no estado de Missouri – EUA	Avaliação em campos MARC com o comparativo do uso de registros de autoridade, a completude das descrições com o acompanhamento da quantidade de atributos nos campos de registros, frequência de valores nestes campos.	Avaliação manual de amostra.
Gaona Garcia, Fermoso García, Sánchez Alonso (2017)	Europeana	Avaliação em campos discricionais, dentre elementos recomendados, opcionais e obrigatórios de 44 mil objetos, utilizando como referência o tesauro AAT.	Análise estatística exploratória.
Stephan, Beat, Angelina (2018)	Portal de Dados do Governo Aberto da Suíça	Realização de entrevistas com especialistas e representantes dos principais grupos das partes interessadas a utilizarem os dados.	Entrevista com especialistas.
Romero (2019)	As bibliotecas: Nacional da França, Nacional da Espanha, Nacional Britânica e Virtual Miguel de Cervantes.	Avaliação em coleções disponíveis em RDF com preceitos LOD em 11 dimensões com total de 35 critérios, para verificação de adequação dos dados.	Expressões regulares, validadores sintáticos e consultas em SPARQL.
Harper (2016)	Biblioteca Pública Digital da América (DPLA)	Visualização do padrão de preenchimento dos campos discricionais em diferentes tipos de recursos, permitindo comparação de grupos de metadados.	*Script* Python
Candela *et al*. (2021)	As bibliotecas: Virtual Miguel de Cervantes, Nacional da França, Nacional Britânica e Nacional de	Avaliação da qualidade dos dados com o uso de *Shape Expressions* (ShEx), permitindo a validação de declaração de restrições no modelo RDF.	*Shape expressions* (ShEx)

DOI http://dx.doi.org/10.1590/1981-5344/46064

	Espanha		
Lorenzini, Rospocher; Tonelli (2021)	Biblioteca digital italiana	Classificação manual de amostra de registro, com objetivo de classificar a qualidade do campo *description* do *Dublin core*. Após classificação de amostra uma Inteligência Artificial é treinada para classificar o restante das descrições.	*Script Python*
Martins *et al*. (2021)	Instituto Brasileiro de Museus (Ibram)	Criação de requisitos para a avaliação da qualidade de dados fundamentados em dimensões associadas a organização e tratamento da informação.	Avaliação manual de amostra.
Phillips, Tarver (2021)	Biblioteca Pública Digital da América (DPLA)	Avaliação focada no campo assunto, com a contagem de registros presentes nos repositórios, quantidade de itens com assuntos preenchidos, quantidade de assuntos por objeto.	Avaliação estatística
Candela (2023)	Biblioteca Nacional da Escócia	Apresentada uma estrutura para transformar conjuntos de dados de organizações GLAM em *Linked Open Data* (LOD) com etapa de avaliação de qualidade de dados.	*Shape expressions* (ShEx)
Lemos, Coelho Júnior (2023)	Instituto Brasileiro de Museus (Ibram)	Avaliação de qualidade de dados em 22 coleções museológicas sob gestão do Ibram. Avaliação semiautomática com regras em expressões regulares (*regex*) baseadas no *Cataloguing Cultural Objects* (CCO). A avaliação é feita por *script Python*.	*Script Python* com expressões regulares (*regex*)

Fonte: elaborado pelos autores (2023).

4 DISCUSSÕES

A análise dos estudos recuperados confirmou a hipótese de que há poucas pesquisas sobre a avaliação de qualidade de dados em acervos de patrimônio cultural, sobretudo quando se considera o cenário específico do Brasil e da América Latina como um todo. As questões regionais relacionadas a qualidade de dados desta ampla região parecem ainda não terem encontrado espaço significativo na pauta de agenda de pesquisa e desenvolvimento acadêmico. Pode-se ressaltar, ainda que de forma meramente inferencial, que questões políticas e sociais de grande relevância na contemporaneidade, como a própria discussão a respeito da decolonização de bases de dados de acervos culturais possa se beneficiar de técnicas e práticas construídas em torno da pesquisa a respeito da análise e melhoria da qualidade de dados dos acervos.

Confirmou ainda que padrões de documentação atuais, que promovem qualidade de dados e, por consequência, recuperação da informação (BATINI; SCANNAPIECA, 2006; BACA *et al*., 2006; ENGLISH, 1999;

SIQUEIRA *et al.*, 2021) ainda não são considerados em estudos mais recentes. Alguns desses padrões documentais são descritos a seguir.

Para avaliar a conformidade dos dados em instituições do domínio cultural, é fundamental o alinhamento com padrões de referência, como o *Machine Readable Cataloging* (MARC)[4] e o *Anglo-American Cataloguing Rules* (AACR)[5], que são formatos padronizados para a codificação de dados bibliográficos em formato de máquina e fornecem informações sobre autor, título, editora, ano de publicação, entre outros. O *Resource Description and Access* (RDA)[6] é outro padrão de referência importante, que estabelece diretrizes para a descrição e acesso a recursos culturais e de conhecimento, incluindo livros, objetos de arte, manuscritos, entre outros. O *Dublin Core* é um padrão de metadados fundamental para a organização e o compartilhamento de informações sobre diferentes tipos de recursos digitais, incluindo imagens, textos e vídeos, na *web*. O padrão fornece uma estrutura simples e flexível para a descrição de recursos digitais, permitindo a interoperabilidade entre diferentes sistemas e a descoberta de recursos na *web*. Já o *Visual Resources Association Core Categories* (VRA Core)[7] é um conjunto de categorias de metadados para a descrição de recursos visuais, como imagens, vídeos e objetos 3D. O *Cataloging Cultural Objects* (CCO)[8] fornece diretrizes para a seleção, a organização e a formatação de dados usados para preencher registros de catálogos, com base em categorias principais no *Categories for the Description of Works of Art* (CDWA) e no VRA *Core*. O *Lightweight Information Describing Objects* (LIDO)[9] é um esquema de colheita XML destinado a fornecer metadados, para uso em uma variedade de serviços on-line, de bancos de dados de coleções on-line de organizações a portais de recursos agregados, além de expor, compartilhar e conectar dados na *web*. Finalmente, o *Encoded Archival Description* (EAD)[10] é um padrão para a codificação de guias de arquivos em XML, mantido pela Biblioteca do Congresso em parceria com a Sociedade Americana de

[4] Para mais informações acesse: https://www.loc.gov/marc/umb/.
[5] Para mais informações acesse: http://www.aacr2.org.
[6] Para mais informações acesse: https://www.loc.gov/aba/rda/.
[7] Para mais informações acesse: https://www.loc.gov/standards/vracore/.
[8] Para mais informações acesse: https://vraweb.org/resourcesx/cataloging-cultural-objects/.
[9] Para mais informações acesse: https://cidoc.mini.icom.museum/working-groups/lido/what-is-lido.
[10] Para mais informações acesse: https://www.loc.gov/ead/.

DOI http://dx.doi.org/10.1590/1981-5344/46064

Arquivistas. É importante destacar que existem outros padrões de referência além desses que também podem ser considerados para a avaliação de conformidade dos dados (Baca *et al*., 2006; Harpring, 2022; Lemos; Coelho-Júnior; Carmo, 2021). A falta de alinhamento com esses padrões pode causar sérios problemas nos processos descritos, já que eles não apresentam critérios claros para a apuração dos dados, reduzindo a confiabilidade e aplicabilidade dos métodos avaliativos propostos. Infere-se deste resultado que essa discussão também precisa ser feita junto aos gestores dos equipamentos culturais públicos e privados e, sobretudo, aos gestores e instituições responsáveis pela formulação de políticas de informação que incentivem e promovam a adoção destes padrões, bem como forneçam o contexto necessário para estimular a formação de profissionais que adotem as boas práticas de documentação oriundas dos padrões mais atuais citados acima.

Muitos estudos apresentam alguma deficiência no processo de avaliação, principalmente em relação ao volume de dados avaliados. Em alguns casos, a avaliação foi realizada de forma manual e por amostragem, o que pode resultar em uma análise incompleta e imprecisa, como, por exemplo, nos estudos de Martins e Martins (2021), Moulaison (2015), Palavitsinis (2013), Stephan, Beat e Angelina (2018) e Westbrook *et al.* (2012). Percebe-se aqui ainda a incipiente incorporação de técnicas automatizadas e semiautomatizadas oriundas de áreas como ciência de dados e mais especificamente da aprendizagem de máquina. Parece haver uma oportunidade de avanço significativa na produção de eficiência nos processos de análise e melhoria na questão da qualidade dos dados ao adotar técnicas oriundas destas áreas.

No entanto, considerando a grande quantidade de bases de dados legadas que precisam de avaliação prévia antes de serem migradas para o ambiente digital (Martins *et al*., 2022), torna-se inviável um processo manual de avaliação. Para superar essa limitação, é essencial utilizar um guia de referência ou padrão de dados. Esses guias fornecem um conjunto de diretrizes que ajudam a padronizar e sistematizar o processo de avaliação, garantindo a confiabilidade e aplicabilidade dos métodos avaliativos propostos.

Assim, podemos destacar os estudos de Candela, 2023), Candela *et al*. (2021) e de Lemos; Coelho Júnior (2023) com uso de procedimentos metodológicos automáticos ou semiautomáticos para o processamento da avaliação diagnóstica. Destaca-se ainda que há falta de iniciativas nacionais acerca da avaliação diagnóstica de qualidade de dados em acervos do patrimônio cultural. Com destaque para os trabalhos de Martins e Martins (2021) e de Lemos; Coelho Júnior (2023) no Ibram e que fazem o uso de padrões de referência para a avaliação diagnóstica. É interessante notar que o aumento da demanda pela publicação de acervos digitalizados na *web*, sobretudo durante o período mais agudo da pandemia do coronavírus, parece ter sido um fator catalisador da geração de pesquisas nesta área no campo museológico brasileiro.

O uso de padrões de dados é fundamental para avaliar grandes volumes de dados de maneira eficiente e confiável, principalmente no domínio cultural, pois a qualidade é baseada no contexto, em que muitas vezes os dados que podem ser considerados adequados para um cenário podem não ser apropriados para outro (Chapman, 2005). Assim, a adoção desses guias permite uma análise mais estruturada e sistemática, o que é essencial para garantir a qualidade dos dados e, consequentemente, a eficácia das análises realizadas.

Acrescenta-se ainda que a avaliação de qualidade de dados é um aspecto importante na disponibilização de dados de acervos culturais online, pois normaliza e padroniza as terminologias e consistência dos dados, sendo de grande valia nos processos de busca e recuperação da informação (Lancaster, 2004), além de ajudar no alcance da interoperabilidade semântica dos dados entre diferentes esquemas de metadados e aplicações disponíveis na *web* (Zeng, 2019).

Finalmente, a avaliação da qualidade de dados é um fator crítico para coleções culturais, já que a precisão dos resultados de buscas e recuperação da informação depende diretamente da qualidade dos dados catalogados. Por isso, é pertinente o desenvolvimento de estudos mais aprofundados com o intuito de estabelecer um arcabouço metodológico reprodutível, automatizado ou semiautomatizado (Wang, 2018), que possa ser utilizado como aliado na melhoria da qualidade e consequente recuperação da

DOI http://dx.doi.org/10.1590/1981-5344/46064

informação. Com uma avaliação mais precisa da qualidade dos dados, é possível economizar recursos e direcionar os esforços dos especialistas para decisões que exijam maior atenção. Resultados diagnósticos de qualidade de dados projetados de forma sistemática e baseados em padrões de referência, com regras claras para avaliação, permitiriam que as instituições invistam em técnicas de Ciência de Dados para melhorar a qualidade dos dados descritivos e temáticos nas bases de dados, sobretudo quando catalogados manualmente. Isso inclui normalização, limpeza, inclusão de valores ausentes e outros tratamentos, sendo bem úteis para aplicações de aprendizagem de máquina não-supervisionadas e supervisionadas (Martins *et al*., 2022).

5 CONSIDERAÇÕES FINAIS

À luz dos resultados foi observado o total de 17 trabalhos que entre 2012 e 2023 relataram acerca da avaliação da qualidade de dados em instituições no domínio da cultura. Este estudo apresenta as diferentes metodologias e processos utilizados na avaliação de qualidade de dados no domínio em questão. Os diferentes estudos recuperados foram avaliados e listadas as metodologias apresentadas, além dos resultados obtidos.

É importante ressaltar, mesmo que algumas importantes considerações apontadas na seção de resultados deste artigo levem ao entendimento de que a automação é um horizonte fundamental de desenvolvimento técnico e científico no tema, que esse processo de avaliação de qualidade de dados não poderia ser totalmente automatizado, sendo a intervenção humana imprescindível em determinadas etapas em que decisões e ajustes são necessários. Também é evidente o importante papel de especialistas em documentação como bibliotecários, museólogos e arquivistas no processo. Desde a definição dos padrões de catalogação, dos modelos de dados e dos vocabulários controlados ao desenvolvimento de políticas de informação e modelos de governança que permitam a instauração de processos que melhorem as práticas de documentação e forneçam horizontes de capacitação para os profissionais da área, os especialistas da área de documentação possuem um papel fundamental em todo processo.

A partir desse diagnóstico, observa-se que grande parte dos métodos de avaliação de qualidade de dados encontra-se disperso, exigindo maior

esforço e dificuldade ao selecionar um método para ser replicado em uma base de dados própria, reduzindo seu aproveitamento em objetos culturais de forma integrada. Incentiva-se, a partir dos resultados, que políticas de informação específicas possam ser criadas pelos órgãos ligadas as políticas culturais e as instituições gestoras de acervos de maneira a dar visibilidade ao tema da qualidade de dados na documentação de acervos culturais.

Ficou claro ainda que há pouca evidência de um processo de garantia de qualidade de metadados testado que comprovasse sua eficácia em um ou mais repositórios. Ressalta-se, ainda, a escassez de procedimentos que utilizam um modelo de catalogação de referência na área da cultura para fundamentar uma avaliação de qualidade de dados em bases de dados.

Espera-se que este tipo de pesquisa forneça o direcionamento e as iniciativas necessárias para qualificar os esforços de documentação e estabelecer melhores critérios para as instituições que desejam publicar seus dados em repositórios digitais para posterior coleta e participação em diferentes redes de informação.

REFERÊNCIAS

BACA, Murtha *et al*. **Cataloging cultural objects**: a guide to describing cultural works and their images. Chicago: American Library Association, 2006.

BALLOU, Donald *et al*. Modeling Information Manufacturing Systems to determine information product quality. **Management Science**, [*s.l.*], v. 44, n. 4, p. 462–484, 1998.

BATINI, Carlo; SCANNAPIECA, Monica. **Data quality**: concepts, methodologies and techniques. Berlin; New York: Springer, 2006.

BELLINI, Emanuele; NESI, Paolo. Metadata quality assessment tool for open access cultural heritage institutional repositories. *In*: NESI, Paolo; SANTUCCI, Raffaella (org.). **Information technologies for performing arts, media access, and entertainment**. Heidelberg: Springer Berlin Heidelberg, 2013. p. 90–103. (series Lecture Notes in Computer Science Berlin, v. 7990).

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked Data: the story so far. **International Journal on Semantic Web and Information Systems (IJSWIS)**, [*s.l.*], v. 5, n. 3, p. 1–22, 2009.

CANDELA, Gustavo. Towards a semantic approach in GLAM Labs: the case of the Data Foundry at the National Library of Scotland. **Journal of Information Science**, [*s.l.*], [*s.n.*], [*s.n.*], 2023. DOI https://doi.org/10.1177/01655515231174386.

DOI http://dx.doi.org/10.1590/1981-5344/46064

CANDELA, Gustavo *et al*. A Shape Expression approach for assessing the quality of Linked Open Data in libraries. **Semantic Web**, *[s.l.]*, [s.*n.*], [s.*n.*], p. 1–21, 2021.

CHAPMAN, Arthur D. **Principles of Data Quality**. Copenhagen: GBIF, 2005. Disponível em: https://www.gbif.org/document/80509. Acesso em: 28 dez. 2022.

CHARLES, Valentine; CLAYPHAN, Robina; ISAAC, Antoine. **Definition of the Europeana Data Model v5.2.8**. [*s.l.*]: Europeana, 2017. Disponível em: https://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.8_102017.pdf. Acesso em: 21 dez. 2022.

ECKERSON, Wayne W. **Data quality and the bottom line**: achieving business success through a commitment to high quality data. [*s.l.*]: The Data Warehousing Institute, 2002. (series TDWI Report Series). http://download.101com.com/pub/tdwi/Files/DQReport.pdf. Acesso em: 28 dez. 2022.

ENGLISH, Larry P. **Improving data warehouse and business information quality**: methods for reducing costs and increasing profits. New York: Wiley, 1999.

FENLON, Katrina; EFRON, Miles; ORGANISCIAK, Peter. Tooling the aggregator's workbench: metadata visualization through statistical text analysis. **Proceedings of the American Society for Information Science and Technology**, [*s.l.*], v. 49, n. 1, p. 1–10, 2012.

FRANCISCO-REVILLA, Luis *et al*. Encoded archival description: Data Quality and analysis. **Proceedings of the American Society for Information Science and Technology**, [*s.l.*], v. 51, n. 1, p. 1–10, 2014.

GAONA GARCÍA, Paulo Alonso; FERMOSO GARCÍA, Ana; SÁNCHEZ ALONSO, Salvador. Exploring the relevance of Europeana digital resources: preliminary ideas on Europeana metadata quality. **Revista Interamericana de Bibliotecología**, [*s.l.*], v. 40, n. 1, p. 59–69, 2017.

THE GETTY RESEARCH INSTITUTE [GETTY]. **Art & architecture thesaurus® online**. California: GETTY, 2017. Disponível em: https://www.getty.edu/research/tools/vocabularies/aat/. Acesso em: 1 ago. 2022.

GUIZZARDI, Giancarlo. Ontology, ontologies and the "I" of FAIR. **Data Intelligence**, [*s.l.*], v. 2, n. 1–2, p. 181–191, jan. 2020.

HARPER, Corey A. Metadata analytics, visualization, and optimization: experiments in statistical analysis of the Digital Public Library of America (DPLA). **The Code4Lib Journal**, [*s.l.*], n. 33, [*s.n.*], 2016. Disponível em: https://journal.code4lib.org/articles/11752?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+c4lj+%28The+Code4Lib+Journal%29. Acesso em: 3 jan. 2023.

HARPRING, Patricia. **Metadata standards crosswalks**. [*s.l.*]: The J. Paul Getty Trust, 2022. Disponível em: https://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html. Acesso em: 11 jan. 2023.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS [IFLA]. **Declaração dos Princípios Internacionais de Catalogação**. Haia: IFLA, 2016. Disponível em: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/icp/icp_2016-pt.pdf. Acesso em: 06 mar. 2023.

LAGOZE, Carl *et al*. **Open archives initiative:** protocol for metadata harvesting - v.2.0. [*s.l.*]: [*s.n.*], 2002. Disponível em: http://www.openarchives.org/OAI/openarchivesprotocol.html. Acesso em: 20 fev. 2023.

LANCASTER, Frederick Wilfrid. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos, 2004.

LEMOS, Daniela Lucas da Silva; COELHO JÚNIOR, Abeil. Qualidade de dados em acervos do patrimônio cultural: uma avaliação diagnóstica semiautomática nos objetos culturais sob gestão do Instituto Brasileiro de Museus. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 28, p. 1–22, 2023.

LEMOS, Daniela Lemos da Silva; COELHO-JÚNIOR, Abeil; CARMO, Daniela do. Ontologias para anotação semântica em mídias: uma construção colaborativa de redes de conhecimento do patrimônio cultural. **Fronteiras da Representação do Conhecimento**, Belo Horizonte, v. 1, n. 1, p. 94–125, 30 set. 2021.

LORENZINI, Matteo; ROSPOCHER, Marco; TONELLI, Sara. Automatically evaluating the quality of textual descriptions in cultural heritage records. **International Journal on Digital Libraries**, [*s.l.*], v. 22, n. 2, p. 217–231, 2021.

MACEDO, Dirceu Flávio; LEMOS, Daniela Lucas da Silva. Dados abertos governamentais: iniciativas e desafios na abertura de dados no Brasil e outras esferas internacionais. **AtoZ**: novas práticas em informação e conhecimento, Curitiba, v. 10, n. 2, p. 14-26, abr. 2021. ISSN 2237-826X. Disponível em: https://revistas.ufpr.br/atoz/article/view/77737. Acesso em: 20 dez. 2022.

MARTINS, Dalton Lopes *et al*. Information organization and representation in digital cultural heritage in Brazil: systematic mapping of information infrastructure in digital collections for data science applications. **Journal of the Association for Information Science and Technology**, [*s.l.*], [*s.n.*], [*s.n.*], p. asi.24650, 2022.

MARTINS, Dalton Lopes; MARTINS, Luciana Conrado. Desafios e aprendizados na implantação do Tainacan nos Museus do Instituto Brasileiro de Museus. **Revista Eletrônica Ventilando Acervos**, Florianópolis, v. [especial], n. 1, p. 91–107, 2021.

DOI http://dx.doi.org/10.1590/1981-5344/46064

BRASIL. Ministério da Cultura. Instituto Brasileiro de Museus. **Resolução Normativa n. 6, de 31 de agosto de 2021**. Estabelece os elementos de descrição das informações sobre o acervo museológico, bibliográfico e arquivístico que devem ser declarados no Inventário Nacional dos Bens Culturais Musealizados, em consonância com o Decreto nº 8.124, de 17 de outubro de 2013. Brasília, DF, Imprensa Nacional, 2021. Disponível em: https://www.in.gov.br/web/dou/-/resolucao-normativa-ibram-n-6-de-31-de-agosto-de-2021-342359740. Acesso em: 10 jan. 2023.

MOULAISON, Heather Lea. The expansion of the personal name authority record under resource description and access: current status and quality considerations. **IFLA Journal**, [*s.l.*], v. 41, n. 1, p. 13–24, 2015.

PALAVITSINIS, Nikos. **Metadata Quality Issues in Learning Repositories**. 2013. 294 f. Thesis (Doctoral) – Departamento de Ciencias de la Computación, Universidade de Alcalá, Espanha, 2013. Disponível em: https://core.ac.uk/download/pdf/58910780.pdf. Acesso em: 1 jan. 2023.

PHILLIPS, Mark Edward; TARVER, Hannah. Investigating the use of metadata record graphs to analyze subject headings in the digital public library of America. **The Electronic Library**, [*s.l.*], v. 39, n. 3, p. 450–468, 2021.

ROMERO, Gustavo Candela. **Publicación y enriquecimiento semántico de datos abiertos en bibliotecas digitales**. 2019. 227 f. Tesis (Doctoral en Informática) – Departamento de Lenguajes y Sistemas Informaticos, Universidad de Alicante, Espanha, 2019. Disponível em: https://rua.ua.es/dspace/handle/10045/97353. Acesso em: 1 jan. 2023.

SIQUEIRA, Joyce *et al*. Elements for the construction of a data quality policy for the aggregation of digital cultural collections: the cases of the Digital Public Library of America. Inc and the Europeana Foundation. *In*: ÁLVAREZ, Edgar Bisset. (eds) **Data and information in online environments**: second EAI International Conference – DIONE 2021. [*s.l.*]: Springer International Publishing, 2021.

ŠLIBAR, Barbara; OREŠKI, Dijana; REĐEP, Nina Begičević. Importance of the open data assessment: an insight into the (Meta) data quality dimensions. **SAGE Open**, v. 11, n. 2, p. 21582440211023178, 2021.

SOCIETY OF AMERICAN ARCHIVISTS [SAA]. **Guidelines for college and university archives**: core archival functions. [*s.l.*]: Society of American Archivists, 2022. Disponível em: https://www2.archivists.org/node/14804. Acesso em: 9 fev. 2023.

STEPHAN, Haller; BEAT, Estermann; DUNGGA-WINTERLEITNER, Angelina. **Study in View of the Further Development of DCAT-AP CH**. [*s.l.*]: Bern University of Applied Sciences, 2018.

TSIFLIDOU, Effie; MANOUSELIS, Nikos. Tools and Techniques for Assessing Metadata Quality. *In*: GAROUFALLOU, Emmanouel; GREENBERG, Jane (org.). Metadata and Semantics Research Conference. 7., 2013, Greece. **Proceedings...** Greece: Springer International Publishing, 2013. v. 390, p. 99–110. Disponível em: http://link.springer.com/10.1007/978-3-319-03437-9_11. Acesso em: 3 dez. 2022.

USAID, U. S. AGENCY FOR INTERNATIONAL DEVELOPMENT. TIPS 12: Data Quality Standards. **USAID**, [*s.l.*], v. 12, n. 2, 2009. Disponível em: https://www.fsnnetwork.org/sites/default/files/tips-dataqualitystandards.pdf. Acesso em: 9 jan. 2023.

VIRKUS, Sirje; GAROUFALLOU, Emmanouel. Data science and its relationship to library and information science: a content analysis. **Data Technologies and Applications**, [*s.l.*], v. 54, n. 5, p. 643-663, 2020.

WANG, Lin. Twinning data science with information science in schools of library and information science. **Journal of Documentation**, [*s.l.*], v. 74, n. 6, p. 1243-1257, 2018.

WESTBROOK, R. Niccole *et al*. Metadata clean sweep: a digital library audit project. **D-Lib Magazine**, [*s.l.*], v. 18, n. 5-6, 2012. Disponível em: http://www.dlib.org/dlib/may12/westbrook/05westbrook.html. Acesso em: 3 dez. 2022.

WILKINSON, Mark D. *et al*. The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, [*s.l.*], v. 3, n. 1, p. 160018, 2016.

ZENG, Marcia Lei. Interoperability. **Knowledge Organization**, [*s.l.*], v. 46, n. 2, p. 122-146, jan. 2019. Disponível em: https://www.isko.org/cyclo/interoperability. Acesso em: 27 dez. 2022.