Tendências da Pesquisa Brasileira em Ciência da Informação

AUTOMAÇÃO DE PROCESSOS DOCUMENTAIS PARA AVALIAÇÃO DIAGNÓSTICA DE DADOS EM INSTITUIÇÕES NO DOMÍNIO DA CULTURA[1]

AUTOMATION OF DOCUMENT PROCESSES FOR DIAGNOSTIC DATA EVALUATION IN CULTURAL DOMAIN INSTITUTIONS

Abeil Coelho Junior[2]
Daniela Lucas da Silva Lemos[3]

Resumo: Destaca como a automação de processos documentais, juntamente com ações humanas, pode facilitar a avaliação diagnóstica de qualidade de dados em instituições que possuem acervos culturais digitais. A pesquisa adota uma abordagem aplicada, combinando elementos qualitativos e quantitativos, para a implementação da ferramenta DataQ Culture, que faz uma avaliação semiautomática por meio da linguagem Python, utilizando expressões regulares da linguagem formal regex, a partir do guia de catalogação *Cataloging Cultural Objects*. A aplicação é testada com sucesso nas coleções do Instituto Brasileiro de Museus, processando mais de 17 mil itens. Os resultados demonstram a importância da avaliação da qualidade de dados para melhorar a indexação e a recuperação da informação em acervos culturais, tornando o patrimônio mais acessível ao público. A DataQ Culture permite que os usuários avaliem a qualidade dos dados de forma simples e interativa, gerando relatórios com métricas de adequação e indicando ações para aprimorar a qualidade dos dados. A adoção de práticas de catalogação baseadas em modelos de referência, como o *Cataloging Cultural Objects*, contribui para a padronização e agregação semântica dos recursos de informação. A ferramenta desenvolvida pode auxiliar profissionais da informação no acompanhamento e melhoria da qualidade dos dados em seus acervos, promovendo uma maior eficiência na análise e recuperação da informação.

Palavras-Chave: qualidade de dados; acervos culturais digitais; catalogação; catalogação de objetos culturais; DataQ Culture.

Abstract: *This paper highlights how the automation of document processes, combined with human actions, can facilitate the diagnostic evaluation of data quality in institutions that hold digital cultural collections. The research adopts an applied approach, combining qualitative and quantitative elements, to implement the DataQ Culture tool, which performs a semi-automatic*

[1] O texto foi submetido, avaliado, aprovado e apresentado no GT8 ENANCIB.

[2]Mestre em Ciência da Informação. Cientista de Dados na Viação Águia Branca. E-mail: abeilc@hotmail.com. ORCID: https://orcid.org/0000-0003-1447-9537.
[3]Doutora em Ciência da Informação. Docente na Universidade Federal do Espírito Santo. E-mail: daniela.l.silva@ufes.br. ORCID: https://orcid.org/0000-0003-1565-7366.

evaluation using the Python programming language, leveraging regular expressions from the formal regex language, based on the Cataloging Cultural Objects guide. The application was successfully tested on the collections of the Brazilian Institute of Museums, processing over 17,000 items. The results demonstrate the importance of data quality evaluation in improving the indexing and retrieval of information in cultural collections, making cultural heritage more accessible to the public. DataQ Culture allows users to assess data quality in a simple and interactive manner, generating reports with adequacy metrics and indicating actions to enhance data quality. The adoption of cataloging practices based on reference models, such as the Cataloging Cultural Objects, contributes to the standardization and semantic aggregation of information resources. The developed tool can assist information professionals in monitoring and improving data quality in their collections, promoting greater efficiency in information analysis and retrieval.

Keywords: *data quality; digital cultural collections; cataloguing; cataloging cultural objects; DataQ Culture.*

1 INTRODUÇÃO

Iniciativas de digitalização de acervos culturais e disponibilização de itens de coleções digitais na internet têm sido uma prática nos últimos anos (Martins *et al*., 2022). Entretanto, investir somente na digitalização de objetos culturais não é suficiente, visto que questões de qualidade de dados frequentemente não são levantadas, considerando os diversos tipos de bancos de dados e sistemas de informação envolvidos em processos de organização, modelagem e representação. Assim, o custo de digitalizar uma coleção em um banco de dados pode ser alto, mas é apenas uma fração do custo de verificar e corrigir os dados posteriormente. É melhor prevenir erros do que corrigi-los posteriormente (English, 1999, p. 282), o que é de longe a opção mais barata (Chapman, 2005).

Os custodiadores e proprietários de dados, como, por exemplo, galerias, bibliotecas, arquivos e museus - GLAMs, acrônimo em inglês - são os principais responsáveis pela qualidade de seus dados, com uma boa catalogação descritiva (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 (Gilliland, 2016), 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.

Entretanto, constata-se que padrões de documentação atuais, que promovem qualidade de dados e, por consequência, recuperação da informação mais eficiente, ainda não são considerados em estudos mais recentes (English, 1999; Batini; Scannapieca, 2006; Baca *et al*., 2006; Martins *et al*., 2022). Nesta direção, destaca-se o CCO[4] (*Cataloging Cultural Objects*), padrão de documentação que fornece diretrizes para a seleção, a organização e a formatação de dados usados para preencher registros de catálogos, com base em categorias genéricas que podem ser empregadas a qualquer conjunto de metadados (Baca *et al*., 2006), inclusive, com os elementos descritivos do experimento adotado na presente pesquisa.

No caso do Instituto Brasileiro de Museus (Ibram), estudo de caso da presente pesquisa, a qualidade de dados dos museus sob sua gestão pode ser mensurada por meio de suas bases de dados modeladas a partir do modelo de dados adotado internamente pela instituição, qual seja o modelo do Inventário Nacional de Bens Culturais Musealizados - INBCM (BRASIL, 2021). Assim, diante do contexto de uso do INBCM na arquitetura das bases de dados dos museus sob gestão do Ibram e da situação problemática associada à qualidade de dados em acervos culturais que aqui se apresenta, a presente pesquisa busca responder a seguinte questão: *como melhorar a qualidade de dados em acervos culturais?* Logo, o objetivo desta pesquisa é apresentar uma ferramenta de avaliação semiautomática de qualidade de dados que possibilite a otimização de resultados diagnósticos nos dados de acervos de instituições culturais, tendo o Ibram como objeto-experimental.

Acredita-se, portanto, que a implementação de uma avaliação semiautomática de qualidade de dados pode aprimorar a indexação, a busca e a navegação nos

4 Fonte: https://vraweb.org/resourcesx/cataloging-cultural-objects/. Acesso em: 12 jul. 2023.

sistemas de recuperação de informações (Lancaster, 2004) de instituições culturais, tornando o patrimônio cultural mais acessível ao público. Além disso, a aplicação desenvolvida permite que qualquer fonte de dados, independentemente do padrão de documentação adotado, possa ser alinhada e avaliada de acordo com as regras de catalogação do CCO. A padronização dos dados em coleções digitais pode facilitar comparações entre diferentes instituições, impulsionando pesquisas acadêmicas e científicas e proporcionando uma compreensão mais profunda da história e cultura do país.

2 QUALIDADE DE DADOS APLICADA NO DOMÍNIO DA CULTURA

A qualidade de dados pode ser vista como um subconjunto da qualidade da informação. Isso ocorre porque a qualidade de 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). Nesse sentido, a Ciência da Informação (CI) traz um arcabouço teórico-metodológico em seu percurso científico bem consolidado e respeitado na produção e na gestão de metadados (Zeng; Qin, 2016; Gilliland, 2016) visando dar significado e contexto aos recursos de informação por meio de instrumentos de organização. Como tal, os metadados ajudam na qualificação de bases de dados, tanto em termos de seu tratamento descritivo quanto temático (Svenonius, 2000; Gilliland, 2016; Hjørland, 2018; International Federation of Library Associations and Institutions, 2016), tornando as fontes de dados adequadas para a realização de experimentos e aplicações de interesse, conforme será elucidado adiante.

Em se tratando da qualidade de dados em documentação, campo no qual o domínio da cultura se faz presente, os padrões de documentação são usados como estratégia adotada pelas instituições para o tratamento documental buscando-se

estruturas padronizadas de descrição do recurso informacional endereçadas às bases de dados dos sistemas de recuperação da informação (SRIS) (Lancaster, 2004), promovendo, logo, mecanismos para coleta automatizada de conteúdos em diversas fontes, compartilhamento de registros e processamento da informação (Harpring, 2022).

Assim, museus, bibliotecas, arquivos e outros espaços culturais demandam bases de dados, consideradas produtos de informação cruciais para a sociedade quando adotados para realizar a mediação entre documentos e públicos, pois referenciam e divulgam o conhecimento a partir do uso qualificado da informação. Para tal, torna-se oportuno pensar a catalogação como processo-chave para a construção de bases de dados, pelo fato de possuir uma história rica e antiga, qual seja, desde os primórdios das formas de arquivamento em pilhas (papirus) e sequencial (tábuas de argila) em bibliotecas como Ebla e Alexandria (Barbosa, 1978; Mey; Silveira, 2009) até a contemporaneidade nos princípios e nos processos de catalogação (Baca *et al*., 2006; International Federation of Library Associations and Institutions, 2016), incluindo a automação e computação digital (ex.: tecnologias de bancos de dados), a rede como paradigma de ambiente informacional (ex.: advento da internet e da Web), a efetiva separação da informação de seu suporte (ex.: processo de digitalização), e a integração e agregação de serviços de informação em rede (ex.: plataforma digital Europeana; plataforma digital Braziliana museus). Em suma, há uma maturidade metodológica na catalogação que deve ser considerada na produção de bases de dados com qualidade em SRIs.

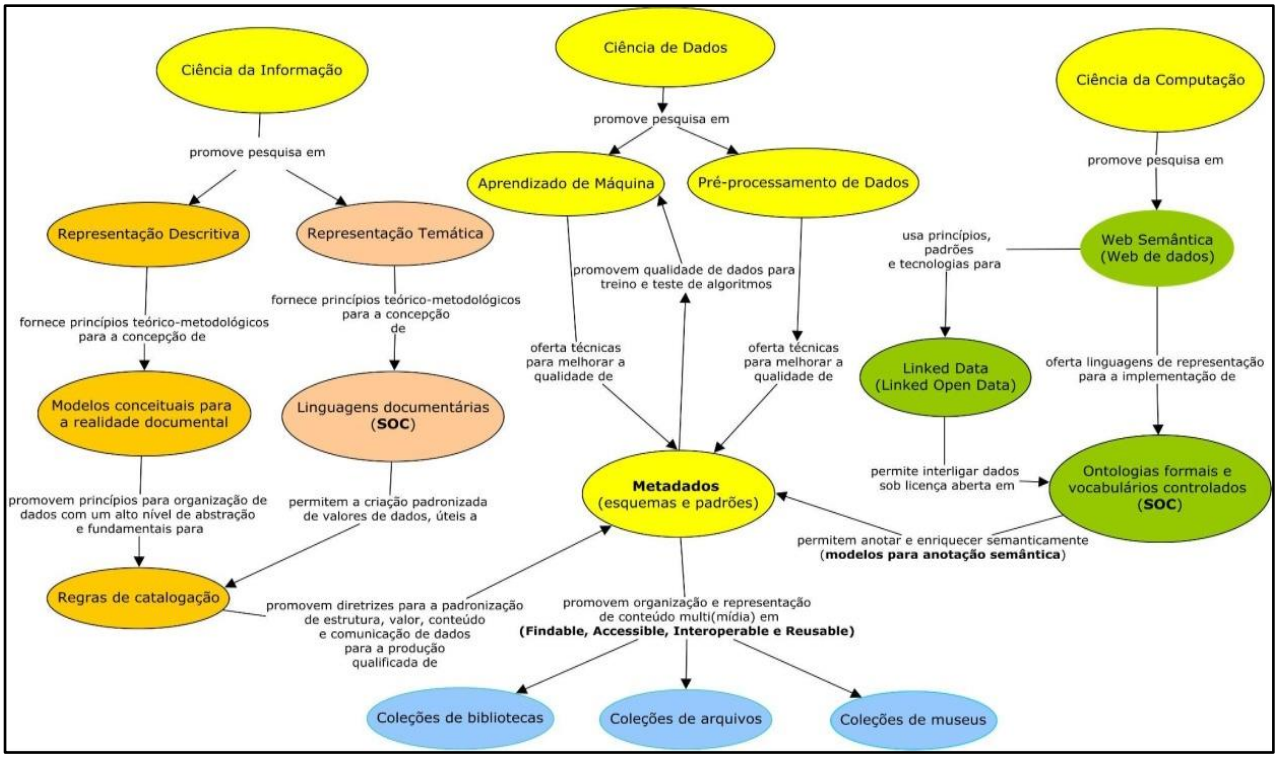
A catalogação, portanto, busca tratar da representação de um objeto informacional, consistindo no levantamento de suas características, de maneira a individualizá-lo, tornando-o distinto dos demais, e criando vínculos entre objetos que compartilham determinadas características (Joudrey; Taylor; Miller, 2015). Para

o cumprimento dessa tarefa, a catalogação precisa apresentar características como integridade, clareza, precisão, lógica e consistência, tendo as linguagens de representação documental (temática e descritiva) como principais agentes para o alcance destas características.

A Figura 1 ilustra essa conjugação na área de CI, cujas Representações Descritivas e Temáticas empregam princípios teórico-metodológicos em processos que se utilizam de instrumentos de organização e representação (modelo conceituais, linguagens documentárias, padrões de metadados, regras de catalogação) para a produção qualificada de bases de dados e, por consequência, metadados destinados a descrever conteúdo em variados tipos de mídia associados a coleções culturais disponíveis na rede, incluindo bibliotecas, arquivos e museus.

Regras de catalogação, padrão de documentação central na presente pesquisa, determinam como elaborar a descrição de um recurso de informação e seus pontos de acesso, tornando-se práticas essenciais na padronização, uniformidade, evitando divergências e duplicações (com recomendação de uso de modelos conceituais, vocabulários controlados e padrões de metadados) na descrição de recursos de informação, com o propósito de viabilizar interoperabilidade, compartilhamento de recursos, intercâmbio contínuo e reutilização de metadados (JOUDREY; TAYLOR; MILLER, 2015; BACA *et al*., 2006; ZENG, 2019). Exemplos de regras de catalogação incluem: *Anglo American Cataloging Rules* (AACR); *Cataloging Cultural Objects* (CCO); *Resource Description and Access* (RDA).

Figura 1 - Modelo teórico-conceitual para a qualidade de objetos culturais digitais



Fonte: Lemos; Martins; Souza (2023).

O CCO, objeto empírico usado no experimento, foi publicado pela American Library Association (ALA) em 2006, como resultado do consenso de profissionais das comunidades de museus, bibliotecas, galerias e arquivos que pesquisam a prática comum entre essas disciplinas (Baca *et al*., 2006, Harpring, 2022). É considerado um padrão de criação de conteúdo descritivo para recursos culturais, derivado do padrão semântico *Categories for the Description of Works of Art* (CDWA) e do padrão de metadados *Visual Resources Association Core Categories* (VRA Core) para a descrição de recursos visuais. O guia de catalogação CCO apresenta três partes principais, a saber: (i) introdução e instruções gerais; (ii) elementos de metadados; e (iii) autoridades.

O guia CCO traz recomendações e regras de catalogação descritas com clareza e bem organizadas em grupos de informação sintetizados em 9 (nove)

capítulos, cada um considerando um conjunto específico de elementos discricionais recomendados (Parte II).

Metadados, portanto, quando produzidos, a partir de regras de catalogação, e organizados e armazenados em bases de dados consideradas qualificadas, tornam-se vocabulários consistentes capazes de descrever, identificar, localizar e facilitar a recuperação, a interoperabilidade, o uso e o gerenciamento de uma variedade de fontes de informação digital disponíveis na rede. Para os propósitos da presente pesquisa, tanto a Representação Descritiva quanto a Representação Temática são consideradas na avaliação da qualidade dos dados das coleções ora envolvidas, sendo avaliado o emprego de vocabulário controlado tanto na descrição da Obra, em seus aspectos de suporte, quanto na descrição de seu conteúdo.

Nesta dinâmica (também apresentada na Figura 1) aparece o paradigma dados abertos ligados (LOD), uma frente de pesquisa da Ciência da Computação (CC), com grande atuação do World Wide Web Consortium (W3C), cuja fundamentação subjaz aos preceitos da Web semântica (Bizer; Heath; Berners-Lee, 2009), e tem como propósito anotar e identificar formalmente conceitos e relações entre conceitos em documentos (por meio de vocabulários controlados e ontologias formais), tornando-os mais inteligentes e facilitando, portanto, o processo de interpretação dos dados pelos sistemas de recuperação de informação.

Outros princípios importantes para se ter em mente ao lidar com LOD são os princípios FAIR (Wilkinson *et al*., 2016), acrônimo para quatro princípios fundamentais: *Findability* (respectivo a possibilidade de o objeto ser encontrado), *Accessibility* (acessibilidade), *Interoperability* (interoperabilidade) e *Reusability* (reutilização). Estes princípios do LOD e do FAIR são em certa medida recomendados no guia CCO (Baca *et al*, 2006, p.30), no sentido de orientar produtores e editores de dados culturais na organização e representação de conteúdo em várias mídias a

partir do uso de vocabulários publicados na internet, como os vocabulários controlados do The Getty Research Institute (Trust, 2024) ou os arquivos de autoridades da Biblioteca do Congresso, ajudando a maximizar a integração de coleções culturais na Web, conforme se pode visualizar no modelo apresentado na Figura 1.

Também digno de nota na Figura 1, no que diz respeito à qualidade de dados, são as pesquisas envolvendo técnicas de Ciência de Dados (CD) aplicadas em processos documentais desenvolvidas ao longo da última década (Harper, 2016; Chardonnens; Rizza; Coeckelbergs; Holland, 2018; Romero, 2019; Liao; Zhao, 2020; Purwitasari *et al*., 2020; Lorenzini; Rospocher; Tonelli, 2021; Candela, 2023) com temas dedicados a automação e processamento de dados e metadados, o que constitui também interesse para a CI, logo, possibilitando visualizar oportunidades interdisciplinares para os profissionais da informação em serviços de informações inovadores e de valor à sociedade. Podem-se citar a automação de tarefas repetitivas envolvendo catalogação (como o proposto na presente pesquisa envolvendo processamento e avaliação diagnóstica de dados, incluindo a técnica de expressões regulares, denominada regex), e 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; Garoufallou, 2020).

Considera-se, portanto, que o esforço empregado em pesquisas interdisciplinares envolvendo campos de conhecimento como a CI, a CD e a CC apóia à estruturação de sistemas de informação inteligentes na área da cultura, como agregadores de dados culturais, repositórios e bibliotecas digitais, adequados à preservação, ao acesso e à recuperação de objetos digitais culturais qualificados na internet. Nesse sentido, a automação em conjunto com o emprego de padrões de

documentação é vista como um mecanismo estratégico nessa estruturação, uma vez que otimiza processos de análise e avaliação de formatos, conteúdos, contextos e estruturas de um documento, evidenciando os metadados, em seu repertório conceitual e operacional, como elemento central de representação e gestão ao longo de todo o ciclo de vida dos documentos.

3 PROCEDIMENTO METODOLÓGICO

Metodologicamente, este estudo adotou uma abordagem aplicada, combinando elementos qualitativos e quantitativos, além de exploratória e descritiva. A abordagem quantitativa foi incluída neste estudo para quantificar a adequação das coleções aos padrões recomendados pelo CCO. Para alcançar isso, foi aplicada uma fórmula matemática para calcular o índice de adequação, conforme será exibida adiante.

Diante do contexto de uso do INBCM na arquitetura das bases de dados do Ibram, algumas decisões metodológicas são importantes de serem elucidadas inicialmente para fins de entendimento dos dados trabalhados na pesquisa. De acordo com a versão mais recente do INBCM (de 31 de agosto de 2021), para a identificação do bem cultural musealizado no INBCM, os elementos específicos de descrição para a área da Museologia são num total de 15, sendo 9 (nove) de entrada obrigatória e 6 (seis) de entrada facultativa.

O processo de alinhamento (mapeamento) foi o primeiro passo crucial deste estudo. O objetivo deste passo foi estabelecer a correspondência entre os elementos descritivos da normativa do INBCM e do guia de catalogação (CCO), conforme apresentado em (LEMOS; COELHO JUNIOR, 2023). Logo, o experimento da presente pesquisa considerou 7 (sete) dimensões analíticas enumeradas e descritas a seguir:

I – *Object Naming*: fornece maneiras de se referir a uma obra, definindo o que está sendo catalogado.

II – *Creator Information*: identifica o criador de uma obra (podendo ser vários), incluindo pessoa, física ou jurídica, conhecida pelo nome ou anônima.

III – *Physical Characteristics*: descreve a aparência de uma obra, apresentando características de sua forma física.

IV – *Stylistic, Cultural, and Chronological Information*: descreve características estilísticas de uma obra, origens culturais e data de design ou criação.

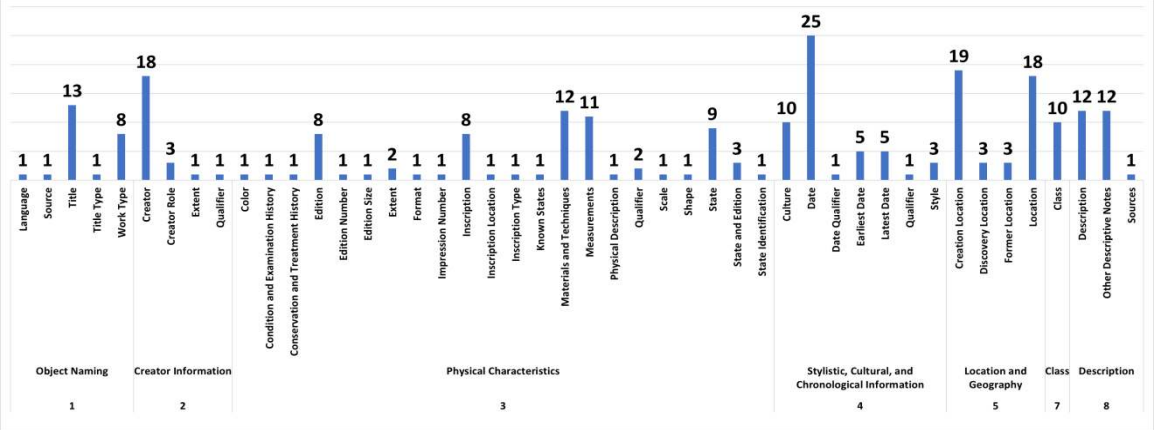
V – *Location and Geography*: trata de elementos que registram informações geográficas e de localização, tais como localização atual, locais ao longo do tempo, localização de criação e localização de descoberta.

VII – *Class*: classifica uma obra específica a outras obras com características semelhantes, muitas vezes com base em esquema organizacional de um determinado repositório ou coleção.

VIII – *Description*: associa campos específicos em todo o registro, consistindo de uma nota descritiva que geralmente é um texto relativamente breve, detalhando o conteúdo e o contexto da obra.

De acordo com o mapeamento realizado em todas as regras explicitadas nos capítulos ora elencados do guia CCO (I, II, III, IV, V, VII e VIII) foram identificadas 244 regras, incluindo 122 regras distintas. A distribuição dessas regras por capítulo e alinhadas ao INBCM pode ser observada na Figura 2. Torna-se importante salientar que o capítulo VI, dedicado ao elemento central “assunto” (*Subject*), não é considerado nos elementos de descrição para identificação do bem cultural de caráter museológico do INBCM, logo, não foi inserido no processo de análise e alinhamento.

Figura 2 - Regras de catalogação CCO alinhadas ao INBCM



Fonte: Elaborado pelos autores (2023).

A avaliação semiautomática foi realizada por meio da linguagem Python, utilizando expressões regulares da linguagem formal regex. O Quadro 1 apresenta as regras de catalogação que foram implementadas pelo algoritmo. A aplicação desenvolvida e denominada DataQ Culture pode ser utilizada por diferentes usuários para avaliar a qualidade dos dados de suas bases de dados e direcionar esforços para ações preventivas e corretivas.

Para cada regra associada ao elemento de metadado pertencente a uma dimensão, o registro de dado correspondente (*string* avaliada) recebeu o valor 0 (zero) ou 1 (um). O valor 1 (um) foi atribuído quando o registro de dado atendeu ao critério (regra) recomendado pelo CCO; e o valor 0 (zero) quando não atendeu. Por fim, o índice de adequação é dado pela fórmula: **índice b = (∑ Valor1 / (∑ Valor1+ ∑ Valor0)) *100** onde **b** é a base com a amostra de dados de uma coleção particular; e **índice** é o percentual de adequação obtido em relação à dimensão, a elemento de metadado e à regra de catalogação para um determinado museu e coleção.

Quadro 1 - Regras de catalogação e regex utilizados na pesquisa

#	Regra	Elemento de descrição	*Regex*
1	Fazer uso de vocabulário controlado	*Class, Creation Location, Creator, Inscription, Location, Materials and Techniques, Measurements, Physical Description, Work Type*	Não se aplica. Utilizado API do Tainacan ou Indicação do usuário
2	Evitar abreviações	Class, Creation Location, Creator, Description, Location, Materials and Techniques, Other Descriptive Notes, Title, Work Type	[A-ZÁÉÍÓÚÜÑ][A-Za-z0-9áéíóúüñ]*\.
3	Usar o mesmo idioma do catálogo	Creation Location, Date, Description, Location, Materials and Techniques, Other Descriptive Notes, Work Type	Não se aplica (Utilizado Python – Langdetect)
4	Abreviar unidades métricas de acordo com o Sistema Internacional (m, cm, mm, g, kg, kb, Mb, Gb)	Measurements	(?i)\b\d+(?:\.\d+)?\s*(?:m|cm|g|kg|B|KB|MB|GB|TB)\b
5	Capitalize as iniciais de nomes próprios e da primeira palavra, para outros termos use letras minúsculas	Creation Location, Date, Description, Location, Materials and Techniques, Other Descriptive Notes, Title, Work Type	^([A-Z0-9]){1}(.)*
6	Medidas geralmente incluem duas casas decimais para medidas métricas	Measurements,	\d+[,.]\d{2}\b
7	Não usar capitalização	Measurements	[A-Z]
8	Utilizar números inteiros ou frações decimais	Measurements	[0-9][,\.]
9	Não pode ficar vazio	Class, Creator, Inscription, Materials and Techniques, Measurements, Work Type, Title, Date, Location	.+
10	Usar singular	Class, Materials and Techniques, Work Type	\b\w+[sS]\b
11	Anos com menos que 4 (quatro) dígitos, inserir 0 (zero) a esquerda	Date	\b\d{4}\b
12	Não usar pontuação, exceto hífen	Work Type	^[a-zA-Z\u00C0-\u00FF 0-9\-\—\–]*$
13	Não utilizar apóstrofo	Date	[\']+
14	Não utilizar artigos	Title	\b(?:o(s)?|a(s)?|um(a)?(s)?|uns)\b|\b(?:O(s)?|A(s)?|Um(a)?(s)?|Uns)\b
15	Seguir padrão para registro de hora, minutos e segundos	Date	(?P<hours>0?[0-9]|1[0-9]|2[0-3]):(?P<minutes>60|[0-5][0-9]):(?P<seconds>60|[0-5][0-9])
16	Seguir padrão pra registro de dia, mês e ano de data	Date	^(?:([0-9]{1,2})(\/|-|.|\S)([0-9]{1,2})(\/|-|.|\S)([0-9]{4}))
17	Use traço para separar intervalo de anos	Date	\b\d{4}\s*-\s*\d{4}\b

Fonte: elaborado pelos autores (2023).

Para o desenvolvimento da aplicação, utilizou-se do modelo de desenvolvimento de *software* em cascata, também conhecido como modelo Cascata (*Waterfall*, em inglês) por ser uma das metodologias de desenvolvimento mais antigas e conhecidas em Engenharia de Software (Pressman, 2014; Sommerville, 2016). Nesta abordagem, as etapas do projeto são realizadas sequencialmente, uma após a outra, e cada etapa é concluída antes que a próxima comece. O modelo de desenvolvimento em cascata é baseado em um conjunto de etapas sequenciais e distintas, sendo geralmente organizadas conforme descritas a seguir:

- Definição de requisitos: nesta etapa, são definidos e documentados os requisitos do *software*, incluindo as funcionalidades que ele deve possuir; as restrições de *design*; e as expectativas do cliente.
- Design: nesta etapa, é desenvolvida a arquitetura do *software*, com base nos requisitos definidos na etapa anterior. O design pode incluir fluxogramas, diagramas, modelos de dados e outros documentos que descrevam a estrutura e o funcionamento do *software*.
- Implementação: na etapa de implementação, o código é escrito de acordo com o design desenvolvido na etapa anterior. A implementação pode incluir a codificação, testes unitários e integração com outros componentes do software.
- Testes: nesta etapa, são realizados testes para garantir que o *software* funcione corretamente, atenda aos requisitos e não apresente falhas. Os testes podem ser automatizados ou manuais, e devem ser realizados em diferentes cenários de uso do *software*.
- Implantação: a implantação é a etapa final do processo, em que o *software* é entregue ao cliente. Isso pode envolver a instalação e configuração do

software em um ambiente de produção, treinamento do usuário final e suporte técnico.

À luz das orientações do método Cascata, os processos listados foram executados durante o desenvolvimento da aplicação. Na primeira etapa, foram identificados quatro requisitos básicos que a ferramenta deveria cumprir: i) receber um conjunto de dados (*dataset*, em inglês); ii) realizar o alinhamento entre os elementos do *dataset* que o usuário forneceu com os elementos do CCO; iii) apresentar os resultados da avaliação de qualidade de dados; e iv) indicar como a qualidade de dados do *dataset* avaliado pode ser melhorada.

A segunda etapa pode ser visualizada no fluxograma dos processos realizados pela aplicação (Figura 3). As ações realizadas pelo usuário estão destacadas em amarelo, enquanto as ações realizadas pela aplicação estão em rosa.

Figura 3 - Diagrama de processos da aplicação

Fonte: Elaborado pelos autores (2023).

Na etapa 3, foram empregadas as linguagens de programação Python no *backend* (processo interno da ferramenta) e HTML, CSS e JavaScript no *frontend* (interface do usuário). Especificamente no *backend*, foram utilizadas diversas bibliotecas, como o Pandas para processamento de dados, o *framework* web Flask[5] para criação das páginas e rotas da ferramenta que o usuário navegará e a biblioteca **re** para processamento de expressões regulares. É importante destacar que,

5 https://flask.palletsprojects.com/en/2.2.x/.

diferentemente da avaliação realizada no estudo de caso com os acervos do Ibram, nesta ferramenta não será utilizado o BeautifulSoup e Requests, isto é,será esperado do usuário o fornecimento de uma base de dados em *Comma Separated Values* (CSV)[6] para a avaliação de qualidade de maneira local, sem necessidade de raspagem ou captação de fontes externas.

Durante a etapa 4, utilizou-se os *datasets* do Ibram exportados em massa. Realizaram-se processos de envio, alinhamento, salvamento do alinhamento e processamento com geração dos resultados, além de testar a recuperação do alinhamento e o *upload* de fontes inválidas para verificar o comportamento adequado da aplicação.

Por fim, na etapa 5, a aplicação é implantada localmente no computador do usuário e fica disponível para acesso a qualquer pessoa na mesma rede. Para possibilitar isso, o código-fonte da aplicação foi disponibilizado no *Github* (Coelho Júnior, 2023) juntamente com um guia passo a passo em texto e em vídeo para a sua execução, permitindo um acesso livre e a implantação por qualquer usuário interessado.

4 RESULTADOS

Com o objetivo de ampliar o acesso e a utilização de ferramentas para avaliação da qualidade de dados em instituições de acervos culturais com base em padrões de referência, tornou-se necessário o desenvolvimento de uma ferramenta de fácil acesso, reprodução e com resultados orientadores para ações mais efetivas e significativas para a melhoria da qualidade de metadados de acervos culturais. Assim, o desenvolvimento da DataQ Culture surge para preencher essa lacuna.

6 https://en.wikipedia.org/wiki/Comma-separated_values.

A funcionalidade básica da ferramenta consiste em processar as *regex* desenvolvidas em uma base de dados fornecida pelo usuário. Desta forma, a funcionalidade inicial é uma interface para envio de um arquivo com os dados a serem avaliados. Assim, a Figura 4 apresenta a interface de envio de arquivo.

Figura 4 - Interface de envio de base de dados para avaliação

Fonte: Elaborado pelos autores (2023).

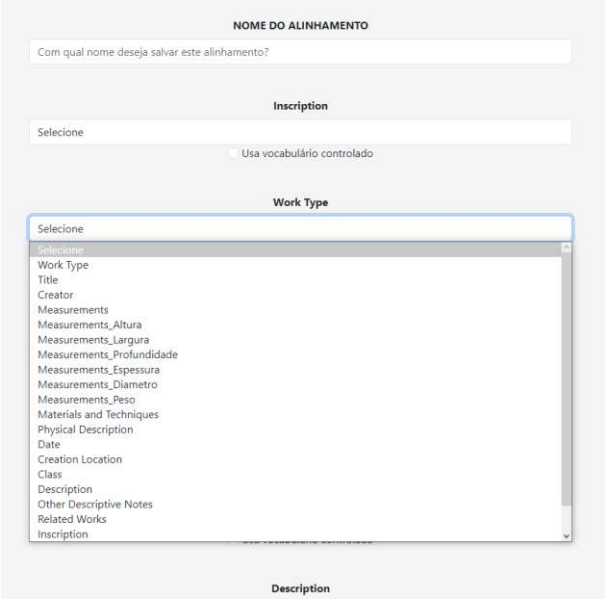
Na interface de envio, é apresentado um campo para seleção de um arquivo no formato CSV. Este formato foi considerado pela sua versatilidade na leitura entre sistemas, já que um arquivo em aplicativo de planilha pode ser exportado neste formato, assim como qualquer banco de dados. Com o *upload* do arquivo CSV, o DataQ Culture avalia o formato do arquivo, validando se é de fato um arquivo CSV; faz a identificação do *encoding*[7] (que se refere à maneira como os caracteres são armazenados em um arquivo de texto) e do delimitador do CSV. O delimitador é o caractere que faz a separação entre as colunas no arquivo, geralmente são utilizados

[7] https://pt.wikipedia.org/wiki/Codificação_de_caracteres.

vírgulas ou ponto e vírgulas, mas também podem ser utilizados um caractere invisível quando se aperta *tab* no teclado. Por isso, é importante ter uma função dedicada à tratativa destes dois pontos, pois com a variedade de sistemas operacionais, diversos tipos de *encoding* podem ser apresentados à ferramenta, além de diferentes delimitadores e arquivos CSV incompletos, corrompidos e inválidos.

Outra tarefa fundamental para a realização da avaliação de qualidade de dados é realizar o alinhamento entre o acervo submetido pelo usuário e as dimensões e elementos discricionais do CCO. Para esse fim, foi elaborada uma tela de alinhamento, conforme apresentado na Figura 5.

Figura 5 - Tela de alinhamento entre elementos discricionais base do usuário com os elementos discricionais do CCO



Fonte: Elaborado pelos autores (2023).

Nesta tela, o usuário pode fazer o alinhamento com o CCO independentemente do padrão de documentação utilizado no *dataset* enviado. Na parte superior da tela, há um campo para inserção do nome do alinhamento. No restante da tela, é possível ver para cada elemento discricional presente no arquivo do usuário, as opções de

seleção para as colunas correspondentes do CCO. Além disso, abaixo de cada um dos elementos discricionais, há a opção de indicar se este faz uso de um vocabulário controlado.

Após o alinhamento, a configuração é salva e pode ser reutilizada sempre que uma base com a mesma configuração de cabeçalho é carregada pelo usuário, reduzindo o retrabalho e otimizando o tempo. É possível ainda editar um alinhamento existente, excluir caso necessário ou, simplesmente, criar um novo, como pode ser visto na Figura 6.

Figura 6 - Tela de alinhamento com indicação de alinhamento já existente

Fonte: Elaborado pelos autores (2023).

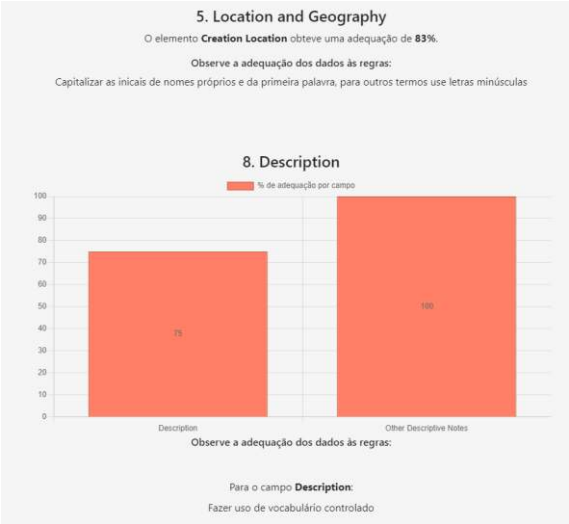
Após o processamento da base, um relatório é gerado com várias métricas, a saber: (i) adequação geral do arquivo avaliado; (ii) adequação por dimensão do CCO; (iii) para cada dimensão que não obteve a pontuação máxima, é exibida a taxa de adequação por elemento discricional. Neste ponto, se houver apenas um elemento discricional em alguma das dimensões, um texto é apresentado com a taxa de adequação do elemento discricional. Caso mais de um elemento discricional pertença à dimensão, é exibido um gráfico com a respectiva pontuação dos elementos; e (iv) para cada elemento discricional que não obteve a adequação máxima, são exibidas as regras que poderiam melhorar a adequação do elemento discricional.

Figura 7 - Tela principal com taxa de adequação de coleção avaliada

Fonte: Elaborado pelos autores (2023).

Essas características podem ser visualizadas nas Figuras 7 e 8. Por fim, no final da página, é possível baixar uma planilha com todos os valores avaliados e a indicação se estava adequado à regra ou não.

Figura 8 - Regras indicadas para elementos que não alcançaram 100% de adequação



Fonte: elaborado pelos autores (2023).

A avaliação diagnóstica resultante do experimento permitiu aferir nas coleções museológicas do Ibram que os dados carecem de um tratamento mais adequado em dimensões como características físicas do objeto de informação, descrição, localização geográfica e informações cronológicas. Por outro lado, as coleções se mostraram qualificadas em termos do uso adequado de taxonomias para a dimensão classificação. Ademais, de uma maneira geral, a ferramenta DataQ Culture permite que um usuário comum realize a avaliação de qualidade de dados de seus acervos de forma simples e interativa, com regras baseadas em padrões de referência, e com geração de relatórios com indicação de ações que trarão resultados efetivos; permite também que gestores de acervos e coleções façam a validação de seus dados sem maiores dificuldades.

5 DISCUSSÃO

O uso de padrões de dados (Gilliland, 2016) em temos de estrutura (ex.: padrão de metadados), valor (ex.: linguagem documentária), conteúdo e formato (ex.: regra de catalogação), e comunicação (ex.: um padrão de metadados num formato legível para a máquina), juntamente com o uso de um guia de referência (como o CCO), 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 de boas recomendações e padrões de documentação de referência 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 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 (por meio de vocabulários controlados) ajudando, assim, a consistência dos dados e auxiliando os 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).

Nesse sentido, a avaliação da qualidade de dados proporcionada pela aplicação DataQ Culture pode ser considerada 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. A aplicação foi estabelecida por meio de um arcabouço metodológico reprodutível e semiautomatizado (Wang, 2018) fundamentado em princípios teórico-metodológicos da Ciência da Computação (Pressman, 2014; Sommerville, 2016) e da Ciência da Informação (Baca *Et Al*., 2006, Harpring, 2022), que pode ser utilizado como aliado na melhoria da qualidade da catalogação e consequente recuperação da informação (International Federation OF Library Associations AND Institutions, 2016). 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.

Por fim, mas não menos importante, o uso de regras de catalogação, como as previstas no CCO, determinam como elaborar o conteúdo da descrição de um recurso de informação, os pontos de acesso e os relacionamentos entre estes, tornando-se práticas essenciais na padronização, na descrição e, portanto, na agregação semântica de recursos de informação (Gilliland, 2016).

6 CONSIDERAÇÕES FINAIS

Os resultados apresentados no presente artigo foram alcançados por meio da tecnologia *regex* juntamente com o uso de um padrão de documentação de

referência, o guia de catalogação COO. A partir do alinhamento entre os elementos descritivos do INBCM e do CCO, foi possível realizar a implementação de uma porção de regras de catalogação do CCO com uso da linguagem Python. A aplicação possibilitou apurar o índice de adequação da qualidade de dados em todos os registros de metadados das 22 coleções museológicas vinculadas ao Ibram, sendo mais de 17 mil itens processados.

Adicionalmente, a aplicação de avaliação da qualidade de dados DataQ Culture foi implementada para que diferentes usuários possam realizar a mesma avaliação em seus acervos, independente do padrão de documentação adotado, levando a uma economia de tempo para o profissional da informação na ação de avaliar a qualidade de bases de dados legadas, direcionando o esforço do usuário para ações preventivas e corretivas a partir das informações diagnósticas levantadas, respondendo, assim, à questão de pesquisa e indicando como melhorar a qualidade de dados em acervos culturais.

Reforça-se que a avaliação da qualidade de dados é incipiente e pouco desenvolvida no domínio da cultura e que a semiautomatização dessa avaliação é um ponto de partida para o direcionamento de esforços para a melhoria da qualidade de dados no domínio. Conclui-se, portanto, que o modelo de avaliação de qualidade de dados proposto nesta pesquisa, com base no guia de catalogação de objetos culturais CCO, mostrou-se eficaz para diagnosticar as discrepâncias e deficiências nos acervos museológicos sob gestão do Ibram. A utilização de práticas de catalogação maduras, oriundas de modelos de referência, pode contribuir para qualificar os atuais padrões de documentação por meio de instrumentos de organização da informação mais sofisticados e orientados para os usuários finais dos sistemas de informação. Além disso, a ferramenta desenvolvida pode auxiliar os profissionais da informação no

acompanhamento da qualidade dos dados de seus acervos e está disponível para uso por outras instituições e profissionais interessados.

REFERÊNCIAS

BACA, Murtha; HARPRING, Patricia; LANZI, Elisa; MCRAE, Linda; WHITESIDE, Ann. **Cataloging cultural objects:** a guide to describing cultural works and their images. Chicago: American Library Association, 2006.

BARBOSA, Alice Príncipe. **Novos rumos da catalogação**. Rio de Janeiro: BNG/BRASILART, 1978.

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

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked Data - the story so far. **International Journal on Semantic Web and Information Systems (IJSWIS)**, v. 5, n. 3, p. 1–22, 2009. Disponível em: https://www.researchgate.net/publication/225070216_Linked_Data_The_Story_so_Far. Acesso em: 4 dez. 2024.

BRASIL. 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: Diário Oficial, 2021. Disponível em: https://www.in.gov.br/web/dou/-/resolucao-normativa-ibram-n-6-de-31-de-agosto-de-2021-342359740. Acesso em: 12 jul. 2023.

CANDELA, Gustavo. Towards a semantic approach in GLAM Labs: the case of the Data Foundry at the National Library of Scotland. **arXiv**, 26 jan. 2023. Disponível em: http://arxiv.org/abs/2301.11182. Acesso em: 26 fev. 2023.

CHAPMAN, Arthur D. **Principles of Data Quality**. Global Biodiversity Information Facility: Copenhagen, 2005. Disponível em: https://www.gbif.org/document/80509. Acesso em: 12 jul. 2023.

CHARDONNENS, Anne; RIZZA, Ettore; COECKELBERGS, Mathias; HOLLAND, Seth Van. Mining user queries with information extraction methods and linked data. **Journal of Documentation**, [*s. l.*], v. 74, n. 5, p. 936-950, 2018. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/jd-09-2017-0133/full/html. Acesso em: 4 dez. 2024.

COELHO JÚNIOR, Abeil. DataQ-Culture. 2023. **Github**. Disponível em: https://github.com/AbeilCoelho/DataQ-Culture. Acesso em: 12 jul. 2023.

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

GILLILAND, Anne J. Setting the Stage. *In*: BACA, Murta. (ed.). **Introduction to metadata**. 3. ed. Los Angeles: Getty Research Institute, 2016. Disponível em: https://www.getty.edu/publications/intrometadata/setting-the-stage/. Acesso em: 12 jul. 2023.

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, 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**. 2022. Disponível em: https://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html#endnote1CCO. Acesso em: 11 jul. 2023.

HJØRLAND, Birger. Data (with big data and database semantics). **Knowledge Organization**, [*s. l.*], v. 45, n. 8, p. 643-652, 2018. Disponível em: https://www.isko.org/cyclo/data. Acesso em: 29 ago. 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: 12 jul. 2023.

JOUDREY, Daniel N.; TAYLOR, Arlene G.; MILLER, David P. **Introduction to cataloging and classification**. 11 ed. Santa Barbara: ABC-CLIO, 2015.

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

LEMOS, Daniela Lucas da Silva; COELHO JUNIOR, 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. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/90510. Acesso em: 4 dez. 2024.

LIAO, Xiaofeng; ZHAO, Zhiming. Unsupervised Approaches for Textual Semantic Annotation, A Survey. **ACM Computing Surveys**, [*s. l.*], v. 52, n. 4, p. 1-45, 2020. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/90510/52571. Acesso em: 4 dez. 2024.

LEMOS, Daniela Lucas da Silva; MARTINS, Dalton Lopes; SOUZA, Renato Rocha. Organização e representação da informação e do conhecimento em contextos informacionais: uma proposta de um modelo teórico-conceitual para a qualidade de objetos culturais digitais. **Fronteiras da Representação do Conhecimento**, Belo Horizonte v. 3, n. 2, ano III, Editorial, mar. 2023. Disponível em: https://periodicos.ufmg.br/index.php/advances-kr/article/view/45974/38882. Acesso em: 4 dez. 2024.

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. Disponível em: https://link.springer.com/article/10.1007/s00799-021-00302-1. Acesso em: 4 dez. 2024.

MARTINS, Dalton Lopes; LEMOS, Daniela Lucas da Silva; OLIVEIRA, Luis Felipe Rosa; SIQUEIRA, Joyce; CARMO, Danielle; MEDEIROS, Vinicius Nunes. 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.], v. 74, n. 6, p. 707-726, 2022. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.24650. Acesso em: 4 dez. 2024.

MEY, Eliane Serrão Alves; SILVEIRA, Naira Christofoletti. **Catalogação no Plural**. Brasília: Briquet de Lemos, 2009.

PRESSMAN, Roger. **Software engineering**: a practitioner's approach. 6th ed. Boston, Mass.: McGraw-Hill, 2005.

PURWITASARI, Diana; FATICHAH, Chastine; SUMPENO, Surya; STEGLICH, Christian; PURNOMO, Mauridhi Hery. Identifying collaboration dynamics of bipartite author-topic networks with the influences of interest changes. **Scientometrics**, v. 122, n. 3, p. 1407-1443, 2020. Disponível em: https://link.springer.com/article/10.1007/s11192-019-03342-2. Acesso em: 4 dez. 2024.

ROMERO, Gustavo Candela. **Publicación y enriquecimiento semántico de datos abiertos en bibliotecas digitales**. 2019. Tese (Doutorado em Informática) – Universidad de Alicante, Espanha, 2019. Disponível em: https://rua.ua.es/dspace/handle/10045/97353. Acesso em: 1 ago. 2022.

SOMMERVILLE, Ian. **Software engineering**. 9th ed. Boston: Pearson, 2011.

SVENONIUS, Elaine. **The intellectual foundation of information organization**. Cambridge: The MIT Press, 2000.

TRUST, Jean Paul Getty. **The Getty Research Institute - Getty Vocabularies**. 2022. Disponível em: https://www.getty.edu/research/tools/vocabularies/. Acesso em: 17 jul. 2024.

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. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/dta-07-2020-0167/full/html. Acesso em: 4 dez. 2024.

WANG, Lin. Twinning data science with information science in schools of library and information science. **Journal of Documentation**, [*s. l.*], v. 74, 2018. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/jd-02-2018-0036/full/html. Acesso em: 4 dez. 2024.

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. Disponível em: https://www.nature.com/articles/sdata201618. Acesso em: 4 dez. 2024.

ZENG, Marcia Lei. Interoperability. **Knowledge Organization**, [*s. l.*], v. 46, n. 2, p. 122-146, jan. 2019. Disponível em: https://www.nomos-elibrary.de/10.5771/0943-7444-2019-2-122.pdf. Acesso em: 4 dez. 2024.

ZENG, Marcia Lei; QIN, Jian. **Metadata**. 2. ed. Atlanta: ALA Neal-Schuman. 2016.

AGRADECIMENTOS

Agradecemos ao Instituto Brasileiro de Museus pelo financiamento a esta pesquisa.

tpbci@ancib.org @anciboficial @ancib_brasil