UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS JURÍDICAS E ECONÔMICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

ABEIL COELHO JÚNIOR

QUALIDADE DE DADOS EM ACERVOS DO PATRIMÔNIO CULTURAL: UMA PROPOSTA DIAGNÓSTICA SEMIAUTOMÁTICA PARA OBJETOS CULTURAIS SOB GESTÃO DO INSTITUTO BRASILEIRO DE MUSEUS

VITÓRIA (ES)
2023

ABEIL COELHO JÚNIOR

QUALIDADE DE DADOS EM ACERVOS DO PATRIMÔNIO CULTURAL: UMA PROPOSTA DIAGNÓSTICA SEMIAUTOMÁTICA PARA OBJETOS CULTURAIS SOB GESTÃO DO INSTITUTO BRASILEIRO DE MUSEUS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal do Espírito Santo (PPGCI/UFES) como requisito para a obtenção do título de Mestre em Ciência da Informação.

Orientadora: Profa. Dra. Daniela Lucas da Silva Lemos

Coorientador: Prof. Dr. Fabrício Martins Mendonça

Linha de pesquisa: 2 Memória, Representação e Informação.

VITÓRIA (ES)
2023

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

---

Coelho Júnior, Abeil, 1997-
C672q Qualidade de dados em acervos do patrimônio cultural: : uma proposta diagnóstica semiautomática para objetos culturais sob gestão do Instituto Brasileiro de Museus / Abeil Coelho Júnior. - 2023.
110 f. : il.

Orientadora: Daniela Lucas da Silva Lemos.
Coorientador: Fabricio Martins Mendonca.
Dissertação (Mestrado em Ciência da Informação) - Universidade Federal do Espírito Santo, Centro de Ciências Jurídicas e Econômicas.

1. Qualidade de dados. 2. Metadados. 3. Catalogação. 4. Coleções culturais. 5. Automação em tratamento documental. 6. Museologia. I. Lemos, Daniela Lucas da Silva. II. Mendonca, Fabricio Martins. III. Universidade Federal do Espírito Santo. Centro de Ciências Jurídicas e Econômicas. IV. Título.

CDU: 001

ABEIL COELHO JÚNIOR

QUALIDADE DE DADOS EM ACERVOS DO PATRIMÔNIO CULTURAL: UMA PROPOSTA DIAGNÓSTICA SEMIAUTOMÁTICA PARA OBJETOS CULTURAIS SOB GESTÃO DO INSTITUTO BRASILEIRO DE MUSEUS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal do Espírito Santo (PPGCI/UFES) como requisito para a obtenção do título de Mestre em Ciência da Informação.

Linha de pesquisa 2: Memória, Representação e Informação.

Aprovado em 24 de março de 2023.

[assinatura digital]
Profa. Dra. Daniela Lucas da Silva Lemos
Orientadora (PPGCI/UFES)

[assinatura digital]
Prof. Dr. Fabrício Martins Mendonça
Coorientador (UFJF)

[assinatura digital]
Prof. Dr. Henrique Monteiro Cristóvão
PPGCI/UFES

[assinatura digital]
Prof. Dr. Dalton Lopes Martins
UnB

[assinatura digital]
Profa. Dra. Renata Cardozo Padilha
UFSC

Datas e horários baseados em Brasília, Brasil
Sincronizado com o NTP.br e Observatório Nacional (ON) em 10/04/2023 às 09:12:55

Folha de aprovação - Abeil

Data e Hora de Criação: 04/04/2023 às 15:40:31

Documentos que originaram esse envelope:
- Folha de aprovação - Abeil.pdf (Arquivo PDF) - 1 página(s)

Hashs únicas referente à esse envelope de documentos

[SHA256]: dfe27973642a178890c917cbc20aa1b3a86fa6038652b374dcaaea8df4aad700
[SHA512]: 1b24ef0a046e72c9e881f396702e049e9b3ad3a0752ace51f2f3e4b9cc7f2fed40b9691cdb518bb29739db17f826e44e23d15cd2c51bb9c7e20ce5bdd43f7b59

Lista de assinaturas solicitadas e associadas à esse envelope

ASSINADO - Daniela Lucas Da Silva Lemos (daniela.l.silva@ufes.br)
Data/Hora: 04/04/2023 - 15:43:28, IP: 200.137.65.107, Geolocalização: [-20.276517, -40.305470]
[SHA256]: 37ee9df04496f16a71158dca484b18377b411ffb178e30dd78eefed10ba00cdd

ASSINADO - Dalton Lopes Martins (daltonmartins@unb.br)
Data/Hora: 04/04/2023 - 15:47:33, IP: 164.41.122.160
[SHA256]: 509def3dd096d2f1d5241f1103902120d359b5d7045efedbc5bb81d25f11c653

ASSINADO - Fabrício Mendonça (fabricio.mendonca@ice.ufjf.br)
Data/Hora: 05/04/2023 - 06:53:44, IP: 194.117.31.200, Geolocalização: [41.1794247, -8.5953701]
[SHA256]: edc10866b13e5a632c98d5aeb629bf0bd385f22efb710c09147430714e8b9972

ASSINADO - Henrique Monteiro Cristovão (henrique.cristovao@ufes.br)
Data/Hora: 09/04/2023 - 23:54:45, IP: 177.104.235.45, Geolocalização: [-20.258744, -40.276258]
[SHA256]: 3204ce69bcc275748a7b7aa29cba51e1d808d427e6216fd4503f586440f8f95e

ASSINADO - Renata Cardozo Padilha (renata.padilha@ufsc.br)
Data/Hora: 10/04/2023 - 09:12:55, IP: 191.191.33.112, Geolocalização: [-27.5973, -48.5496]
[SHA256]: acb952a75b5a6bfc3f4ff2ad1c382d2f4889def709bf08cb694c5f766c0c97fa

Histórico de eventos registrados neste envelope

10/04/2023 09:12:55 - Envelope finalizado por renata.padilha@ufsc.br, IP 191.191.33.112
10/04/2023 09:12:55 - Assinatura realizada por renata.padilha@ufsc.br, IP 191.191.33.112
10/04/2023 09:12:07 - Envelope visualizado por renata.padilha@ufsc.br, IP 191.191.33.112
09/04/2023 23:54:45 - Assinatura realizada por henrique.cristovao@ufes.br, IP 177.104.235.45
09/04/2023 23:54:19 - Envelope visualizado por henrique.cristovao@ufes.br, IP 177.104.235.45
05/04/2023 06:53:44 - Assinatura realizada por fabricio.mendonca@ice.ufjf.br, IP 194.117.31.200
05/04/2023 06:53:34 - Envelope visualizado por fabricio.mendonca@ice.ufjf.br, IP 194.117.31.200
04/04/2023 15:47:33 - Assinatura realizada por daltonmartins@unb.br, IP 164.41.122.160
04/04/2023 15:47:30 - Envelope visualizado por daltonmartins@unb.br, IP 164.41.122.160
04/04/2023 15:43:28 - Assinatura realizada por daniela.l.silva@ufes.br, IP 200.137.65.107
04/04/2023 15:43:23 - Envelope visualizado por daniela.l.silva@ufes.br, IP 200.137.65.107
04/04/2023 15:42:15 - Envelope registrado na Blockchain por daniela.l.silva@ufes.br, IP 200.137.65.107
04/04/2023 15:42:13 - Envelope encaminhado para assinaturas por daniela.l.silva@ufes.br, IP 200.137.65.107
04/04/2023 15:40:31 - Envelope criado por daniela.l.silva@ufes.br, IP 200.137.65.107

AGRADECIMENTOS

Durante toda a minha vida, sempre acreditei que a oportunidade é como um cavalo selvagem galopando ao nosso redor. Enquanto nenhum deles se apresenta acessível, devemos nos preparar e capacitar para que, quando a oportunidade passar perto o suficiente, possamos laçá-la com habilidade e montá-la em direção a novos lugares e conquistas. Assim como um bom cavaleiro, é necessário ter paciência, perseverança e treinamento constante para estar pronto para agarrar a oportunidade assim que ela aparecer.

Graças às oportunidades que se apresentaram ao longo do caminho, consegui superar minhas próprias expectativas e alcançar resultados além do que eu imaginava. Fui agraciado com a chance de cursar minha graduação na renomada Universidade Federal do Espírito Santo e, posteriormente, ingressar em seu programa de mestrado. Essa jornada me proporcionou a realização de uma defesa de dissertação de mestrado diante de uma banca de prestígio, permitindo-me avançar ainda mais em minha trajetória acadêmica e profissional.

Apesar de todas as dificuldades e desafios que enfrentei ao longo da caminhada, sou grato a Deus pela sabedoria que me concedeu nos momentos sombrios e pela perseverança que me sustentou nos momentos difíceis. Obrigado, Senhor, por guiar meus passos e me dar forças para seguir em frente.

Gostaria também de expressar minha profunda gratidão à minha esposa, Samantha, por ser um pilar de apoio inabalável ao longo dessa jornada árdua. Sua presença constante e suporte incansável me deram a força necessária para continuar em frente. Muito obrigado, minha amada!

Não poderia deixar de expressar minha gratidão à minha fiel companheira de quatro patas, Zoey. Ela foi uma fonte constante de alegria e leveza em meio a todas as adversidades, e me ajudou a manter o foco no presente e no que realmente importa. Zoey, obrigado por sempre estar ao meu lado.

Gostaria de expressar minha gratidão à minha família por entenderem e apoiarem meu período de ausência e introspecção pessoal. Apesar da distância, sei que estão sempre ao meu lado. Amo vocês.

Também gostaria de agradecer à Universidade Federal do Espírito Santo, ao Programa de Pós-Graduação em Ciência da Informação, ao corpo docente e aos meus colegas de turma pela ajuda e apoio ao longo dessa jornada. Agradeço também

ao Laboratório de Inteligência de Redes (UnB/Ufes) pela oportunidade de contribuir com seus projetos, e à Viação Águia Branca pela confiança em meu trabalho.

Por fim, quero fazer um agradecimento especial à minha professora orientadora, Daniela, por sempre me incentivar, compartilhar seus conhecimentos e não deixar que eu desanimasse diante das adversidades. Obrigado por tudo que fez por mim ao longo dessa jornada, serei eternamente grato. E não posso deixar de agradecer também ao meu mentor, Elias, cujos ensinamentos me permitiram chegar até aqui.

“Carry on, my wayward son; There'll be peace when you are done; Lay your weary head to rest; Don't you cry no more.”

Kansas

RESUMO

Nos últimos anos, houve um aumento na digitalização e disponibilização de dados de acervos culturais na internet. No entanto, a qualidade dos dados frequentemente não é levada em consideração, o que pode prejudicar a indexação, a busca e a navegação dos usuários em sistemas de recuperação da informação. O Instituto Brasileiro de Museus utiliza o Inventário Nacional de Bens Culturais Musealizados para a descrição e publicação de dados, mas este modelo não é direcionado para catalogação visando padrões e vocabulários padronizados. A presente pesquisa utiliza o guia *Cataloging Cultural Objects* como referência para avaliar a qualidade dos dados das bases de dados dos museus sob gestão do Ibram. O objetivo da pesquisa é desenvolver uma aplicação de avaliação diagnóstica semiautomática que permita a otimização da qualidade dos dados em acervos culturais. A metodologia utilizada foi a pesquisa aplicada, qualiquantitativa, exploratória e descritiva, a partir de um estudo de caso em 22 coleções de museus digitais. A avaliação semiautomática foi realizada por meio da linguagem *Python*, utilizando expressões regulares, e os resultados evidenciaram problemas de catalogação em relação às características físicas do objeto, informações cronológicas, localização geográfica e descrição. Por fim, a aplicação desenvolvida 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.

PALAVRAS-CHAVE: Qualidade de dados. Metadados. Catalogação. Coleções culturais. Automação em tratamento documental. Museologia. Coleções digitais.

ABSTRACT

In recent years, there has been an increase in the digitization and availability of cultural heritage data on the internet. However, the quality of the data is often not taken into consideration, which can hinder indexing, search, and navigation for users in information retrieval systems. The *Instituto Brasileiro de Museus* uses the *Inventário Nacional de Bens Culturais Musealizados* for the description and publication of data, but this model is not focused on cataloging for standardized standards and vocabularies. This research uses the Cataloging Cultural Objects guide as a reference to assess the quality of data in museum databases managed by Ibram. The objective of the research is to develop a semi-automatic diagnostic evaluation application that allows for the optimization of data quality in cultural heritage collections. The methodology used was applied, qualitative-quantitative, exploratory, and descriptive research, based on a case study of 22 collections of digital museums. The semi-automatic evaluation was performed using the Python language, using regular expressions, and the results revealed cataloging problems related to the physical characteristics of the object, chronological information, geographic location, and description. Finally, the developed application can be used by different users to evaluate the quality of data in their databases and direct efforts towards preventive and corrective actions.

KEYWORDS: Data quality. Metadata. Cataloging. Cultural collections. Automation in document processing. Museology. Digital collections.

LISTA DE FIGURAS

Figura 1 – Etapas do modelo Cascata 69
Figura 2 – Diagrama de processos da aplicação 70
Figura 3 – Regras de catalogação CCO alinhadas ao INBCM.....73
Figura 4 – Diagnóstico da adequação dos metadados dos museus Ibram às dimensões CCO 78
Figura 5 – Adequação de coleções Ibram ao uso de vocabulário controlado 80
Figura 6 – Página inicial da ferramenta.....83
Figura 7 – Interface de envio de base de dados para avaliação 84
Figura 8 – Tela com sinalização de sucesso para identificar *encoding* e delimitador85
Figura 9 – Tela com sinalização de erro para identificar *encoding* e delimitador 85
Figura 10 – Tela de alinhamento entre elementos discricionais da base do usuário com os elementos discricionais do CCO.....87
Figura 11 – Tela de alinhamento com indicação de alinhamento já existente 88
Figura 12 – Tela de espera enquanto os dados são avaliados 89
Figura 13 – Tela principal com taxa de adequação de coleção avaliada 90
Figura 14 – Regras indicadas para elementos discricionais que não alcançaram 100% de adequação.....91
Figura 15 – Opção de baixar relatório completo em Excel.....91

LISTA DE QUADROS

Quadro 1 – Elementos recomendados pelo guia CCO 29
Quadro 2 – Regras de catalogação mapeadas do guia CCO 33
Quadro 3 – Tecnologias para extração automática de metadados 40
Quadro 4 – Quadro sinóptico sobre projetos de avaliação de qualidade de dados no domínio cultural 53
Quadro 5 – Quantidade de resultados por serviço de busca 57
Quadro 6 – Critérios de inclusão dos artigos.....58
Quadro 7 – Total de trabalhos recuperados por fonte 60
Quadro 8 – Passos de inclusão dos artigos selecionados 60
Quadro 9 – Elementos INBCM para identificação do item de caráter museológico ..63
Quadro 10 – Quantidade de Itens por coleção dos museus sob gestão do Ibram 64
Quadro 11 – Alinhamento entre elementos descritivos – INBCM e CCO.....72
Quadro 12 – Regras de catalogação e *regex* utilizados na pesquisa.....74

SUMÁRIO

1 INTRODUÇÃO 10

1.1 Situação problemática, questão e hipótese de pesquisa 14

1.2 Objetivos 16

2 FUNDAMENTOS TEÓRICO-METODOLÓGICOS 17

2.1 Catalogação e Metadados 18

2.2 Instrumentos de representação da informação 21

2.3 Padrões de Documentação para Tratamento Documental 23

2.4 Automação em Processos de Tratamento Documental 39

2.5 Expressão Regular para Tratamento Documental 43

3 ESTUDO DO ESTADO DA ARTE: MODELOS DE DIAGNÓSTICO DE QUALIDADE DE DADOS NO DOMÍNIO DO PATRIMONIO CULTURAL 45

4 METODOLOGIA DA PESQUISA 55

4.1 Fundamentos da pesquisa 55

4.2 Estudo de Caso: Instituto Brasileiro de Museus 61

4.3 Etapas do processo de avaliação diagnóstica 62

4.3.1 Alinhamento entre elementos de descrição: INBCM e CCO 65

4.3.2 Exploração semiautomática das bases de dados de coleções 67

4.4 Modelo Cascata de desenvolvimento de software 68

5 RESULTADOS 71

5.1 Alinhamento de padrões 72

5.2 Índice de adequação das coleções do Ibram 75

5.3 Índice de adequação das coleções do Ibram frente ao uso de vocabulários controlados 79

5.4 *DataQ Culture:* ferramenta de avaliação de qualidade de dados 82

6 DISCUSSÃO DOS RESULTADOS 92

7 CONCLUSÕES E TRABALHOS FUTUROS 97

REFERÊNCIAS.....100

1 INTRODUÇÃO

Nos últimos anos, houve uma crescente adesão das instituições de patrimônio cultural à digitalização e disponibilização de seus dados de acervos na internet, proporcionando maior acesso e democratização do conhecimento científico e cultural à sociedade. Com o advento das novas tecnologias da informação, esse aumento se tornou ainda mais significativo na produção e intercâmbio de registros em diversas áreas do conhecimento. Diante desse fato, a representação de dados no contexto da web é um tema de grande relevância atual a se explorar, e considerações acerca da importância da qualidade para publicação de conjuntos de dados abertos na internet também surgiram nas últimas décadas em contextos diversos (BIZER; HEATH; BERNERS-LEE, 2009; WILKINSON et al., 2016; GUIZZARDI, 2020; SIQUEIRA et al., 2021; MACEDO; LEMOS, 2021; MARTINS et al., 2022).

Apesar dessas iniciativas, investir somente na digitalização de objetos culturais não é suficiente (MARTINS et al., 2022), 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. 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 à desconfiança quanto à veracidade ou podem ainda levar à 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 (WANG, 2018; ŠLIBAR; OREŠKI; BEGIČEVIĆ REĐEP, 2021). 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) (VIRKUS; GAROUFALLOU; 2020; MARTINS et al., 2022).

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 é necessariamente dados isentos de erros, sendo esta, apenas uma das dimensões. Deve-se, portanto, considerar outras dimensões, tais como a precisão, a integridade e a consistência dos dados, que, por fim, corroborarão com a medida da adequação dos dados a um propósito específico (ECKERSON, 2002). 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).

Eckerson (2002), no relatório publicado pelo The Data Warehouse Institute, indica 7 (sete) dimensões que caracterizam a qualidade dos dados, a saber:
Precisão: pondera sobre a representação da realidade por meio dos dados ou se estes são provenientes de uma fonte confiável;

- **Integridade**: define se os relacionamentos desses dados com a estrutura são mantidos de forma sólida;
- **Consistência**: define se os entendimentos dos elementos são definidos de forma consistente;
- **Completude**: define se todos os dados necessários estão presentes;
- **Validade**: define se os valores dos dados estão dentro dos critérios aceitáveis definidos pelo negócio;
- **Pontualidade**: define se os dados estarão presentes em tempo hábil quando necessários; e por fim

- **Acessibilidade**: define se os dados estão acessíveis, de fácil acesso e se são compreensíveis.

Em particular, dimensões como precisão, integridade, pontualidade e consistência têm sido amplamente citadas na literatura como algumas das dimensões de qualidade mais importantes para os usuários de informação (WANG; STRONG, 1996; BATINI; SCANNAPIECA, 2006 p. 38; WANG, 2018). Correção, confiabilidade e usabilidade são dimensões interessantes em áreas como processo de modelagem de modelos preditivos[1] (KENETT; REDMAN, 2019, p. 25).

Os efeitos da má qualidade dos dados são vivenciados em diversas ocasiões, todavia, na maior parte das vezes não se faz a conexão do efeito com a sua causa (BATINI; SCANNAPIECA, 2006). O relatório da IBM (2014) apontou que dados com baixa qualidade custam só a empresas dos Estados Unidos da América, anualmente, 3,1 trilhões de dólares ao ano.

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 (ECKERSON, 2002). É melhor prevenir erros do que corrigi-los posteriormente (ENGLISH, 1999, p. 282), o que é de longe a opção mais barata (CHAPMAN, 2005). Como exemplo do custo de correção, pode-se citar o "problema do ano 2000", que levou à modificação de aplicativos de *software* e bancos de dados que utilizavam um campo de dois dígitos para representar anos. Foi um problema de qualidade de dados e, portanto, os custos para modificar tais aplicativos de *software* e bancos de dados foram estimados em cerca de 1,5 trilhão de dólares americanos (BATINI; SCANNAPIECA, 2006, p. 5; ENGLISH, 1999).

Além do eventual custo de reparação, a falta de qualidade em dados tem como efeito ainda a interoperabilidade reduzida, pois não deixa claro quais outros *datasets* um determinado *dataset* está relacionado ou pode ser vinculado (MACEDO; LEMOS, 2021). Por consequência, a capacidade de descoberta desses dados se torna limitada, dificultando os usuários a encontrar conjuntos de dados relevantes apenas consultando o portal de metadados, sem examinar os conjuntos de dados reais, e, por fim, menor satisfação do cliente, devido à dificuldade em encontrar dados.

[1] Modelo usado para evidenciar padrões capazes de apontar as próximas tendências.

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. 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. 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).

Segundo Wang (2018), o controle da qualidade de dados é um dos papéis mais importantes do profissional da informação para com a Ciência de Dados, cujo principal propósito deste campo é identificar e extrair conhecimento a partir de grandes volumes de dados existentes em diversos tipos de fontes disponíveis em ambientes digitais, o que promove, inclusive, a abertura de espaços para o crescimento de disciplinas emergentes como as Humanidade Digitais (LIU, 2012; KOLTAY, 2016; POOLE, 2017; CLEMENT; CARTER, 2017), que buscam a extração de valor em objetos digitais disponíveis na rede por meio de ações de reúso para o desenvolvimento econômico, social e humano, a exemplo da indústria criativa, ciência aberta, turismo, educação, entre outros.

No caso do Instituto Brasileiro de Museus (Ibram), objeto de estudo 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 padrão de dados adotado internamente pela instituição - o modelo do Inventário Nacional de Bens Culturais Musealizados – INBCM (BRASIL, 2021). O Ibram faz a gestão do patrimônio museológico a nível federal de 23 coleções digitais de 22 museus espalhados pelo país por meio do *software* Tainacan[2], uma ferramenta de organização e gestão de acervos digitais (MARTINS; LEMOS; ANDRADE, 2021).

[2] https://tainacan.org

1.1 Situação problemática, questão e hipótese de pesquisa

As bases de dados desses museus foram modeladas no Tainacan a partir do INBCM. Contudo, o INBCM não pode ser considerado um guia de catalogação por não almejar, primariamente, requisitos descritivos únicos e singulares, vocabulários padronizados, indexação, localização, acesso e navegação em sistemas de recuperação da informação (SRIs) (WINAR, 1985; MEY, 1995; INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2016; LANCASTER, 1986; 2004).

O INBCM não detalha e nem orienta, por exemplo, os aspectos sintáticos e semânticos para os elementos de descrição sugeridos, deixando em aberto a forma como os instrumentos de organização da informação (modelos conceituais, padrões de metadados, linguagens documentárias e regras de catalogação) deveriam ser implementados pela instituição que aderiu aos mesmos.

Adicionalmente, cabe destacar, para um possível cenário de agregação de dados, que o preenchimento de campos ou elementos de metadados deveria ser seguido com rigor a partir de uma política de catalogação da instituição, de modo a atender os elementos mínimos obrigatórios orientados para um registro museal coerente e consistente a ser mapeado para o modelo de metadados de referência adotado pelo agregador.

Logo, a ausência de padrões e práticas de catalogação pode comprometer a identificação de informações cruciais e necessárias para descrever um item de modo a localizá-lo no acervo para fins de busca e recuperação (WYNAR, 1985; MEY, 1995; INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2016). A utilização de regras de catalogação pelas instituições como museus é essencial, uma vez que orientam os formatos e os valores adequados de preenchimento (GILLILAND, 2016) acerca dos elementos de metadados constitutivos de suas bases de dados, que podem, inclusive, serem utilizados como possíveis índices numa interface de busca e navegação (ABADAL; CODINA, 2005).

Dentre os padrões de catalogação recomendados no campo do patrimônio cultural, incluindo modelos de dados, modelos ontológicos, padrões de metadados, guias de gestão de acervos e guias de catalogação (SPECTRUM, FRBROO, CIDOC-CRM, CCO, CDWA, DUBLIN CORE, OBJECT ID, entre outros), a pesquisa destaca o guia *Cataloging Cultural Objects* (CCO).

O CCO surgiu a partir do resultado do consenso de profissionais das comunidades GLAM que pesquisam a prática comum entre essas disciplinas (BACA *et al.*, 2006, HARPRING, 2022), fornecendo diretrizes para selecionar, ordenar e formatar dados usados para preencher registros de catálogo. O guia apresenta ainda conceitos genéricos, permitindo seu uso com outros conjuntos de metadados, como, por exemplo, o *Machine-Readable Cataloging* (MARC)[3], o *Metadata Object Description Schema* (MODS)[4], o *Dublin Core*[5], e os próprios elementos do INBCM, conforme poderá ser visto adiante em procedimentos metodológicos.

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 apresentam, a presente pesquisa busca responder a seguinte questão: *como melhorar a qualidade de dados em acervos culturais?*

A partir do estudo do estado da arte (Capítulo 3) sobre modelos de diagnóstico de qualidade de dados no domínio do patrimônio cultural, foi observada na última década (2012 a 2023) a falta de alinhamento com modelos de referência para levantamento de regras para avaliação da conformidade dos dados de instituições do domínio cultural. Percebeu-se também alguma deficiência no processo em termos de volume de dados a ser verificado, com uso, em alguns casos, de avaliação por amostragem. Com a utilização de um guia de referência ou padrão de dados, considerando a grande quantidade de bases de dados legadas que necessitaria de avaliação prévia para conseguir ser endereçada para ambiente digital, torna-se pertinente a semiautomatização do processo repetitivo de avaliação da qualidade de dados, auxiliando o especialista em sua tarefa, economizando recursos e direcionando seus esforços para tomada de decisão em questões mais complexas e efetivas.

Assim sendo, a presente pesquisa parte da hipótese de 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. Nesse sentido, a utilização de um guia de catalogação de objetos culturais, como o CCO, é uma

[3]https://www.loc.gov/marc/. Acesso em 12/03/2023

[4]http://www.loc.gov/standards/mods/. Acesso em: 12/03/2023.

[5]https://www.dublincore.org. Acesso em 12/03/2023.

referência importante para o estabelecimento de critérios e padrões de qualidade de dados para o domínio. Apesar dos esforços desenvolvidos pelo Ibram na qualificação dos dados e disponibilização em ambiente digital, esses dados não estão de acordo com as recomendações indicadas pelo CCO, comprometendo: i) o uso desses recursos informacionais em ambientes de dados abertos ligados; ii) a interoperabilidade semântica e sintática; e iii) a recuperação eficiente das informações pelos usuários humanos e agentes computacionais.

1.2 Objetivos

O objetivo geral desta pesquisa é desenvolver um modelo de avaliação semiautomática de qualidade de dados que possibilite o diagnóstico nos dados de acervos de instituições culturais, tendo o Ibram como objeto-focal, buscando melhorar a qualidade das fontes documentais ora envolvidas.

Como objetivos específicos almejam-se:

- Explorar como a avaliação da qualidade de dados vem sendo realizada em acervos de instituições do patrimônio cultural.
- Implementar o modelo de avaliação de qualidade de dados nos acervos digitais das instituições culturais vinculadas ao Ibram.
- Apresentar um panorama-situacional dos dados das coleções do Ibram em relação a padrões de qualidade à luz de princípios teórico-metodológicos da Catalogação no âmbito da Ciência da Informação, incluindo incompletude, inconsistência de dados e uso de vocabulário controlado.

Desta forma, com o objetivo de preencher lacunas observadas na literatura, justifica-se a realização desta pesquisa. Acredita-se que a implementação de uma avaliação semiautomática de qualidade de dados poderia aprimorar a indexação, a busca e navegação nos sistemas de recuperação de informações 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, possa ser alinhada e avaliada de acordo com as regras de catalogação CCO. A padronização dos dados em coleções digitais poderia facilitar comparações entre diferentes instituições, impulsionando pesquisas acadêmicas e

científicas e proporcionando uma compreensão mais profunda da história e cultura do país.

Esta dissertação está organizada da seguinte maneira:

- O Capítulo 1 apresenta o contexto da pesquisa, a situação problemática, sua questão e hipótese; além dos objetivos (geral e específicos), justificativa e contribuições para a ciência e a sociedade.
- O Capítulo 2 disserta sobre a fundamentação teórico-metodológica, tratando de conceitos e teorias importantes ao desenvolvimento da pesquisa.
- O Capítulo 3 discorre sobre o estado da arte envolvendo modelos de diagnóstico de qualidade de dados em acervos culturais.
- O Capítulo 4 apresenta o estudo de caso e os procedimentos metodológicos empregados na implementação do modelo para a avaliação da qualidade de dados nas bases de dados dos museus sob gestão do Ibram.
- O Capítulo 5 apresenta os resultados associados à aplicação do modelo de diagnóstico de qualidade de dados, e, como produto da dissertação, a aplicação desenvolvida permitindo a reprodução dos resultados em diferentes acervos culturais.
- O Capítulo 6 compreende as discussões relacionadas a todos os resultados alcançados a partir da condução da pesquisa.
- Por fim, o Capítulo 7 conclui a dissertação tecendo as considerações finais e expondo as limitações, contribuições e indicações de trabalhos futuros.

2 FUNDAMENTOS TEÓRICO-METODOLÓGICOS

O presente capítulo constitui-se de uma abordagem teórica conceitual fundamental para embasar os procedimentos metodológicos e o desenvolvimento dos produtos da pesquisa. Assim, visando a um encadeamento salutar no percurso teórico metodológico adotado, o Capítulo foi organizado em três seções, conforme apresentado a seguir. A Seção 2.1 trata da catalogação, metadados e representação da informação, apontando assuntos seminais aos campos da Ciência da Informação e da Ciência da Computação sobre desafios inerentes à catalogação e representação da informação. A Seção 2.2 explana sobre padrões de documentação para o tratamento documental, discorrendo e apresentando alguns dos padrões existentes e

recomendados, incluindo o padrão usado no desenvolvimento da pesquisa, o CCO. E a Seção 2.3 discorre sobre extração e criação de metadados de forma automática e semiautomática, incluindo a Subseção 2.3.1 abordando sobre a utilização de expressões regulares para o tratamento documental, sendo uma das técnicas importantes usadas na pesquisa.

2.1 Catalogação e Metadados

Padrões de documentação são um conjunto de regras e diretrizes que devem ser seguidas ao criar e manter documentos. Eles podem incluir orientações sobre o formato, o conteúdo e a estrutura de um documento, além de como este deve ser armazenado e gerenciado. Como subconjunto de padrões de documentação temos a catalogação (MEY, 1995; JOUDREY; TAYLOR; MILLER, 2015).

A catalogação, segundo Mey (1995), trata-se da representação de um objeto informacional, consistindo no levantamento de características deste, de maneira a individualizá-lo, tornando-o distinto dos demais, e também criando vínculos entre objetos que compartilham determinadas características. Desta forma, permite-se que o usuário veja, a partir de um objeto, as alternativas que estão presentes para seu uso e permita sua localização no acervo. Para o cumprimento dessa tarefa, a catalogação precisa apresentar algumas qualidades, tais como integridade, clareza, precisão, lógica e consistência.

A necessidade de catalogar uma coleção surge à medida que a recuperação de um desses objetos se torna complexa devido ao volume de objetos no repositório. Além disso, um catálogo deve ser construído de forma que todos os metadados possam ser encontrados de forma rápida e fácil. Assim, a catalogação auxilia tanto na recuperação da informação, sendo instrumento eficaz e efetivo para o usuário em um SRI, quanto para o controle e gerenciamento da coleção (JOUDREY; TAYLOR; MILLER, 2015; INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2016).

Barbieri (2001) esclarece que os metadados surgiram impulsionados pela tecnologia de banco de dados na década de 80, além da ampliação de uso com a chegada dos sistemas de apoio à decisão (SADs). Entretanto, outras comunidades que trabalham com informação como bibliotecários, arquivistas e museólogos sempre produziram metadados para criar registros sobre informação (ABBAS, 2010). Assim,

considera-se que o registro de metadados é uma prática antiga, principalmente na área de Biblioteconomia, uma vez que o processo de catalogação e indexação sempre foi realizado no intuito de organizar, descrever e melhorar o acesso aos recursos de informação.

Os recursos de informação, nesse caso, independentemente da forma física ou abstrata que assumem, possuem três características que podem ser representadas por meio de metadados, a saber: conteúdo, contexto e estrutura (GILLILAND, 2016). O conteúdo refere-se ao que o recurso contém ou sobre a natureza ontológica do mesmo; o contexto trata dos elementos ligados à criação do recurso de informação, quais sejam, quem, o quê, porquê, onde e como; e a estrutura refere-se ao conjunto formal de associações dentro e/ou entre recursos de informação.

O investimento na construção qualificada de metadados para a representação consistente de recursos de informação torna-se útil para: (i) a organização de coleções de acervos;(ii) a comunicação entre pessoas que estabelecem relações com esses recursos; (iii) a comunicação entre sistemas de informação; e (iv) a comunicação entre instituição, pessoas e sistemas de informação (GILLILAND, 2016). Isso nos remete ao conceito de interoperabilidade, seminal e imprescindível de entendimento às estratégias de agregação de museus por parte do Ibram. De acordo com a NISO (2004), interoperabilidade é a capacidade de diversos sistemas trocarem dados entre si (interoperar) com perda mínima de conteúdo e funcionalidade, de modo a garantir que pessoas, organizações e sistemas de informação (tal como os repositórios digitais) interajam de forma satisfatória.

Nesta dinâmica aparece o paradigma dados abertos ligados (do inglês, *Linked Open Data* - LOD), com grande atuação do *World Wide Web Consortium* (W3C) (W3C LINKED DATA, 2022), cuja fundamentação subjaz aos preceitos da Web semântica (BIZER; HEATH; BERNERS-LEE, 2009), e tem como propósito identificar formalmente conceitos e relações entre conceitos em documentos, 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, acrônimo para quatro princípios fundamentais: *Findability* (respectivo a possibilidade de ser encontrado), *Accessibility* (acessibilidade), *Interoperability* (interoperabilidade) e *Reusability* (reutilização). Estes princípios servem para guiar produtores e editores de dados, ajudando a maximizar o valor adquirido pela

disponibilização de dados digitais com o uso de códigos de catalogação contemporâneos, a exemplo do RDA[6], que recomenda seus modelos de dados em linguagem RDF para interligação de objetos digitais na web focada no usuário (SILVA; SOUZA, 2014; WILKINSON et al., 2016; HENDLER; GANDON; ALLEMANG, 2020; LEMOS; SOUZA, 2020).

Logo, com objetivo de viabilizar a interoperabilidade, instituições vêm investindo na adoção de padrões e práticas recomendadas para a produção de metadados, sendo, portanto, uma tentativa de se obter um vocabulário comum e consistente para descrever uma variedade de estruturas de dados capazes de satisfazer a várias comunidades. Nesse sentido, Boughida (2005) e Gilliland (2016) promovem orientações acerca do uso de padrões para tratamento nos dados focando padronização, normalização, qualidade e intercâmbio de metadados em ambiente digital, conforme elucidado nos exemplos abaixo:

- Padrões para estrutura de dados: conjunto de elementos de metadados ou esquemas de categorias que formam um registro de informação (como exemplo: MARC; Dublin Core; VRA Core; LIDO; MPEG-7).
- Padrões para valores dos dados: linguagens documentárias, vocabulários controlados, arquivos de autoridade e ontologias de domínio usados para preencher os dados nos elementos de metadados (como exemplo: *Library of Congress Authority Files*; *Union List of Artist Names* – Getty ULAN; *Art & Architecture Thesaurus* - *Getty* AAT; *Media Ontology*; CIDOC/CRM).
- Padrões para conteúdo dos dados: regras e códigos de catalogação que orientam em formatações, sintaxes e relacionamentos para os valores de dados usados para preencher os elementos de metadados (como exemplo: *Cataloging Cultural Objects* - CCO;*Descriptive Cataloging of Rare Materials* - DCRM; *Anglo-American Cataloguing Rule* - AACR2.
- Padrões para comunicação de dados: padrões de metadados expressados em uma linguagem de representação legível para a máquina (exs.: MARC21; *Dublin Core* RDF/XML; LIDO XML *Schema*; *VRA Core* 4.0 XML).

[6] https://www.rdatoolkit.org. Acesso em 12/03/2023.

2.2 Instrumentos de representação da informação

Linguagens documentárias compreendem a comunicação entre usuários e SRIs, sendo consideradas linguagens artificialmente construídas a partir da linguagem natural presente nos documentos, buscando-se obter um vocabulário controlado de um assunto específico (DODEBEI, 2002; LANCASTER, 2004). Já vocabulário controlado é definido por Lancaster (2004) como uma lista de termos autorizados, em que o indexador somente pode atribuir a um documento termos que existem na lista adotada pela unidade de informação em que trabalha. O vocabulário controlado tem por funções: i) o controle de sinônimos, através da definição de um termo padrão, com remissivas para os sinônimos; ii) a diferenciação entre os homógrafos; e iii) o agrupamento de termos em que os significados apresentem uma relação mais estreita entre si, como por exemplo, relações hierárquicas ou não hierárquicas.

A ANSI (2005) enfatiza que o principal propósito dos vocabulários controlados é fornecer um significado para a organização da informação, incluindo tradução, consistência, relacionamentos, visualização e recuperação perante a informação.

Dentro do universo dos Sistemas de Organização do Conhecimento (SOCs) (HJØRLAND, 2007), consideram-se também as linguagens documentárias como resultantes da análise semântica de conceitos e relacionamentos de um domínio que busca, a priori, a organização de recursos de informação nos aspectos de: i) associação, gerando relacionamentos; ii) representação, gerando pontos de acesso e índices em processos de catalogação e indexação; iii) classificação, promovendo colocação e ordenação para os documentos; e iv) categorização, gerando esquemas de categorias. Logo, podem-se citar alguns SOCs alinhados com o conceito de linguagens documentárias ou vocabulários controlados usados em ambiência digital, porém numa perspectiva de caracterização por níveis de complexidade estrutural, isto é, pelo formalismo promovido pela adequabilidade no uso de tecnologias semânticas, especialmente no que se refere aos relacionamentos (HJØRLAND, 2007), a saber: taxonomias, tesauros e ontologias.

Taxonomias, por exemplo, são coleções de termos classificados em uma estrutura hierárquica, na qual se emprega relacionamentos de generalização e especialização. O Ibram utiliza o software Tainacan, que possui o recurso de produção de taxonomias para normalização de valores de seus tipos de metadados (MARTINS; LEMOS; ANDRADE, 2021). Já nos tesauros, as relações semânticas se estendem em

equivalência, hierárquicas e associativas entre os termos, em que são claramente mostradas e identificadas através de indicadores de relação padrão (ANSI, 2005). Ontologias possuem os mesmos princípios dos vocabulários controlados (SILVA; SOUZA; ALMEIDA, 2008), isto é, trabalham com linguagem natural e fazem a delimitação de termos e de relações. Entretanto, a semântica envolvida na terminologia da ontologia se difere dos vocabulários controlados (dentro do contexto das linguagens documentárias) por incluir axiomas formais (através de declarações lógicas) que restringem a utilização do vocabulário. Outra distinção está no fato de as especificações de relações no contexto das ontologias serem em número superior às dos tesauros, por exemplo.

Os vocabulários controlados vistos como SOCs viabilizam, portanto, a recuperação de informação nos mais diversos ambientes de informação, a exemplo dos museus digitais, padronizando a entrada de dados, facilitando a estratégia de busca e, consequentemente, melhorando a interação do usuário com o SRI.

Para tal, Svenonius (2000) acrescenta que para a informação ser organizada no âmbito de um SRI, precisa ser descrita e que, o produto desse processo descritivo é a representação da informação. Ressalta que alguns tipos de representação da informação são construídos através do uso de linguagens, as quais são subdivididas em linguagens que descrevem a informação (o conteúdo) e linguagens que descrevem o documento (o suporte), no todo ou em partes. A linguagem usada para descrever o documento está relacionada à representação descritiva, também tratada como processo de catalogação, que busca retratar aspectos específicos do documento que permitam a sua individualização e determinação dos pontos de acesso para proporcionar aos usuários a condição de encontrar, identificar, selecionar e obter o item por meio de um catálogo (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2016). Já a linguagem destinada à descrição do conteúdo do documento está relacionada à representação temática que vislumbra aspectos intelectuais e semânticos, portanto, subjetivos, como a compreensão do assunto do documento para fins de tradução para uma linguagem documentária (LANCASTER, 2004; ANSI, 2005). Para os propósitos desta pesquisa, ambas as representações são consideradas na avaliação da qualidade dos dados das coleções ora envolvidas, sendo o emprego de vocabulário controlado avaliado tanto na descrição da Obra em seus aspectos de suporte quanto na descrição de seu conteúdo.

Metadados, portanto, quando produzidos a partir de padrões de documentação, incluindo regras de catalogação, padrões de metadados e linguagens documentárias, tornam-se elementos centrais na produção de bases de dados com qualidade, sendo estas, produtos de informação considerados cruciais para a sociedade quando adotados para realizar a mediação entre documentos e comunidade de usuários interessada na preservação, busca, recuperação, acesso e reúso, pois referenciam e divulgam o conhecimento a partir do uso qualificado da informação.

Instituições de patrimônio cultural (CHARLES *et al.,* 2017; FINK, 2018; DIJKSHOORN *et al*., 2018; HARPRING, 2022) geralmente aderem a padrões de documentação que produzem descrições de coleções de objetos culturais digitalizados ou nato digitais, os quais necessitam padronizar aspectos únicos de coleções culturais tanto fisicamente quanto digitalmente, bem como fornecer dados administrativos para descrever a digitalização, os direitos autorais e as disposições de uso dos objetos. Nesse sentido, Abbas (2010) destaca dois desses padrões para tratamento documental que vislumbram a produção de bases de dados no âmbito do patrimônio cultural, especialmente para museus, a saber: i) o padrão semântico *Categories for the Descriptionof Works of Art* (CDWA) e sua extensão *Cataloging of Cultural Objects* (CCO); e ii) o padrão de metadados *Visual Resources Association Core Categories* (VRA Core). A próxima seção se incube de elucidar tais padrões, dando destaque ao padrão de referência usado na presente pesquisa, o CCO.

2.3 Padrões de Documentação para Tratamento Documental

O CDWA compreende um conjunto de diretrizes para as melhores práticas na catalogação e descrição de obras de arte, arquitetura, outros materiais culturais, grupos e coleções de obras e imagens relacionadas. O CDWA não é um modelo de dados em si, mas é organizado em uma estrutura conceitual que pode ser usada para projetar modelos de dados e bancos de dados para acessar informações, como é o caso do CONA[7]. O CDWA inclui ainda cerca de 540 categorias e subcategorias de informações. Um pequeno subconjunto de categorias é considerado central, pois representa as informações mínimas necessárias para identificar e descrever uma obra (TRUST; COLLEGE ART ASSOCIATION, 2022).

[7]https://www.getty.edu/research/tools/vocabularies/cona/index.html

Baseado nos elementos recomendados pelo CDWA, surgem os padrões de metadados para descrever registros para obras de arte e do domínio da cultura material, o *CDWA Lite* e o *VRA* Core. O CDWA *Lite*, como o nome sugere, surge com base nos elementos de dados e diretrizes contidas no CDWA e no guia *Cataloging cultural objects* (CCO). Em 2010, o esquema *CDWA Lite* foi ampliado e integrado ao esquema *Lightweight Information Describe Objects* (LIDO), disponível no site do CIDOC (Comitê Internacional de Documentação do Conselho Internacional de Museus – ICOM)[8]

O VRA Core é um conjunto de elementos de dados utilizado para descrever obras de arte e outros objetos culturais. Embora baseado no CDWA, o VRA Core surgiu da necessidade de uma proposta mais satisfatória para a descrição de imagens e, especialmente, para cobrir todos os elementos necessários para a descrição da arquitetura e outras obras específicas (WHITESIDE, 1999). Desde sua primeira versão em 1996, o *VRA Core* passou por atualizações e está atualmente na versão 4. Esse padrão é capaz de documentar objetos e seus desdobramentos digitais, bem como eventos, pinturas, desenhos, esculturas, arquitetura, fotografias, obras de arte, decoração e performances. O conjunto de elementos do padrão fornece uma organização categórica para a descrição dessas obras (VISUAL RESOURCES ASSOCIATION, 2021).

Apesar de haver vários padrões de metadados, além dos citados acima, os catalogadores mesmo sabendo o local de inserção de informações catalográficas específicas e com quais termos descreveriam obras ou imagens, à luz de um conjunto de metadados, diretrizes, padrões e guias, não sabiam como formatar os termos para inseri-los nos seus respectivos campos. Questionamentos como “deveria o nome de um determinado autor vir antes ou depois do sobrenome? o título descritivo de uma determinada obra deve estar em caixa alta?” ainda traziam inconsistência no conteúdo dos dados e diferença na representação dos metadados de uma obra ou imagem (VISUAL RESOURCES ASSOCIATION, 2021).

A comunidade do patrimônio cultural carecia de diretrizes publicadas que suprissem os requisitos descritivos únicos e singulares de objetos culturais, com foco principal em evitar a catalogação inadequada e inconsistente, que ocorria devido ao processo decisório dos catalogadores estar fundamentado unicamente em sua

[8]https://cidoc-crm.org

percepção de como as obras são definidas (COBURN et al., 2010). Logo, em 1999, quando um grupo de trabalho do Council On *Library and Information Resources* (CLIR) começou a implementar o novo conjunto de elementos *VRA Core* para catalogar informações visuais, percebeu-se a ausência de um componente para responder a esse tipo de dúvida e padronizar a entrada de dados na catalogação. Assim, viu-se a necessidade da existência de um guia para o conteúdo e a formatação de dados (VISUAL RESOURCES ASSOCIATION, 2021). Surgia então o CCO.

O CCO 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) e fornece diretrizes para selecionar, ordenar e formatar dados usados para preencher registros de catálogo com base em categorias principais em CDWA e VRA Core. Porém, embora tenha sido inspirado no desenvolvimento dos elementos VRA Core e do *Getty Vocabularies*, o CCO apresenta conceitos mais genéricos que podem ser utilizados com outros conjuntos de metadados, como, por exemplo, o *Machine-Readable Cataloging* (MARC), *Metadata Object Description Schema* (MODS), *Dublin Core*, etc. O guia de catalogação CCO apresenta três partes principais, a saber: (i) introdução e instruções gerais; (ii) elementos de metadados; e (iii) autoridades.

Na parte introdutória (parte I) do guia, é estabelecido o objetivo principal deste, declarando que foi desenvolvido para os profissionais que documentam objetos culturais e suas imagens, com foco em metadados descritivos e controle de autoridade. Foi informado ainda que o guia abrange muitos tipos de obras culturais, incluindo arquitetura, pinturas, esculturas, gravuras, manuscritos, fotografias e outras mídias visuais, arte performática, sítios e artefatos arqueológicos e vários objetos funcionais do reino da cultura material. Acrescenta ainda que o guia se destina a ajudar a instituição a planejar, a implementar e a usar bancos de dados e regras de catalogação locais. O CCO tem como objetivo ser uma referência durante a catalogação, sem a necessidade de ser lido completamente. São elencados ainda, nesta parte do guia, 10 (dez) princípios importantes, a saber:

1. Os Registros de obras devem ter um foco lógico;
2. Inclua todos os elementos CCO necessários;
3. Siga as regras do CCO;

4. Use vocabulários controlados;
5. Crie autoridades locais;
6. Use padrões de metadados estabelecidos;
7. Compreenda a diferença e a relação entre catalogação, classificação, indexação e exibição da informação;
8. Seja consistente no estabelecimento de relações entre obras e imagens, entre um grupo ou coleção e obras, entre obras e entre imagens;
9. Seja consistente no uso de letras maiúsculas e minúsculas, pontuação e sintaxe;
10. Use valores de dados do idioma da catalogação sempre que possível.

Ademais, foi realizada uma distinção importante a respeito da diferença entre obra e imagem. É importante deixar essa distinção clara no momento da catalogação, pois uma obra é uma criação intelectual ou artística distinta, limitada principalmente a objetos e estruturas feitas por humanos, incluindo obras construídas, obras de arte visual e artefatos culturais. As imagens são representações visuais de obras, geralmente em formato de slide, fotografia ou digital. Modelos físicos tridimensionais, desenhos, pinturas ou esculturas não são imagens por se tratarem de obras em si.

Em seguida, na parte (ii) são elencados os elementos de metadados discutidos no guia, incluindo 9 (nove) capítulos, cada um considerando um conjunto específico de elementos discricionais recomendados. Os grupos de elementos discricionais apresentados na parte (ii) o guia são:

1. *Object Naming*, fornece maneiras de se referir a uma obra, seção em que se define o que está sendo catalogado.
2. *Creator Information*, seção específica para identificar o criador de uma obra, este pode ser uma pessoa, conhecida pelo nome ou anônima. Vários criadores podem ser responsáveis por projetar e fazer uma obra. Um criador também pode ser uma pessoa jurídica.
3. *Physical Characteristics*, esta seção descreve a aparência de uma obra e apresenta características de sua forma física.
4. *Stylistic, Cultural, and Chronological Information*, esta refere-se às características estilísticas de uma obra, origens culturais e data de design ou criação.

5. *Location and Geography* trata de elementos que registram informações geográficas. Pelo menos quatro perguntas sobre localização são geralmente de interesse ao descrever um objeto cultural ou obra: Onde está agora? Onde estava antes? Onde foi feito? Onde foi descoberto?
6. *Subject*, assunto deve ser registrado para todas as obras e imagens. Este elemento contém uma identificação, descrição ou interpretação do que é representado por uma obra ou imagem. Os assuntos incluem coisas, lugares, atividades, formas abstratas, decorações, histórias e eventos da literatura, mitologia, religião ou história.
7. *Class*, é usado para relacionar uma obra específica a outras obras com características semelhantes, muitas vezes com base no esquema organizacional de um determinado repositório ou coleção. O objetivo é colocar a obra em um contexto mais amplo, categorizando-a com base em características semelhantes. Os termos de classe podem representar uma hierarquia, uma tipologia ou algum outro agrupamento de itens, implicando semelhanças entre os trabalhos dentro da lógica da classificação.
8. *Description*, pode ser associado a campos específicos em todo o registro. O elemento consiste em uma nota descritiva que geralmente é um texto ensaístico relativamente breve, detalhando o conteúdo e o contexto da obra.
9. *View Information*, incluem detalhes sobre a exibição da obra conforme ela aparece em uma imagem (substituta) da obra. É importante registrar informações sobre a visualização de uma imagem independentemente de seu formato ou tipo de instituição. Os substitutos visuais podem fornecer acesso a obras que, de outra forma, estariam indisponíveis devido a suas localizações remotas ou outras restrições que limitariam o contato direto.

Por fim, a parte (iii) fornece diretrizes para criar registros de autoridade, que são usados para registrar informações sobre entidades como pessoas, entidades corporativas, locais e assuntos em um formato padronizado. A saber:

1. Pessoas e organizações: os registros de autoridade para pessoas são usados para descrever indivíduos que são criadores ou sujeitos de materiais de patrimônio cultural. Esses registros geralmente incluem o nome da pessoa, nomes variantes, informações biográficas e relacionamentos com outras pessoas ou organizações. Os registros de autoridade para organizações são

usados para descrever instituições ou grupos que são criadores ou sujeitos de materiais de patrimônio cultural. Esses registros geralmente incluem o nome da organização, nomes variantes, histórico e relacionamentos com outras organizações ou pessoas.

2. Lugares: os registros de autoridade para lugares são usados para descrever localizações geográficas que são significativas para os materiais de patrimônio cultural. Esses registros geralmente incluem o nome do local, nomes variantes, coordenadas geográficas, história e relacionamentos com outros locais ou entidades.
3. Autoridades de conceito: registros de autoridade para conceitos são usados para descrever as ideias ou temas abstratos ou intangíveis que são expressos em materiais de patrimônio cultural. Esses registros geralmente incluem o nome do conceito, nomes variantes, definição, notas de escopo e relacionamentos com outros conceitos.
4. Autoridades de assunto: os registros de autoridade para assuntos são usados para descrever os tópicos ou assuntos que são abordados em materiais de patrimônio cultural. Esses registros geralmente incluem o nome do assunto, nomes variantes, definição, notas de escopo e relacionamentos com outros assuntos.

No geral, a terceira parte do CCO fornece uma estrutura para a criação de registros de autoridade que permitem aos usuários descrever de forma precisa e consistente em materiais de patrimônio cultural, aprimorando o acesso e a compreensão desses recursos.

Para além dos conjuntos de metadados, o guia também indica os campos mínimos recomendados, além de apontar se deveria ser utilizado vocabulário controlado e lista de autoridades, conforme apresentado no Quadro 1. Cada um desses campos possui regras específicas de uso e preenchimento, com vista a nortear o catalogador com regras claras para o preenchimento do conteúdo dos dados.

Quadro 1 – Elementos recomendados pelo guia CCO

Grupo	Elemento	Elemento obrigatório	Uso de lista controlada	Descrição
I	*Work Type*	Sim	Sim	Identificação do tipo da obra, que é o foco do registro do catálogo, como pintura, escultura, etc..
I	*Title*	Sim	Não	Título da obra. Pode ser fornecido pelo autor, um atribuído ou em outro idioma.
I	*Title Type*	Não	Sim	Especificação do tipo do título. Exemplos: título do repositório, título inscrito, título do criador, título descritivo.
I	*Language*	Não	Sim	Idioma do título.
I	*Source*	Não	Sim	Fonte do título.
II	*Creator*	Sim	Sim	Autor/Criador.
II	*Extent*	Não	Sim	Papel do autor na obra.
II	*Qualifier*	Não	Sim	Adição de detalhar sobre a autoria, como incertezas e inspirações.
II	*Creator Role*	Sim	Sim	Registra o papel ou atividade desempenhada pelo criador.
III	*Measurements*	Sim	Sim	Informação sobre as medidas, dimensões, tamanho ou escala da obra.
III	*Value*	Não	Não	Valor da medida.
III	*Unit*	Não	Não	Unidade da medida.
III	*Type*	Não	Não	Tipo da medida com altura, largura, etc.
III	*Extent*	Não	Sim	Refere-se a parte da obra medida.
III	*Qualifier*	Não	Sim	Natureza das dimensões da obra, como máximo, mínimo, aproximado, etc.
III	*Shape*	Não	Sim	Registra contorno ou forma da obra ou de parte da obra.
III	*Format*	Não	Sim	Inclui formatos técnicos da obra, como VHS, CD, etc. Pode ser igual ao Work Type.
III	*Scale*	Não	Sim	Expressão da relação entre o tamanho da representação de algo e a coisa representada.
III	*Materials and Techniques*	Sim	Sim	Inclui as substâncias ou materiais usados na criação de uma obra, bem como técnicas.
III	*Material*	Não	Não	Para materiais, registre a matéria, materiais ou substâncias usadas para criar uma obra.
III	*Material type*	Não	Não	Também conhecida como meio e suporte, especifica o tipo de material aplicado e em qual superfície.
III	*Technique*	Não	Não	Meio, método, processo ou técnica pela qual um material foi usado.

III	*Color*	Não	Sim	Observa a cor, tonalidade ou matiz do material do qual a obra é composta.
III	*Mark*	Não	Não	Identifica marcas d'água, carimbos e outras marcas inerentes ou aplicadas.
III	*Extent*	Não	Sim	Refere-se à parte específica de uma obra composta por um determinado material ou criada usando uma determinada técnica.
III	*Qualifier*	Não	Sim	Não apresentado pelo guia.
III	*Stateand Edition*	Não	Não	Referem-se principalmente a obras produzidas em séries.
III	*Edition*	Não	Não	Registra a edição, notação que identifica uma gravura específica ou outra obra no contexto de um número limitado de obras idênticas.
III	*Impression Number*	Não	Sim	Registra o número atribuído a um item específico em uma edição ou execução de produção específica.
III	*Edition Size*	Não	Sim	Número total de obras criadas em uma determinada execução de produção.
III	*Edition Number*	Não	Sim	Registra o termo para a edição específica à qual a obra pertence.
III	*State*	Não	Não	Registra uma indicação da relação da obra com outras etapas da mesma obra. Estado de impressão não numéricos.
III	*State Identification*	Não	Sim	Apresenta o número da edição da obra.
III	*Known States*	Não	Sim	Apresenta o número total de estados da obra.
III	*Source of State*	Não	Não	Breve referência ou outra fonte usada para identificar o estado da obra.
III	*Additional Physical Characteristics*	Não	Não	Transcreve quaisquer letras físicas, anotações, textos, marcações inseridos ou fixados à obra.
III	*Inscription*	Sim	Sim	Transcreve quaisquer letras físicas, anotações, textos, marcações inseridos ou fixados à obra.
III	*Inscription Type*	Não	Sim	Apresenta o tipo de inscrição apresentada na obra, como assinado, datado, etc.
III	*Inscription Location*	Não	Sim	Indicação do local da inscrição na obra.
III	*Inscription Author*	Não	Não	Autor da inscrição. Semelhante a Creator.
III	*Type face or Letter form*	Não	Não	Apresenta característica da inscrição, fonte, estilo da fonte, etc.
III	*Facture*	Não	Não	Apresenta métodos de construção utilizados ou as aplicações específicas de técnicas.

III	*Physical Description*	Não	Sim	Registra uma descrição da aparência de uma obra expressa em termos genéricos.
III	*Condition and Examination History*	Não	Sim	Registra uma descrição avaliando a condição física geral, características e integridade de uma obra.
III	*Conservation and Treatment History*	Não	Sim	Registra os procedimentos ou ações que um trabalho sofreu para repará-lo, conservá-lo ou estabilizá-lo.
IV	*Style*	Não	Sim	Identifica o estilo nomeado, definido, período histórico ou artístico, movimento, grupo ou escola cujas características estão representadas.
IV	*Qualifier*	Não	Sim	Refere ao estilo, período, grupo, movimento ou dinastia, diferenciando, assim, cada tipo.
IV	*Culture*	Não	Sim	Registra o nome da cultura, povo ou nacionalidade de origem da obra.
IV	*Date*	Sim	Não	Registra a data ou intervalo de datas associado à criação, design, etc. de obra ou de parte da obra.
IV	*Earliest Date*	Sim	Sim	Campo não exibido para o usuário; data mais antiga do intervalo da data.
IV	*Latest Date*	Sim	Sim	Campo não exibido para o usuário; data mais recente do intervalo da data.
IV	*Date Qualifier*	Não	Sim	Utilizado para rotular os vários conjuntos de datas.
V	*Location*	Sim	Sim	Inclui a localização geográfica da obra de arte ou arquitetura e o edifício ou repositório que atualmente abriga a obra.
V	*Creation Location*	Não	Sim	Onde a obra ou seus componentes foram criados, projetados ou produzidos.
V	*Discovery Location*	Não	Sim	Local geográfico onde uma obra foi escavada ou descoberta.
V	*Former Location*	Não	Sim	Local anterior para o registro da obra, incluindo locais relacionados à propriedade ou histórico de coleta/transferência da obra.
VI	*Subject*	Sim	Sim	Identificação, descrição ou interpretação do que é representado em e por uma obra ou imagem.
VI	*Controlled Subject*	Sim	Sim	Campo controlado com termos de assuntos indexados.
VI	*Extent*	Não	Não	Designar a parte do trabalho para a qual os termos do assunto são pertinentes, como lado a, lado b, verso, etc.

VI	*Subject Type*	Não	Não	Distinguir entre assuntos que refletem o que é a obra, assuntos descritivos ou interpretativos.
VII	*Class*	Sim	Sim	Usado para relacionar uma obra específica a outras com características semelhantes.
VIII	*Description*	Não	Não	Consiste em uma nota descritiva que geralmente é um ensaio relativamente breve como texto, detalhando o conteúdo e o contexto da obra.
VIII	*Sources*	Não	Não	Fontes de informação publicadas, podendo incluir obras de referência geral, enciclopédias, dicionários de arte, etc.
VIII	*Other Descriptive Notes*	Não	Não	Notas de texto livre, específicas de elementos adicionais que explicam ou qualificam informações em vários elementos da obra.
IX	*View Description*	Sim	Não	Incluem detalhes sobre a exibição da obra conforme ela aparece em uma imagem (substituta) da obra.
IX	*View Type*	Sim	Sim	Registra o ponto de vista específico ou perspectiva de uma imagem de uma obra.
IX	*View Subject*	Sim	Sim	Inclui termos ou frases que caracterizam o assunto da obra conforme descrito em uma imagem específica.
IX	*View Date*	Não	Não	Inclui qualquer data ou intervalo de datas associado à criação ou produção da imagem.
IX	*View Earliest Date*	Não	Não	Campo não exibido para o usuário; data mais antiga do intervalo da data.
IX	*View Latest Date*	Não	Não	Campo não exibido para o usuário; data mais recente do intervalo da data.

Fonte: elaborado pelo autor

Para os elementos indicados pelo CCO, pode-se observar algumas das regras de catalogação indicadas pelo guia no Quadro 2 abaixo, a fonte com todas as regras traduzidas pode ser acessada em (COELHO JÚNIOR, 2023) ou na íntegra no CCO.

Quadro 2 – Regras de catalogação mapeadas do guia CCO

Capítulo CCO	**Elemento CCO**	**Regras de catalogação**
Object Naming	*Work Type*	É obrigatório o preenchimento de valores; neste caso, não pode ficar vazio.
		Capitalize as iniciais de nomes próprios e a primeira letra do texto; para outros termos, use apenas letras minúsculas.
		Evite o uso de abreviações.
		Não utilize pontuação, com exceção do hífen.
		Use o mesmo idioma do catálogo.
		Use a ordem natural das palavras.
		Pode haver mais de um tipo por mudanças na obra ao longo do tempo; registre em ordem cronológica inversa: o mais recente ou relevante deve vir na frente.
		Utilize apenas palavras no singular.
		Faça uso de vocabulário controlado.
Object Naming	*Title*	É obrigatório o preenchimento de valores; neste caso, não pode ficar vazio.
		Deve ser conciso e descritivo.
		Caso a obra possua múltiplos títulos, o título de preferência deve ser destacado.
		Use o mesmo idioma do catálogo.
		Deve ser título dado pela instituição custodiadora ou título inscrito na obra, ou título providenciado pelo autor/artista, caso seja conhecido e descritivo o suficiente.
		Capitalize as iniciais de nomes próprios e da primeira palavra; para outros termos, use letras minúsculas.
		Títulos em outros idiomas, siga as regras de capitalização destes.
		Evite o uso de abreviações.
		Evite o uso de artigos, a não ser quando for crítico para o entendimento.
		Caso a obra possua mais de um título, o título mais popularmente conhecido deve ser utilizado como o preferencial.
		Pode se referir a assuntos históricos ou religiosos, pessoas, trabalhos ou lugares, e tipo do trabalho, dono, local ou história, nomes de edifícios.
		Obras podem ter múltiplos títulos; neste caso, especifique com o subelemento tipo do título.
		Os vários títulos de uma obra podem estar em diferentes idiomas; neste caso, identifique com o subelemento Idioma.
		Faça uso de vocabulário controlado para tipo do título, idioma e fonte.
Creator Information	*Creator*	Caso o autor seja desconhecido, este campo deve receber o valor 'desconhecido' ou a denominação da cultura que criou a obra.
		Caso haja ambiguidade ou incerteza de algum dado, deve-se registrar e apresentar todas as possibilidades.
		Autores podem participar em múltiplos papéis na criação de uma obra.
		Devem ser exibidos o nome do criador/autor e a biografia composta da nacionalidade e datas de nascimento e falecimento.
		O nome do autor deve vir de uma fonte de autoridade.
		Evite o uso de abreviações.
		Use o mesmo idioma do catálogo.

		Use a ordem natural das palavras.
		Deveria ser composto por papel, nome em ordem natural, nacionalidade (ou cultura) e data de nascimento e falecimento ou datas de atividade do autor.
		Caso o autor/artista tenha trocado de nome, deve-se registrar o mesmo nome utilizado na época da criação da obra.
		Caso haja dúvidas ou incertezas quanto as datas, deve-se indicar com: "ca.", "após", "depois" ou com a apresentação de intervalo de datas/períodos.
		Local de atividade pode ser indicado após a data, caso o local seja diferente da nacionalidade do autor.
		Em caso de criadores anônimos, deve-se utilizar nacionalidade deduzida e utilizar datas de nascimento e falecimento ou atividades aproximadas.
		No caso de várias entidades envolvidas na criação de uma obra, todas devem ser citadas. Porém, caso sejam muitas, deve-se citar as mais proeminentes ou mais relevantes.
		Caso órgãos corporativos façam parte da autoria da obra, estes devem ser citados na autoria junto ao nome da pessoa.
		Em caso de autoria desconhecida, o método utilizado para este caso deve ser consistente.
		É obrigatório o preenchimento de valores; neste caso, não pode ficar vazio.
		Faça uso de vocabulário controlado, incluindo papel e nacionalidade do autor.
Physical Characteristics	*Measurements*	É obrigatório o preenchimento de valores; neste caso, não pode ficar vazio.
		Recomenda-se utilizar o sistema métrico e também o imperial.
		É recomendado dar preferência ao sistema métrico.
		Caso um sistema métrico secundário seja utilizado, coloque valores entre parêntesis.
		Não use letras maiúsculas.
		Não repita unidade de medida para todas as dimensões, exceto quando for necessário para não causar confusão.
		Utilize números inteiros ou frações decimais.
		Deve-se utilizar abreviação nas unidades métricas, conforme indicado pelo Sistema Internacional, incluindo km, m, cm, mm, kg, g, kb, etc.
		Exemplos de tipos de medição podem ser altura, largura, profundidade, comprimento, circunferência, diâmetro, volume, peso, área e tempo de execução.
		Para medidas métricas, geralmente é apresentado duas casas decimais.
		Os catalogadores de recursos visuais ou de outros tipos que não estão medindo o objeto original, não devem arredondar as dimensões, mas devem registrar com precisão as medições encontradas em fonte autorizada.
		Faça uso de vocabulário controlado, incluindo características como forma (*shape*), escala (*scala*), formato (*format*), entre outras.
Physical Characteristics	*Materials and Techniques*	É obrigatório o preenchimento de valores; neste caso, não pode ficar vazio.
		Utilize apenas palavras no singular.
		Evite o uso de abreviações.
		Capitalize as iniciais de nomes próprios e a primeira letra do texto; para outros termos, use apenas letras minúsculas.
		Use o mesmo idioma do catálogo.

		Use a ordem natural das palavras.
		Liste primeiramente o meio ou a mídia, seguido pelo suporte, caso seja pertinente.
		Em caso de mais de um material ou técnica, a listagem deve seguir uma ordem lógica, podendo ser por importância ou ordem de aplicação/uso.
		Para trabalhos tridimensionais com uso de vários materiais, registre os mais proeminentes ou importantes.
		Para descrição de conjuntos de obras, liste os materiais e técnicas mais importantes ou mais evidentes no conjunto.
		Se for incerto qual meio foi utilizado, liste um meio mais abrangente.
		Caso haja ambiguidade ou incerteza de algum dado, deve-se registrar e apresentar todas as possibilidades.
		Faça uso de vocabulário controlado.
Physical Characteristics	*Physical Description*	Faça uso de vocabulário controlado para registrara descrição da aparência de uma obra expressa em termos genéricos, sem referência ao assunto retratado.
Physical Characteristics	*Inscription*	A abreviação e a capitalização de termos devem refletir a mesma presente na obra.
		A transcrição da inscrição deve ser idêntica à apresentada na obra.
		Deve apresentar a posição da inscrição na obra.
		Tradução de inscrições deve ser apresentada em colchetes.
		Inscrições transcritas parcialmente, geralmente por serem extensas, devem apresentar elipse "[...]".
		Caso haja ambiguidade ou incerteza de algum dado, deve-se registrar e apresentar todas as possibilidades.
		Partes ilegíveis devem ser apresentadas em colchete com o valor possível seguida de interrogação ([-?], [4?]).
		É obrigatório o preenchimento de valores; neste caso, não pode ficar vazio.
		Faça uso de vocabulário controlado, incluindo características de tipo e localização da inscrição na obra.
Stylistic, Cultural, and Chronological Information	*Date*	É obrigatório o preenchimento de valores; neste caso, não pode ficar vazio.
		Deve seguir formato consistente.
		Use a ordem natural das palavras.
		Evite o uso de abreviações.
		Capitalize as iniciais de nomes próprios e a primeira letra do texto, para outros termos use apenas letras minúsculas.
		Use o mesmo idioma do catálogo.
		Use traço para separar intervalo de anos.
		Expresse incerteza com "ca", “designado” ou “possivelmente”.
		Caso o ano possua menos que 4 (quatro) dígitos, zeros à esquerda devem ser inseridos.
		Siga padrão para registro da data com dia, mês e ano.
		Siga padrão para registro de tempo com hora, minutos e segundos.
		O registro de fuso horário deve ser consistente. Caso nenhum fuso horário seja indicado, será subtendido o fuso horário local.
		Não é obrigatória a indicação d.C. para datas do ano 1 (um) em diante, a não ser que possa causar confusão.
		Utilize a indicação a.C. para datas antes do ano 1.

		Caso seja utilizado outro calendário diferente do gregoriano, deve-se deixar clara esta informação.
		Caso a data informada não seja a data de conclusão, especifique a referência da data.
		Caso sejam registrados intervalos de tempo, especifique o tipo do intervalo, como, por exemplo, de construção, de design, lançamento, etc.
		Incerteza, datas aproximadas ou até mesmo quando as datas forem desconhecidas, estas devem ser apresentadas junto de: "provavelmente", "ou", "cerca de", "por volta de", etc.
		Obras do último século devem ter a *earliest date* e *latest date* um intervalo de 10 (dez) anos.
		Obras antigas devem ter a *earliest date* e *latest date* um intervalo de 100 anos.
		Para obras muito antigas, use a palavra como "por volta" em vez de "ca.".
		Caso a data exata seja desconhecida, pode-se utilizar datas relativas a um limite máximo e/ou mínimo, como por exemplo "antes de" ou "depois de".
		Não utilize apóstrofo.
		Datas de períodos ou eras podem receber o nome destes (neste caso, use vocabulário controlado). Porém, os campos *earliest date* e *latest date* devem ser preenchidos com valores respectivos a estes períodos.
		Uma obra pode apresentar mais de um intervalo de datas. Neste caso, repita os campos *earliest date* e *latest date* com um qualificador (*qualifier*) para discriminar o tipo de data.
Location and Geography	*Creation Location*	A designação de local incluirá cidade, subdivisão administrativa como estado (se aplicável) e nação, precedida pelo nome do repositório.
		Caso haja ambiguidade ou incerteza de algum dado, deve-se registrar e apresentar todas as possibilidades.
		Capitalize as iniciais de nomes próprios e a primeira letra do texto; para outros termos, use apenas letras minúsculas.
		Evite o uso de abreviações.
		Use o mesmo idioma do catálogo.
		Não utilize palavras obsoletas.
		Busque se basear em fontes de referências modernas.
		Utilize nomes diacríticos quando não houver adaptação ou quando é mais popular do que o nome no idioma da catalogação.
		Caso o local (*Location*) não possua fonte de autoridade para ser utilizada, crie uma baseada no *Anglo-American Cataloguing Rules* (AACR).
		Obras móveis podem possuir todos os tipos de localizações preenchidos.
		Obras estacionárias podem ter apenas a localização atual preenchida.
		É obrigatório o preenchimento de valores; neste caso, não pode ficar vazio.
		Obras de performance podem ter o local de criação preenchido e a localização atual recebe "não se aplica".
		Para registrar coleções privadas, cite o nome da coleção dado pelo dono, ou caso este queira permanecer anônimo, preencha com "coleção privada" e indique a localização geográfica.
		Caso o local não possua um nome, registre o local mais próximo que possua nome.
		Se a localização for incerta, indique e registre o lugar ou lugares prováveis.

		Manter histórico de empréstimos (locais) registrados.
		Utilize os nomes dos locais referentes à época.
		Faça uso de vocabulário controlado.
Subject	*Subject*	É obrigatório o preenchimento de alguns valores, neste caso não pode ficar vazio.
		Não inclua informações de cunho interpretativo se você não tiver base acadêmica para apoiá-la.
		É recomendado ser amplo e preciso do que específico e incorreto.
		Em caso de incerteza, deve-se utilizar termos mais abrangentes dos quais tenha certeza.
		Caso haja ambiguidade ou incerteza de algum dado, deve-se registrar e apresentar todas as possiblidades.
		Deve-se utilizar apenas palavras no singular.
		Capitalize as iniciais de nomes próprios e a primeira letra do texto, para outros termos use apenas letras minúsculas.
		Evitar o uso de abreviações.
		Deve ser utilizado o mesmo idioma que o catálogo é escrito.
		Deve-se utilizar a ordem natural das palavras.
		Informações adicionais devem estar entre parêntesis.
		Use palavras sensíveis ao contexto do trabalho catalogado.
		Deve-se incluir termos gerais e termos específicos.
		Inclua termos que descrevam o assunto de uma forma geral.
		Inclua termos para descrever o assunto da forma mais específica.
		Escolha termos adequados ao tipo de assunto que está sendo catalogado.
		Termos indexados não podem conter vieses, por exemplo um evento pode ser religioso ou mitológico, neste exemplo indexe as duas formas.
		Inclua termos para descrever conceitos temáticos e alegóricos.
		Descreva o assunto como retratado na obra.
		Deve ser utilizado vocabulário controlado.
Class	*Class*	É obrigatório o preenchimento de valores; neste caso, não pode ficar vazio.
		Não é recomendado que valores presentes no tipo da obra (*work type*) se repita na classe (*class*).
		Faça uso de vocabulário controlado.
		Deve seguir formato consistente.
		Utilize apenas palavras no singular.
		Quando apropriado, use conceitos compostos para o elemento classe (*class*) de uma coleção específica.
		Capitalize as iniciais de nomes próprios e a primeira letra do texto, para outros termos use apenas letras minúsculas.
		Evite o uso de abreviações.
		Se possível, não duplique nenhum termo usado no elemento tipo da obra (*work type*).
		Caso seja necessário, atribua várias classes (dado multivalorado).
Description	*Description*	Caso uma nota contenha qualquer informação significativa para recuperação, essa informação também deve ser registrada no elemento de metadados apropriado para indexação.
		Insira as informações de forma clara e concisa. Capture pontos distintos ainda não totalmente descritos em outros elementos.
		Use a ordem natural das palavras.

		Use frases completas.
		Liste as informações em ordem de importância, cronologicamente ou do geral para o específico, dependendo do que for apropriado para a obra específica.
		Liste as informações nesta ordem: qual é o trabalho com tipo da obra (*work type*), assunto (*subject*) e estilo (*style*). Seguido de quem é o responsável pela obra, onde foi feito e quando foi feito.
		Capitalize as iniciais de nomes próprios e a primeira letra do texto, para outros termos use apenas letras minúsculas.
		Evite o uso de abreviações.
		Use o mesmo idioma do catálogo.
		Inclua um esclarecimento de questões controversas ou incertas relacionadas à atribuição, localização original, identificação de assuntos, datas ou outras informações históricas relevantes, se apropriado.
		Pode-se citar as fontes de autoridades utilizadas entre as notas.
		Idealmente, um campo fonte (*source*) será associado à nota, e vinculado a um arquivo de autoridade bibliográfica para controlar os valores.
Description	*Other Descriptive Notes*	Caso uma nota contenha qualquer informação significativa para recuperação, essa deve ser registrada no elemento de metadados apropriado para indexação.
		Insira as informações de forma clara e concisa. Capture pontos distintos ainda não totalmente descritos em outros elementos.
		Use a ordem natural das palavras.
		Use frases completas.
		Liste as informações em ordem de importância, cronologicamente ou do geral para o específico, dependendo do que for apropriado para a obra específica.
		Liste as informações nesta ordem: qual é o trabalho com tipo da obra (*work type*), assunto (*subject*) e estilo (*style*). Seguido de quem é o responsável pela obra, onde foi feito e quando foi feito.
		Capitalize as iniciais de nomes próprios e a primeira letra do texto, para outros termos use apenas letras minúsculas.
		Evite o uso de abreviações.
		Use o mesmo idioma do catálogo.
		Inclua um esclarecimento de questões controversas ou incertas relacionadas à atribuição, localização original, identificação de assuntos, datas ou outras informações históricas relevantes, se apropriado.
		Pode-se citar as fontes de autoridades utilizadas entre as notas.
		Idealmente, um campo fonte (*source*) será associado à nota, e vinculado a um arquivo de autoridade bibliográfica para controlar os valores.
Location and Geography	*Location*	É obrigatório o preenchimento de valores; neste caso, não pode ficar vazio.
		A designação de local incluirá cidade, subdivisão administrativa como estado (se aplicável) e nação, precedida pelo nome do repositório.
		Caso haja ambiguidade ou incerteza de algum dado, deve-se registrar e apresentar todas as possibilidades.
		Capitalize as iniciais de nomes próprios e a primeira letra do texto, para outros termos use apenas letras minúsculas.
		Evite o uso de abreviações.
		Use o mesmo idioma do catálogo.
		Não utilize palavras obsoletas.

		Busque se basear em fontes de referências modernas.
		Utilize nomes diacríticos quando não houver adaptação ou quando é mais popular do que o nome no idioma da catalogação.
		Caso o local (*Location*) não possua fonte de autoridade para ser utilizada, crie uma baseada no *Anglo-American Cataloguing Rules* (AACR).
		Obras móveis podem possuir todos os tipos de localizações preenchidos.
		Obras estacionárias podem ter apenas a localização atual preenchida.
		Obras de performance podem ter o local de criação preenchido e a localização atual recebe "não se aplica".
		Para registrar coleções privadas, cite o nome da coleção dado pelo dono, ou caso este queira permanecer anônimo, preencha com "coleção privada" e indique a localização geográfica.
		Caso o local não possua um nome, registre o local mais próximo que possua nome.
		Se a localização for incerta, indique e registre o lugar ou lugares prováveis.
		Manter histórico de empréstimos (locais) registrados.
		Faça uso de vocabulário controlado.
View Information	*View Description*	Deve seguir formato consistente.
		Descreva os aspectos espaciais, cronológicos ou contextuais da obra, conforme capturados na visualização da imagem.
		Capitalize as iniciais de nomes próprios e a primeira letra do texto, para outros termos use apenas letras minúsculas.
		Não capitalize direções cardeais (leste, oeste, norte e sul).
		Evite o uso de abreviações.
		Use a ordem natural das palavras.
		Use o mesmo idioma do catálogo.
		É obrigatório o preenchimento de valores; neste caso, não pode ficar vazio.

Fonte: elaborado pelo autor

2.4 Automação em Processos de Tratamento Documental

A automação tornou-se cada vez mais prevalente em resposta ao crescimento exponencial de informações geradas a cada ano. Essa tendência é alimentada pelas ineficiências e erros inerentes às tarefas manuais de processamento de informações que exigem esforço humano significativo (BRYNJOLFSSON; MCAFEE, 2014, p. 177). Frey e Osborne (2017) argumentam que a automação pode ajudar a enfrentar esses desafios, melhorando a eficiência e a precisão do processo de produção de informações. Os autores sugerem que a automação pode permitir uma análise mais rápida e precisa de grandes conjuntos de dados, levando a uma tomada de decisão mais precisa e mais produtiva.

Um dos aspectos mais importantes da automação é sua capacidade de lidar com grandes volumes de dados de forma rápida e eficiente. E isso é fundamental para a produção informacional, que cresce exponencialmente a cada ano. De fato, o crescimento exponencial da produção informacional foi observado pela primeira vez por Bush (1945) durante a Segunda Guerra Mundial, e esse grande volume de produção documental tem sido documentado por muitos outros pesquisadores desde então (SARACEVIC, 1996; LYMAN; VARIAN, 2003).

Assim, o Quadro 3 apresenta o resultado de um levantamento de tecnologias mais recentes que vêm sendo utilizadas para extração, processamento e validação de metadados na literatura, tendo em vista que a presente pesquisa visa propor uma solução de processamento de dados e metadados para fins de obtenção de resultados diagnósticos de bases de dados culturais em sistemas documentais geridos pelo Ibram. Neste Quadro, pode-se observar a autoria, a ferramenta/metodologia utilizada e em qual contexto foi utilizado.

Quadro 3 – Tecnologias para extração automática de metadados

Autoria	**Estratégia de pesquisa**	**Síntese do estudo**
(CHARDONNENS et al., 2018)	Reconhecimento de entidades nomeadas (NER)	O artigo propõe um método para minerar consultas de usuários usando reconhecimento de entidades nomeadas (NER) e processamento de linguagem natural (NLP) para extrair informações relevantes e dados vinculados para fornecer informações contextuais.
(KIRCHHOFF et al., 2018)	Expressões regulares (*regex*) e Reconhecimento ótico de caracteres (OCR)	O artigo usa *regex* e OCR para automatizar a extração de informações de espécimes de herbário, o que inclui a extração de informações de várias fontes, como rótulos de espécimes, imagens e dados associados.
(JIANG et al., 2018)	*PDFBox*	O artigo propõe um método eficiente e preciso de extração e correspondência de metadados para documentos acadêmicos, usando a ferramenta PDFBox e semântica implícita em um grande volume de documentos em PDF.

(NASAR; JAFFRY; MALIK, 2018)	Modelos baseados em regras (como *regex*) e modelos de rede neural convolucional (CNN)	O artigo apresenta um levantamento dos métodos de extração de informação (IE) para artigos científicos. Os autores discutem modelos estatísticos e modelos baseados em regras, como modelos ocultos de Markov, campos aleatórios condicionais, máquinas de vetores de suporte e classificação Naive-Bayes, e também avanços recentes em métodos IE baseados em aprendizado profundo, como redes neurais convolucionais, redes neurais recorrentes, e modelos de transformadores.
(GIL et al., 2018)	Aprendizado de máquina (ML)	O artigo identifica direções principais de pesquisa para aprendizado de máquina em geociências, com auxilio na coleta de dados, indicação de uso e integração de observações isoladas em estudos mais amplos.
(TARMIZI et al., 2019)	ML	O estudo tem objetivo de avaliar a veracidade de notícias online e identificar possíveis *fake news*. Assim, usa o aprendizado de máquina para identificar palavras que carregam emoções e calcular o peso de cada estado emocional presente no conteúdo da notícia.
(OTMANI et al., 2019)	Ontologia e Processamento de linguagem natural (NLP)	O artigo propõe uma abordagem baseada em ontologia para aprimorar a extração de informações sobre conceitos médicos da web. Faz uso de processamento de linguagem natural (NLP). Os autores avaliaram a abordagem proposta em um conjunto de dados de páginas da web médicas, e os resultados mostram que ela supera os métodos existentes em termos recuperação da informação.
(DE CLERCQ et al., 2019)	NLP	O artigo propõe um método para classificação multi-rótulo e visualização interativa de dados de patentes de veículos elétricos, utilizando NLP. O método proposto combina algoritmos de aprendizado de máquina com técnicas de NLP para classificar patentes em várias categorias e visualizar os resultados usando uma interface interativa.
(LIAO; ZHAO, 2020)	ML	O artigo discute o uso de anotação semântica e aprendizado de máquina não supervisionado para extração de informações de texto não estruturado. O artigo descreve as subtarefas envolvidas na anotação semântica, incluindo reconhecimento de entidades e extração de relações, e discute várias ferramentas e técnicas usadas nessas tarefas.
(TANASIJEVIĆ; PAVLOVIĆ-LAŽETIĆ, 2020)	ML	O artigo apresenta o HerCulB, um sistema de extração e recuperação de informações baseado em conteúdo que emprega técnicas de aprendizado de máquina, como processamento de linguagem natural e visão computacional, para o patrimônio cultural dos Bálcãs. HerCulB usa modelos de aprendizado profundo para extrair informações de textos e imagens e para identificar semelhanças e conexões entre recursos do patrimônio cultural.
(PURWITASARI et al., 2020)	*NLP e ML*	O artigo propõe um método para identificar dinâmicas de colaboração em publicações científicas usando redes bipartidas autor-tópico e o Natural Language Toolkit (NLTK) e o *Scikit-learn*. O método proposto combina várias técnicas de aprendizado de máquina, como agrupamento e classificação, para identificar as influências das mudanças de interesse na dinâmica de colaboração.
(C; MAHESH, 2021)	*OpenRefine*	O artigo discute o uso da análise baseada em gráficos para entender e modelar sistemas complexos, com foco nos dados do COVID-19 do estado de Karnataka, na Índia. Os autores usaram o conjunto de dados KaTrace, que foi limpo usando o OpenRefine, mesclado com dados do censo e dados do centróide e usando análises baseadas em gráficos.

(ZHANG et al., 2021)	NER e NLP	O artigo discute o desenvolvimento de um novo algoritmo de incorporação de gráficos baseado em GeoAI para ajudar a construir relações semânticas entre diferentes dados espaciais para recuperação de informação em dados espaciais. O artigo também descreve o uso de NER para processo de extração de entidade geral e processo de extração de entidade de produto de dados. NLP foi usada para construir um gráfico de conhecimento de dados baseado em semelhanças semânticas e espaço-temporais de metadados.
(BRACK et al., 2022)	Aprendizado profundo (DL)e ML	A abordagem proposta emprega um modelo de aprendizado profundo com um codificador compartilhado que aprende representações comuns entre domínios e um decodificador específico de tarefa que aprende a classificar sentenças em cada domínio. O codificador compartilhado melhora a precisão da tarefa de classificação aproveitando as informações de diferentes domínios.
(NAMGUNG; CHIANG, 2022)	Modelo BART e OCR	O artigo propõe um método para melhorar a precisão do OCR em imagens de mapas, utilizando o modelo BART para incorporar informações de contexto espacial. Os autores demonstram que o OCR em imagens de mapas é desafiador devido à presença de vários objetos espaciais, como estradas, rios e edifícios. O modelo BART é usado para prever a probabilidade de cada pixel pertencer a uma classe de objeto espacial específica, que é então usada para ponderar a importância de cada pixel no processo de OCR.
(FELL et al., 2022)	NLP	O artigo apresenta o corpus WASABI, um conjunto de dados de letras de músicas e um gráfico de conhecimento para análise de letras de músicas. Os autores propõem um método que combina modelos de NLP com representação gráfica de conhecimento para extrair e analisar características linguísticas de letras de músicas, como sentimento, emoção e assuntos.
(LI et al., 2023)	DL e CNN	O artigo discute o CoRec, uma estrutura de recomendação baseada em comportamento da Internet que combina computação de borda e computação em nuvem usando redes neurais convulsionais profundas (CNNs). O CoRec visa fornece recomendações personalizadas e eficientes, capturando dados de comportamento dos usuários na Internet e usando um modelo CNN profundo para modelar as interações usuário-objeto e o conteúdo do objeto informacional.

Fonte: elaborado pelo autor

Com base na revisão bibliográfica, torna-se evidente que há várias metodologias descritas na literatura para automatizar ou semiautomatizar a criação, extração e processamento de dados e metadados.

Algumas dessas metodologias incluem o processamento de texto, como expressões regulares (*regex*), reconhecimento ótico de caracteres (OCR), reconhecimento de entidades nomeadas (NER) e processamento de linguagem natural (NLP). Além disso, existem metodologias baseadas em aprendizado de máquina, como aprendizado profundo e redes neurais, bem como ferramentas de limpeza e transformação de dados, como *OpenRefine* e PDFBox.

A aplicação dessas diversas tecnologias abrange diversas áreas, desde Ciências Naturais e Biológicas até Ciência da Computação e Ciência da Informação. Com o uso dessas técnicas as informações podem ser analisadas de forma rápida e eficiente, permitindo identificar padrões e tendências que podem ser úteis para pesquisadores e profissionais (CHEN; MAO; LIU, 2014).

Nesta pesquisa, será utilizada a tecnologia *regex*, um método de processamento de texto baseado em regras ideal para processar resultados de processos de catalogação registrados em bases de dados, o qual será elucidado a seguir.

2.5 Expressão Regular para Tratamento Documental

As expressões regulares, também conhecidas como *regex*, são uma linguagem formal utilizada para análise e processamento de texto. Com o auxílio de um processador de expressão regular, que funciona como um analisador sintático, é possível identificar as partes do texto que correspondem à especificação dada pela *regex*. Esse conceito foi formalizado pelo matemático americano Stephen Cole Kleene na década de 1950, quando utilizou expressões regulares para descrever uma linguagem regular (KLEENE, 1956).

As expressões regulares são ferramentas poderosas e eficientes para manipulação de texto, permitindo a descrição e análise de padrões em uma ampla variedade de dados. Com uma notação de padrão geral semelhante a uma minilinguagem de programação, as *regex* podem ser usadas para adicionar, remover e isolar texto de forma precisa e rápida. Por essa razão, as expressões regulares são amplamente utilizadas em diversas áreas, sendo uma ferramenta indispensável para programadores que necessitam processar grandes quantidades de texto de forma eficiente (FRIEDL, 2002, p. 22).

Embora a linguagem de programação, como PHP, Java, .NET e *Python*, por exemplo, forneçam suporte para processamento de texto, o real poder das expressões regulares está na sua capacidade de identificar o texto desejado e ignorar o desnecessário. Combinando essas expressões com a construção de suporte da linguagem, é possível realizar diversas ações no texto, como adicionar códigos de destaque apropriados, remover ou alterar o texto, entre outras. Dessa forma, é

possível explorar todo o potencial desse recurso e utilizá-lo para a validação de dados em formulários e sistemas.

As *regex* se tornaram populares para uso prático em editores de texto na década de 1960 e atualmente várias linguagens de programação possuem formas diferentes de lidar com esse recurso (THOMPSON, 1968). Na biblioteca *re* do *Python*, por exemplo, é possível utilizar expressões regulares para processar textos.

Desta forma, as expressões regulares são uma técnica desenvolvida em Ciência da Computação Teórica e Teoria da Linguagem Formal que se tornou uma ferramenta indispensável para diversas áreas. Com sua capacidade de manipular grandes quantidades de texto de forma precisa e eficiente, as *regex* são amplamente utilizadas em campos como busca, validação de dados, substituição de texto e registro de atividades (EXPRESSÃO REGULAR, 2020; FRIEDL, 2002; CROCHEMORE; RYTTER, 1994, p. 157).

Assim, no contexto das coleções museológicas do Ibram, esse recurso será utilizado, à luz das regras de catalogação do CCO, para validar a adequação dos dados. Como exemplo de uma regra *regex*, podemos observar a seguinte situação:

Para a procura de ocorrências de abreviações em textos, podemos utilizar a seguinte *regex*:

```
([A-ZÁÉÍÓÚÜÑ][A-Za-z0-9áéíóúüñ]*\.)
```

Conforme abaixo, apresenta-se o que cada bloco de elemento faz:

- O primeiro elemento, `[A-ZÁÉÍÓÚÜÑ]`, é uma classe de caracteres que representa uma letra maiúscula presente no alfabeto latino, incluindo letras acentuadas. Assim, qualquer letra do nosso alfabeto na forma maiúscula com ou sem acentos que apareça uma única vez.
- O segundo elemento, `[A-Za-z0-9áéíóúüñ]`, é uma classe de caracteres que representa uma letra maiúscula, incluindo letras acentuadas, ou um dígito numérico. Este elemento é seguido por um quantificador asterisco (*), que representa a ocorrência de zero ou mais vezes do elemento anterior. Diferente do primeiro elemento, neste bloco é considerado números também.

- O terceiro elemento, `\.`, é o caractere "ponto" ou "ponto final", que é utilizado para indicar o final de uma abreviação. Este elemento é seguido por um quantificador +, que representa uma ou mais ocorrências do elemento anterior.

Desta forma, o texto que atender a regra da *regex* pode ser considerado adequado ou não, dependendo da avaliação realizada e do objetivo. No caso da catalogação, os valores de dados ou registros de metadados disponíveis em uma base de dados, por exemplo, seriam processados a partir de regras da *regex*, fundamentadas nas orientações advindas de códigos de catalogação específicos.

3 ESTUDO DO ESTADO DA ARTE: MODELOS DE DIAGNÓSTICO DE QUALIDADE DE DADOS NO DOMÍNIO DO PATRIMONIO CULTURAL

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 EUA. Em suas coleções e conteúdos digitais (FENLON et al. 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. 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.

A Biblioteca Digital da Universidade de Houston (WESTBROOK, 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*. 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 avaliação 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 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 usuários externos que consomem os dados. Por fim, a metodologia foi implantada em um período de 26 meses, com a realização do total de 8 (oito) experimentos, e pode-se observar que a qualidade dos registros de metadados nos repositórios melhorou em média 16% quanto à 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 2 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 *HighCharts* JS 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 (dez) 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, Estados Unidos da América (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 6 (seis) meses e 1 (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 GARCIA et al. 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 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 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, 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 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.

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 possibilidades 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 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 et al. 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 4 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 et al. 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 5 (cinco) níveis, sendo 0 (zero) o nível mais baixo e 4 (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 INBCM (MINISTÉRIO DA CULTURA, 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 1(um) de qualidade. Para as linguagens documentárias receberam o nível 2 (dois). Para os tipos de mídia e licenças o nível 2 (dois) foi o evidenciado, com exceção de um dos classificados em nível superior aos demais, recebendo o nível 3 (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 apresentada uma estrutura para transformar conjuntos de dados de organizações GLAM em *Linked Open Data* (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 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 Instituto Brasileiro de Museus (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 4 a seguir sintetiza as iniciativas de projetos voltados à avaliação da qualidade de dados em acervos culturais.

Quadro 4 – Quadro sinóptico sobre projetos de avaliação de qualidade de dados no domínio cultural

Autoria	**Tipo de instituição**	**Escopo**	**Método de avaliação**
(FENLON et al. 2012)	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 destacando pontos de atenção.	Modelos estatísticos.
(WESTBROOK, 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)	Universidade de Pisa, Universidade de Roma e Universidade de Turin - Itália	Baseado em regras elencadas, são avaliados os requisitos completude, precisão e consistência do acervo cultural.	Algoritmo de avaliação desenvolvido 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 em marcações XML, com indicação de ausência de elementos obrigatórios, erros de marcação e também a avaliação do padrão de conteúdo de dados como relacionamento inconsistente nas marcações, dados duplicados, entre outros.	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, e frequência de valores nestes campos.	Avaliação manual de amostra.

(GAONA GARCIA et al. 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)	Biblioteca Nacional da França, Biblioteca Nacional da Espanha, Bibliografia Nacional Britânica e Biblioteca Virtual Miguel de Cervantes.	Avaliação em coleções disponíveis em RDF com preceitos *Linked Open Data* (LOD) em 11 dimensões com total de 35 critérios, para verificação de adequação dos dados.	Expressões regulares (*regex*), validadores sintáticos externos e consultas estruturadas 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)	A Biblioteca Virtual Miguel de Cervantes, a Biblioteca Nacional da França, a Biblioteca Nacional Britânica e a Biblioteca Nacional de Espanha	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*)
(LORENZINI et al. 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	É apresentado 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	É realizado a avaliação de qualidade de dados de 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 pelo autor

4 METODOLOGIA DA PESQUISA

Metodologicamente, este estudo adotou uma abordagem aplicada, combinando elementos qualitativos e quantitativos, além de exploratória e descritiva, baseado em um estudo de caso envolvendo 22 coleções de museus digitais gerenciadas pelo Ibram. 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. O estudo qualitativo e quantitativo visa obter uma compreensão mais ampla e completa dos dados coletados, o que pode resultar em uma avaliação mais precisa e confiável. Assim, a Seção 4.1 se incube de apresentar as bases de dados científicas exploradas para fundamentar e desenvolver a pesquisa; a Seção 4.2 discorre acerca do Ibram e de seu surgimento; a Seção 4.3 apresenta as etapas metodológicas projetadas para o desenvolvimento do modelo de avaliação diagnóstica implementado para o processamento de 22 coleções culturais representadas em bases de dados pertencentes ao Ibram; e a Seção 4.4 apresenta o modelo Cascata de desenvolvimento de *software*.

4.1 Fundamentos da pesquisa

Para a fundamentação teórica e metodológica da pesquisa, e o levantamento do estado da arte apresentado no Capítulo 3, usou-se de levantamento bibliográfico em bases de dados de documentos científicos, sendo a principal delas o portal de

periódicos da Capes para recuperar artigos recentes, relevantes e com fator de impacto considerável aos propósitos desta pesquisa. A pesquisa documental também foi considerada neste estudo, tendo em vista o uso de documentações acerca de padrões de documentação, incluindo o guia de catalogação CCO e o inventário INBCM.

Para o levantamento do estado da arte sobre modelos de diagnóstico de qualidade de dados, 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 selecionar documentos que não são indexados nas bases de dados mais institucionais.

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. Com recorte temporal de estudos que foram publicados entre 2012 e 2023. 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 5) nas bases de dados de pesquisa acadêmica online, que, após a retirada das duplicatas, foi reduzido para 438.

Quadro 5 – Quantidade de resultados por serviço de busca

Fonte		**Idioma da busca**	**Quantidade de resultados**	**Data**
Base de Dados Referenciais de Artigos de Periódicos em Ciência da Informação (BRAPCI)		EN	*0*	Janeiro de 2012 a fevereiro de 2023
		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 pelo autor.

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; 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 e foram selecionados na revisão (Quadro 6).

Quadro 6 – 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 pelo autor

Para o levantamento de tecnologias recentes usadas em processos automáticos ou semiautomáticos de extração ou criação de metadados no fluxo de produção informacional (Seção 2.3), os resultados obtidos no Portal de Periódicos CAPES foram complementados com os resultados da *ACM Digital Library* e *IEEE Xplore Digital Library*. Por meio da busca avançada, foram utilizados os descritores "*Information extraction*", "*Metadata*" e "*Information*".

Os critérios de inclusão das publicações no estudo foram artigos originais e de revisão, publicados de janeiro de 2018 a fevereiro de 2023, nos idiomas português, espanhol ou inglês. Foram excluídos estudos repetidos, editoriais, teses, dissertações, livros, resumos e trabalhos que não descreviam qual metodologia utilizada.

A partir da estratégia de busca, foram identificados um total de 105 trabalhos (Quadro 7).

Quadro 7 – Total de trabalhos recuperados por fonte

Fontes		**Total de resultados**
Portal de periódicos Capes	*Biodiversity Information Science and Standards*	2
	Database: the journal of biological databases and curation	1
	DESIDOC journal of library & information technology	1
	Electronic library	2
	IJGIS	1
	Ingénierie des systèmes d'Information	1
	International journal of Web information systems	1
	Journal of database management	1
	Journal of documentation	1
	Language resources and evaluation	1
	Pizhūhishnāmah-i pardāzishvamudiriyyat-i iṭṭilā'āt	1
	Scientometrics	2
	SIGIR forum	2
	Springer	1
	World patent information	1
ACM		86
IEEE		0

Fonte: elaborado pelo autor

Após adoção dos critérios de inclusão e exclusão, remoção de duplicatas e leitura flutuante, a busca resultou em 17 artigos para leitura na íntegra, sendo 9 (nove) trabalhos no Periódico CAPES e 8 (oito) trabalhos no *ACM Digital Library*. Nenhum trabalho foi encontrado na base de dados *IEEE Xplore Digital Library* que atendesse os critérios de inclusão (Quadro 8).

Quadro 8 – Passos de inclusão dos artigos selecionados

Fase	**Critérios**	**Estudos restantes**
Busca	Quantidade total de resultados obtidos	105
Filtro de idioma e duplicatas	Artigos completos em inglês, espanhol ou português sem duplicatas	99
Filtro por tipo de estudo	Estar acessível e possuir resumo Não ser pesquisa de opinião, resumo ou capítulos de livro	39
Filtro de estudos no escopo de interesse	Apresentam metodologia utilizada	17

Fonte: elaborado pelo autor

4.2 Estudo de Caso: Instituto Brasileiro de Museus

Durante o período de 2003 a 2015, houve um notável avanço no setor museológico brasileiro, em grande parte devido ao aumento dos investimentos em políticas públicas culturais. Nesse ano, foi criada a Política Nacional de Museus (PNM), o primeiro documento a estabelecer diretrizes conceituais sobre o papel dos museus e o direito à memória da sociedade brasileira (SANTANA, 2020).

Como resultado da PNM, em 2003 foi criado pelo Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN) o Departamento de Museus e Centros Culturais (DEMU). A criação do DEMU se deu devido à singularidade do conjunto de museus do IPHAN e à ausência formal de um setor na área federal voltado às ações no campo da museologia (BRASIL, 2007, p. 29).

O DEMU tinha como principal propósito a criação de políticas relacionadas à preservação do patrimônio material, imaterial, arqueológico e aos museus. Entretanto, segundo Santana (2020), a gestão do IPHAN sobre os museus apresentava problemas, o que motivou a criação do Instituto Brasileiro de Museus (Ibram).

Em janeiro de 2009, foram sancionadas as Leis nº 11.904, que cria o Estatuto de Museus, e nº 11.906, que cria o Ibram. Com isso, o DEMU foi extinto e 28 unidades museológicas passaram a ser de responsabilidade do Ibram.

O Ibram é uma autarquia federal vinculada ao Ministério do Turismo e tem como missão promover a valorização dos museus e do campo museal, a fim de garantir o direito à memória, o respeito à diversidade e a universalidade de acesso aos bens musealizados. Além disso, é responsável pela Política Nacional de Museus (PNM) e pela administração direta de 30 museus distribuídos em nove estados brasileiros (BRASIL, 2023). Sua visão é ser referência na gestão de políticas públicas e na geração e difusão de conhecimento para o campo museal, regulamentado pela Portaria nº 66 de 22 de fevereiro de 2018.

Entre seus projetos, destaca-se o programa Acervo em Rede, que visa democratizar o acesso digital aos bens culturais musealizados, promovendo a digitalização e a documentação dos acervos das instituições museológicas na internet (ACERVO EM REDE, 2023).

Atualmente, o Ibram disponibiliza online dados de 23 coleções, abrangendo um total de 30 museus e mais de 200 mil itens em tratamento das informações (MARTINS;

MARTINS, 2021). Além disso, mais de 19 mil itens estão disponíveis para consulta (IBRAM, 2023), conforme constatado no Quadro 8.

Esse acesso aos dados permite que o público interessado conheça mais sobre o patrimônio musealizado brasileiro, ao mesmo tempo que auxilia na gestão e preservação desse patrimônio. Ademais, o Ibram promove diversos eventos e programas de capacitação voltados para profissionais da área museológica, a fim de aprimorar a gestão dos museus e a valorização do patrimônio cultural do país.

A criação da Política Nacional de Museus e do Instituto Brasileiro de Museus representou um avanço significativo na valorização e preservação do patrimônio cultural brasileiro. Apesar dos desafios enfrentados durante o processo, hoje o Ibram atua de forma abrangente e diversa, promovendo a democratização do acesso aos museus e bens culturais, assim como aprimorando a gestão e preservação do patrimônio musealizado do país.

4.3 Etapas do processo de avaliação diagnóstica

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; da área da Biblioteconomia são num total de 19 elementos, sendo 15 de entrada obrigatória e 4 (quatro) de entrada facultativa; por fim, da área da Arquivologia são num total de 16 elementos, sendo 7 (sete) de entrada obrigatória e 9 (nove) de entrada facultativa. Informações completas sobre esses elementos de descrição podem ser conferidas na íntegra em sua referência (MINISTÉRIO DA CULTURA, 2021).

No que diz respeito ao INBCM, foram considerados apenas os 15 elementos (Quadro 9) de descrição para identificação do bem cultural de caráter museológico.

Quadro 9 – Elementos INBCM para identificação do item de caráter museológico

Elemento de descrição	Indicação	Descrição
Número de registro	Obrigatório	Registro individual definido pelo museu para identificação e controle do objeto dentro do acervo.
Outros números	Facultativo	Numerações anteriores atribuídas ao objeto, tais como números antigos e números patrimoniais.
Situação	Obrigatório	Informação da situação em que se encontra o objeto, o seu status dentro do acervo do museu, com a marcação das opções: 1 (um)- localizado; 2 (dois) - não localizado; 3 (três)– excluído.
Denominação	Obrigatório	Nome que identifica o objeto.
Título	Facultativo	Denominação dada ao objeto atribuído pelo autor, curador ou pelo profissional da documentação.
Autor	Obrigatório	Nome do autor do objeto (individual ou coletivo).
Classificação	Facultativo	Classificação do objeto segundo o "Thesaurus para Acervos Museológicos" ou outros vocabulários controlados.
Resumo descritivo	Obrigatório	Resumo da descrição textual do objeto, apresentando as características que o identifique inequivocamente e sua função original.
Dimensões	Obrigatório	Dimensões físicas do objeto, considerando-se as medidas bidimensionais (altura x largura); tridimensionais (altura x largura x profundidade); circulares (diâmetro x espessura) e peso.
Material / Técnica	Obrigatório	Materiais do suporte que compõem o objeto, hierarquizando sempre a sua maior área confeccionada/manufaturada e a técnica empregada na sua manufatura.
Estado de conservação	Obrigatório	Estado de conservação em que se encontra o objeto na data da inserção das informações.
Local de produção	Facultativo	Indicação geográfica do local onde o objeto foi confeccionado.
Data de produção	Facultativo	Data ou período de confecção/produção/manufatura do objeto.
Condições de reprodução	Obrigatório	Descrição das condições de reprodução do objeto, indicando se há alguma restrição que possa impedir a reprodução/divulgação da imagem do objeto nos meios ou ferramentas de divulgação.
Mídias relacionadas	Facultativo	Informação acerca da inserção de arquivos de imagem, sons, vídeos e/ou textuais relacionados ao objeto.

Fonte: elaborado pelo autor

Tal decisão foi feita após análise prévia dos dados captados dos acervos à luz das orientações do INBCM. Apenas o Museu Solar Monjardim apresentava acervo do tipo Arquivístico. Todos os outros acervos (22 coleções no total) dos 20 museus utilizam metadados especificados pelo INBCM com caráter museológico, conforme demonstra o Quadro 10.

Quadro 10 – Quantidade de Itens por coleção dos museus sob gestão do Ibram

#	**Museu**	**Coleção**	**Endereço eletrônico**	**Quantidade de itens**
1	Museu da Inconfidência	Coleção museológica	https://museudainconfidencia.acervos.museus.gov.br/acervo-museologico/	4.622
2	Museu Villa-Lobos	Fotografias	https://museuvillalobos.acervos.museus.gov.br/fotografias/	1.812
3	*Museu Solar Monjardim*	**Coleção arquivística**	*http://museusolarmonjardim.acervos.museus.gov.br/documentos-historicos/*	*1.750*
4	Museu Casa da Hera	Coleção museológica	https://museucasadahera.acervos.museus.gov.br/acervo-museologico/	1.220
5	Museu Histórico Nacional	Moedas de ouro	https://mhn.acervos.museus.gov.br/moedas-de-ouro/	1.215
6	Museu de Arqueologia de Itaipu	Coleção museológica	https://museudearqueologiadeitaipu.museus.gov.br/museu-itaipu/	1.040
7	Museu Histórico Nacional	Coleção museológica	https://mhn.acervos.museus.gov.br/reserva-tecnica/	1.040
8	Museu Casa de Benjamin Constant	Coleção museológica	http://museucasabenjaminconstant.acervos.museus.gov.br/acervo-museologico/	983
9	Museu do Diamante	Coleção museológica	https://museudodiamante.acervos.museus.gov.br/acervo-museologico/	895
10	Museu Casa da Princesa	Coleção museológica	https://museusibramgoias.acervos.museus.gov.br/museu-casa-da-princesa/	799
11	Museu de Arte Sacra da Boa Morte	Coleção museológica	https://museusibramgoias.acervos.museus.gov.br/museu-casa-da-boa-morte	790
12	Museu Casa Histórica de Alcântara	Coleção museológica	https://museucasahistoricadealcantara.acervos.museus.gov.br/objetos/	631
13	Museu Regional Casa dos Ottoni	Coleção museológica	https://museuregionalcasadosottoni.acervos.museus.gov.br/acervo-museologico/	463
14	Museu das Bandeiras	Coleção museológica	https://museusibramgoias.acervos.museus.gov.br/museu-das-bandeiras/	415
15	Museu Regional de São João del-Rei	Coleção museológica	https://museuregionaldesaojoaodelrei.acervos.museus.gov.br/acervo_museologico/	328

16	Museu da Abolição	Coleção museológica	https://museudaabolicao.acervos.museus.gov.br/acervo_museologico/	301
17	Museu Regional de Caeté	Coleção museológica	https://museuregionaldecaete.acervos.museus.gov.br/acervo/	243
18	Museu Victor Meirelles	Coleção museológica	https://museuvictormeirelles.acervos.museus.gov.br/mvm-acervo/	237
19	Museu de Arte Religiosa e Tradicional	Coleção museológica	https://museudeartereligiosaetradicional.acervos.museus.gov.br/acervo-museologico/	132
20	Museu do Ouro	Coleção museológica	https://museudoouro.acervos.museus.gov.br/acervo/	126
21	Museu das Missões	Coleção museológica	https://museudasmissoes.acervos.museus.gov.br/acervo-museologico/	90
22	Museu Solar Monjardim	Coleção museológica	https://museusolarmonjardim.acervos.museus.gov.br/acervo/	77
23	Museu Casa da Hera	Indumentárias	https://museucasadahera.acervos.museus.gov.br/indumentaria/	67
TOTAL				19.276

Fonte: IBRAM, 2023. Coletado em 30/07/2022.

Outro detalhe importante a ser destacado é que foram considerados 8 (oito) dos 9 (nove) capítulos do CCO. O Capítulo 9 (nove), denominado *View Information*, é endereçado à catalogação do substituto digital de uma obra, a exemplo de uma imagem. Os dados de catalogação dos acervos museais vinculados ao Ibram são referentes às obras presentes nos museus, logo, não descreve as imagens representativas dessas obras.

As etapas do método de avaliação dos dados dos acervos elencados para o alcance dos resultados são descritas nas subseções a seguir.

4.3.1 Alinhamento entre elementos de descrição: INBCM e CCO

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 no Quadro 11. O alinhamento foi realizado através de um procedimento manual e intelectual que requereu a aquisição de conhecimento sobre ambos os

instrumentos de pesquisa. Este processo de mapeamento foi baseado na correspondência dos campos descritivos em sua descrição e função.

É importante destacar que o capítulo VI, na seção 2 (dois) do guia CCO, dedicado ao elemento central "assunto" (*Subject*), não é considerado funcionalmente e com clareza semântica, conforme se prevê no guia, para os elementos de descrição do INBCM no contexto de caráter museológico. O elemento “classificação” do INBCM pode até ser indicado como um assunto controlado pelos tesauros recomendados pela instrução normativa, mas a mesma deixa bem em aberto a forma como o catalogador pode interpretar a função deste elemento. Percebe-se que para os outros dois contextos, bibliográfico e arquivístico, há recomendações do INBCM para o tratamento da representação temática “assunto”.

Sugere-se, portanto, que esse tipo de representação temática não é considerado relevante para o contexto dos museus associados ao Ibram, ou pode não ter sido explorado o seu potencial pelos especialistas em documentação no que diz respeito aos SRIs (LANCASTER, 2004). Além disso, é importante mencionar que o capítulo IX, na seção 2 (dois) do guia, que trata dos detalhes sobre a exibição da obra com base em uma imagem substituta da obra, também não foi considerado, uma vez que o Ibram fornece dados descritivos sobre as obras e não há dados relacionados a imagens das obras. Logo, o experimento da presente pesquisa considerou 7 (sete) dimensões analíticas enumeradas e descritas a seguir:

- Object Naming: fornece maneiras de se referir a uma obra, definindo o que está sendo catalogado.
- Creator Information: identifica o criador de uma obra (podendo ser vários), podendo ser uma pessoa, física ou jurídica, conhecida pelo nome ou anônima.
- Physical Characteristics: descreve a aparência de uma obra, apresentando características de sua forma física.
- Stylistic, Cultural, and Chronological Information: descreve características estilísticas de uma obra, origens culturais e data de design ou criação.
- 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.

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

4.3.2 Exploração semiautomática das bases de dados de coleções

Para a obtenção dos dados para a presente pesquisa, foi desenvolvido um *script* (COELHO JÚNIOR, 2023) por meio da utilização da linguagem de programação *Python*, e com o uso das bibliotecas *Pandas*[9], *BeautifulSoup*[10] e *Requests*[11], para realizar a exportação em massa de todos os dados dos acervos dos museus no formato “CSV: inbcm-ibrammapper”. Ressalta-se que esse formato é um dos disponíveis para exportação no *software* Tainacan e, portanto, segue as recomendações do INBCM. Porém, há também exportação em formatos JSON e HTML disponibilizados pela API (GOV.BR, 2021).

Algumas técnicas de pré-processamento de dados foram usadas a fim de deixar a base de dados do experimento padronizada em termos de quantidade de campos, a saber: foi removido das bases o elemento de descrição “Outros Números”, conforme o procedimento de alinhamento; foram renomeados os títulos dos elementos discricionais para o padrão utilizado no CCO; e por fim, todas as bases das coleções trabalhadas foram agrupadas em uma mesma base para fins de processamento.

Em seguida, a avaliação envolveu as seguintes etapas:

- Mapeamento de todas as regras explicitadas nos capítulos I, II, III, IV, V, VII e VIII do guia CCO, identificando 244 regras, incluindo 122 regras distintas.
- Seleção das regras pertencentes aos elementos descritivos alinhados com o INBCM (Inventário Nacional de Bens Culturais Móveis), que não

[9] https://pandas.pydata.org
[10] https://beautiful-soup-4.readthedocs.io/
[11] https://requests.readthedocs.io/en/master/

apresentassem fator subjetivo e que, portanto, pudessem inviabilizar tecnicamente a avaliação por algoritmo computacional.

- Implementação das regras mapeadas em *Python* seguindo os fundamentos do campo da catalogação descritiva.
- Utilização de padrões para tratamento nos dados, a saber: padrão de conteúdo de dados e padrão de valor de dados nos acervos dos museus envolvidos na análise.
- Implementação do padrão de conteúdo de dados no algoritmo a partir de regras da *regex*, utilizando a biblioteca *re* do *Python*.
- Implementação do padrão de valor de dados por meio da avaliação da utilização de vocabulário controlado, a partir dos dados disponibilizados pela API do Tainacan.

Dessa forma, o processo sistemático envolveu desde o mapeamento e seleção de regras até a implementação dessas regras em *Python*, utilizando padrões para tratamento dos dados e avaliação de vocabulário controlado.

Com a avaliação de adequação dos dados, para cada regra (documentada em COELHO, 2023) 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 seguinte fórmula:

$$\textbf{índice}_{b} = (\sum \textbf{Valor1} \,/\, (\sum \textbf{Valor1} + \sum \textbf{Valor0})) * 100$$

Onde:

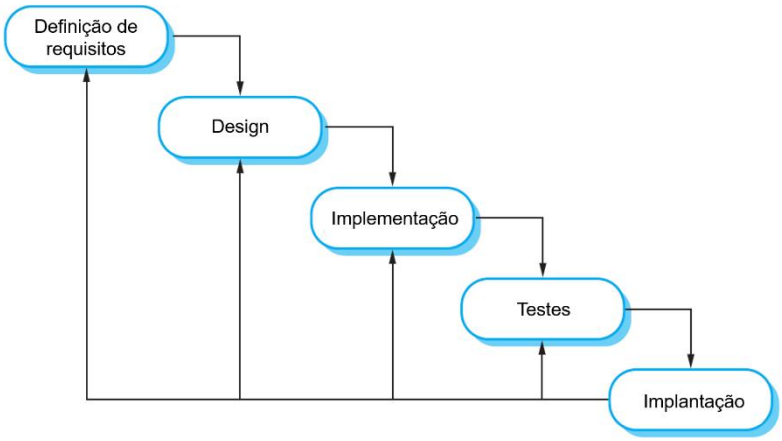
- **b** é a base com a amostra de dados de uma coleção particular;
- **índice** é o percentual de adequação obtido em relação a dimensão, a elemento de metadado e a regra de catalogação para um determinado museu e coleção;
- **valor1** é a indicação de ocorrência do registro de dado que se adequou a regra;
- **valor0** é a indicação de ocorrência do registro de dado que não atendeu a regra.

4.4 Modelo Cascata de desenvolvimento de software

O modelo de desenvolvimento de software em cascata, também conhecido como modelo Cascata, em inglês *Waterfall*, é uma das metodologias de desenvolvimento mais antigas e conhecidas. O modelo foi desenvolvido originalmente na década de 1970, como uma forma de gerenciar projetos complexos de 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 apresentado na Figura 1.

Figura 1 – Etapas do modelo Cascata



Fonte: Adaptado de Sommerville (2016)

As etapas podem ser descritas da seguinte forma e com as seguintes funções:

- 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 bugs. 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.

A luz das orientações do método Cascata, os processos listados foram executados durante o desenvolvimento da aplicação, como pode ser observado abaixo:

Na primeira etapa, foram identificados quatro requisitos básicos que a ferramenta deveria cumprir: receber um conjunto de dados (*dataset*); realizar o alinhamento entre os elementos do *dataset* que o usuário forneceu com os elementos do CCO; apresentar os resultados da avaliação de qualidade de dados e indicar como a qualidade de dados do *dataset* avaliado pode ser melhorada.

Na segunda etapa, desenhamos um fluxograma dos processos realizados pela aplicação e o apresentamos na Figura 2. As ações realizadas pelo usuário estão destacadas em amarelo, enquanto as ações realizadas pela aplicação estão em rosa.

Figura 2 – Diagrama de processos da aplicação

Fonte: elaborado pelo autor

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*[12]

[12]https://flask.palletsprojects.com/en/2.2.x/

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, diferentemente da avaliação realizada no estudo de caso com os acervos do Ibram, nesta ferramenta não será utilizado o *BeautifulSoup* e *Requests*. Será esperado do usuário o fornecimento de uma base de dados em *Comma Separated Values* (CSV)[13] para a avaliação de qualidade de maneira local, sem necessidade de raspagem ou captação de fontes externas.

Durante a etapa 4, utilizamos os *datasets* do Ibram exportados em massa na seção 4.3.2 deste trabalho. Realizamos 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.

5 RESULTADOS

O presente capítulo apresenta os índices de adequação de qualidade de dados obtidos a partir das 22 coleções museológicas sob gestão do Ibram, frente às 7 (sete) dimensões analíticas oriundas dos Capítulos do CCO, os quais passaram pelo processo de alinhamento do INBCM e CCO com o levantamento das regras herdadas neste processo (em 5.1), a luz do alinhamento, as coleções museológicas do Ibram tiveram seus elementos descritivos avaliados com os elementos do guia de catalogação CCO (em 5.2). O Índice de adequação focado no uso de vocabulários controlados também é apresentado (em 5.3). Apresenta ainda a ferramenta desenvolvida com interface para usuário com resultados e relatórios de qualidade de dados das coleções avaliadas (em 5.4).

[13]https://en.wikipedia.org/wiki/Comma-separated_values

5.1 Alinhamento de padrões

O alinhamento ente o INBCM e CCO foi uma atividade crucial que viabilizou a aplicação das regras de catalogação do guia nos elementos de metadados do INBCM presentes nos acervos captados do Ibram. Como resultado deste alinhamento, como explanado na seção 4.3.1, pode ser observado no Quadro 11 abaixo.

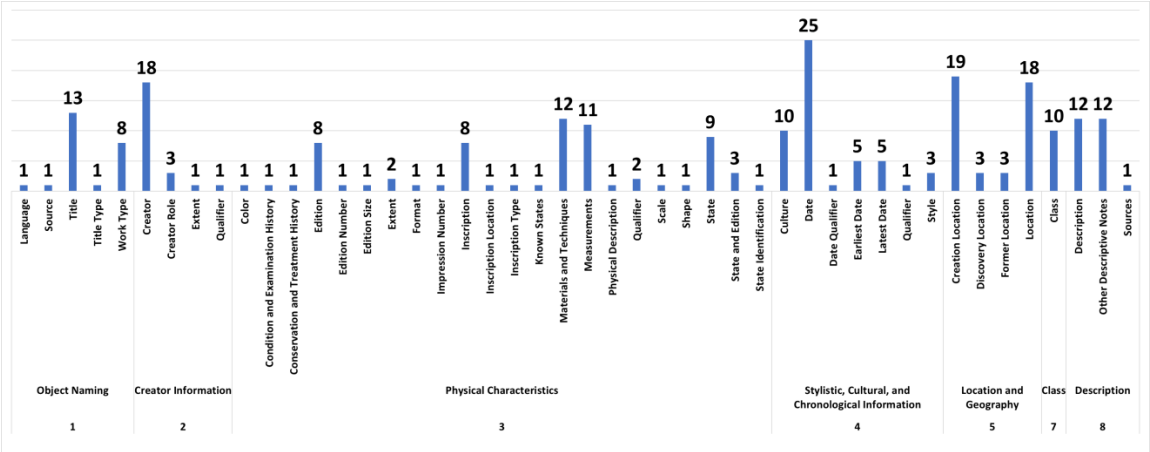
Quadro 11 – Alinhamento entre elementos descritivos – INBCM e CCO

Capítulo CCO	**Elemento CCO**	**Obrigatório CCO**	**Vocabulário Controlado CCO**	**Elemento INBCM**	**Obrigatório INBCM**
I-Part 2	*Work Type*	Sim	Sim	Denominação	Sim
I-Part 2	*Title*	Sim	Não	Título	Não
II-Part 2	*Creator*	Sim	Sim	Autor	Sim
III-Part 2	*Measurements*	Sim	Sim	Dimensões	Sim
III-Part 2	*Materials and Techniques*	Sim	Sim	Material/Técnica	Sim
III-Part 2	*Physical Description*	Não	Sim	Estado de Conservação	Sim
III-Part 2	*Inscription*	Sim	Sim	Número de Registro	Sim
IV-Part 2	*Date*	Sim	Não	Data de Produção	Não
V-Part 2	*Creation Location*	Não	Sim	Local de Produção	Não
V-Part 2	*Location*	Sim	Sim	Situação	Sim
VII-Part 2	*Class*	Sim	Sim	Classificação	Não
VIII-Part 2	*Description*	Não	Não	Resumo Descritivo	Sim
VIII-Part 2	*Other Descriptive Notes*	Não	Não	Condições de Reprodução	Sim
IX-Part 2	*View Description*	Sim	Não	Mídias Relacionadas	Não
NA	NA	NA	NA	Outros Números	Não
Legenda: NA (não aplicado)					

Fonte: elaborado pelo autor

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 3.

Figura 3 – Regras de catalogação CCO alinhadas ao INBCM



Fonte: elaborado pelo autor

Contudo, dentre o conjunto de regras mapeadas, foram elencadas apenas as regras pertencentes aos elementos descritivos alinhados com o INBCM (Quadro 11), que não apresentassem fator subjetivo e que, portanto, pudessem inviabilizar tecnicamente a avaliação por algoritmo computacional. As regras mapeadas foram então implementadas em *Python* seguindo os fundamentos do campo da catalogação descritiva (GILLILAND, 2016) quanto às orientações acerca do uso de padrões para tratamento nos dados, a saber: padrão de conteúdo de dados e padrão de valor de dados nos acervos dos museus envolvidos na análise.

No caso do padrão de conteúdo de dados, a avaliação do dado foi implementada no algoritmo a partir de regras da *regex*, conforme referenciadas na Seção 2.3.1. No *Python*, há uma biblioteca chamada *re*[14] que trabalha bem com *regex*, e esta foi utilizada no algoritmo do experimento.

Já para o padrão de valor de dados, a avaliação da utilização de vocabulário controlado foi feita a partir dos dados disponibilizados pela API do Tainacan, disponível no painel de exportação com nome “API do Tainacan em formato JSON". Essa API disponibiliza dados para além dos elementos de metadados do INBCM, e que indicam se a configuração do elemento de metadado é do tipo taxonomia para uma determinada coleção. O Quadro 12 apresenta as regras de catalogação que foram implementadas pelo algoritmo, incluindo a linguagem formal em *regex.*

[14] https://docs.python.org/3/library/re.html

Quadro 12 – 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 pelo autor

Vale acrescentar, como já mencionado, que a análise do guia CCO resultou na identificação de 244 regras, das quais 122 eram distintas. No entanto, apenas 17 dessas regras foram consideradas viáveis para automação por software, indicando a necessidade de avaliar outras metodologias para ampliar o número de regras avaliáveis. Como exemplo do critério de seleção das regras viáveis, pode-se observar a regra "*Caso o local não tenha nome, registre o local mais próximo*" ilustra as limitações da automação atual. A tarefa de registrar o local mais próximo requer uma análise subjetiva que não pode ser realizada de forma automatizada com o uso de software atualmente disponível. No entanto, é importante destacar que a automação de análise de regras é uma área em constante evolução e que novos métodos e técnicas estão sendo desenvolvidos a cada dia. Portanto, é possível que, com o uso de outras metodologias de análise, mais regras possam ser empregadas para avaliação automatizada no futuro.

5.2 Índice de adequação das coleções do Ibram

Os índices de adequação das coleções selecionadas podem ser observados na Figura 4 com o índice de adequação das coleções de caráter museal do Ibram.

Cada linha do mapa de calor representa uma coleção e cada coluna representa uma dimensão do CCO. Os quadrados mais claros representam as dimensões com menor índice de adequação e as mais escuras as com maior índice de adequação. Os valores representam porcentagens de 0 (zero) a 100.

Na dimensão *Object Naming*, algumas coleções se destacaram com índice de adequação acima de 90%, como o Museu solar Monjardim, Museu da Abolição, Museu Histórico Nacional - coleção de moedas de ouro e o Museu das Bandeiras. Contribuíram para essa adequação, as regras “evite abreviações”, “use o mesmo idioma do catálogo” e “não pode ficar vazio” nos elementos *Title* e *Work Type*. Como adequação de qualidade negativa, destaca-se as coleções com índice de adequação inferior a 40%, como o Museu das Missões (36%) e o Museu de Arte Religiosa e Tradicional (20%). Esses valores são justificados devido à ausência do valor de dado no elemento *Work Type* na descrição dos recursos. Outro destaque negativo é a completa ausência de título nas coleções do Museu de Arqueologia de Itaipu, Museu de Arte Sacra da Boa Morte, Museu Regional Casa dos Ottoni e Museu Casa da Hera - Indumentárias.

Na dimensão *Creator information*, o destaque é dado às coleções do Museu Regional de Caeté, Museu do Ouro, Museu Casa da Princesa, Museu Solar Monjardim, Museu Casa de Benjamin Constant, Museu de Arte Sacra da Boa Morte e Museu Regional de São João del-Rei com o índice de qualidade em 100%, atendendo completamente as regras “evite abreviações”, “não pode ficar vazio” e “faça uso de vocabulário controlado”. Como adequação de qualidade negativa, destaca-se as coleções com índice em 0% do Museu Histórico Nacional e do Museu de Arqueologia de Itaipu. Esses valores são justificados devido à ausência do valor de dado no elemento *Creator* na descrição dos recursos.

Na dimensão *Physical Characteristics*, observa-se que o maior índice de adequação chegou a 31% e 29% respectivamente para as coleções Museu das Missões e Museu Histórico Nacional, sendo que grande parte das coleções ficou igual ou abaixo de 27%. O baixo índice de adequação nesta dimensão ocorreu devido às regras de catalogação “faça uso de vocabulário controlado”, “medidas incluem duas casas decimais para medidas métricas”, “abrevie unidades métricas de acordo com o Sistema Internacional” e “não use capitalização” com índices inferiores a 10%.

Na dimensão *Stylistic, Cultural, and Chronological Information* o destaque é dado às coleções que apresentaram o índice de adequação superior a 50%, sendo

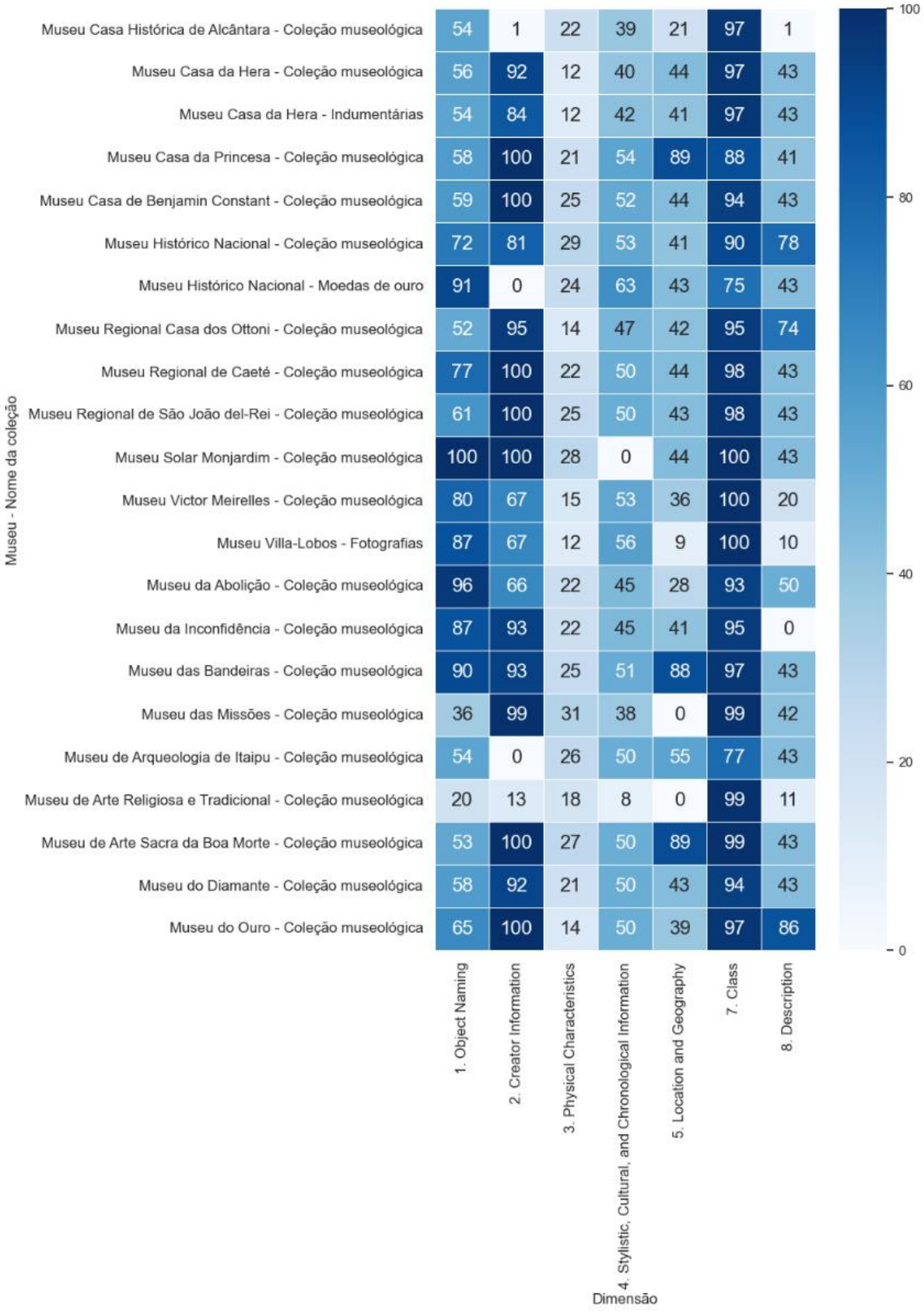
estas as coleções do Museu Histórico Nacional com 63% e do Museu Villa-Lobos com 56%. Na contramão, grande parte das coleções ficaram iguais abaixo de 54% por não ter atendido as regras elencadas para esta dimensão, tais como “anos com menos que 4 (quatro) dígitos, inserir 0 (zero) a esquerda”, “use traço para separar intervalo de anos”, “siga padrão pra registro de dia, mês e ano de data”, “siga padrão para registro de hora, minutos e segundos” e “não utilize apóstrofo” no elemento *Date*.

Na dimensão *Location and Geography*, algumas coleções se destacaram com índice de adequação superior a 80% por terem atendido consideravelmente as regras elencadas para esta dimensão, a saber: “capitalize as iniciais de nomes próprios e a primeira letra do texto; para outros termos, use apenas letras minúsculas”; “use o mesmo idioma”; e “evite abreviações” nos elementos *Creation Location* e *Location*. Foram os casos das coleções do Museu de Arte Sacra da Boa Morte e do Museu Casa da Princesa, ambos com 89% no índice de qualidade, e do Museu das Bandeiras com 88%. Nestas coleções, houve perda de pontos pelo não cumprimento da regra “faça uso de vocabulário controlado” no elemento *Location*. Como adequação de qualidade negativa, destaca-se as coleções do Museu das Missões e do Museu de Arte Religiosa e Tradicional, ambas com adequação em 0%. Esse índice é justificado pela ausência do valor de dado nos elementos *Creation Location* e *Location* na descrição dos recursos.

Na dimensão *Class* pode-se observar os índices de qualidade mais altos dentre as dimensões avaliadas no experimento com índices acima de 75%, destacando 19 das 22 coleções avaliadas acima de 90% por terem atendido consideravelmente as regras “evite abreviações”, “não pode ficar vazio” e “faça uso de vocabulário controlado” no elemento *Class*.

Na dimensão *Description*, as coleções pertencentes ao Museu do Ouro (86%), ao Museu Histórico Nacional (78%) e ao Museu Regional Casa dos Ottoni (74%) obtiveram maior índice de qualidade. Tal cenário evidenciou-se devido ao cumprimento das regras “evite abreviações”, “capitalize as iniciais de nomes próprios e a primeira palavra” e “use o mesmo idioma”, perdendo pontos, por outro lado, na regra “faça uso de vocabulário controlado”. Como adequação de qualidade negativa, destaca-se a completa ausência de valores nessa dimensão para a coleção Museu da Inconfidência (0%). Por fim, as demais coleções também foram impactadas no índice de qualidade devido à ausência de valor de dado nos elementos descritivos dos recursos.

Figura 4 – Diagnóstico da adequação dos metadados dos museus Ibram às dimensões CCO

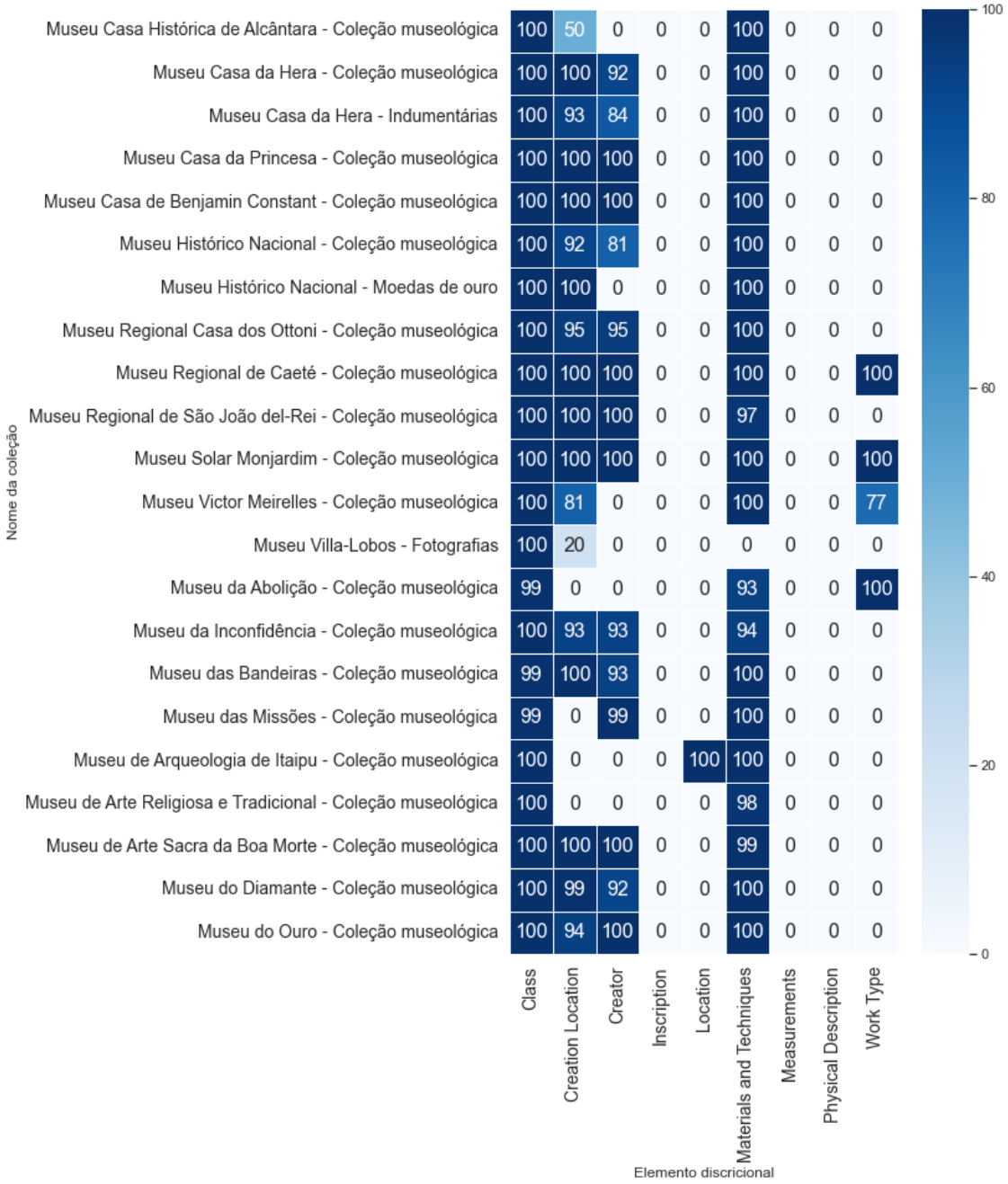


Fonte: elaborado pelo autor

5.3 Índice de adequação das coleções do Ibram frente ao uso de vocabulários controlados

A taxa de adequação dos elementos de metadados das bases de dados do Ibram que tiveram a indicação de uso de taxonomia pode ser observada a seguir na Figura 5. Observam-se nas linhas, as coleções museológicas disponíveis online pelo Ibram; e nas colunas, os elementos de metadados alinhados com as regras de catalogação de uso de vocabulário controlado à luz do CCO. Cada célula apresenta a taxa de adequação das coleções à regra de uso de vocabulário controlado, com valores indo de 0 (zero) - onde foi observada a completa inadequação da coleção à regra; a 100 - em que houve a completa adequação da regra pela coleção em questão. Assim, as cores mais claras representam menores taxas de adequação e as mais fortes maiores taxas de adequação.

Figura 5 – Adequação de coleções Ibram ao uso de vocabulário controlado



Fonte: elaborado pelo autor

Pode-se destacar o maior índice do uso de vocabulário controlado no elemento de metadado *Class*, sendo a taxa de adequação em 99% nas coleções dos museus da Abolição, Bandeiras e Missões. O restante das coleções com taxa de adequação em 100%.

O elemento *Creation Location* apresentou menor taxa de adequação nas coleções dos museus da Abolição, Missões, Arqueologia de Itaipu e de Arte Religiosa

e Tradicional com adequação em 0%, devido ao elemento não apresentar o uso de taxonomia nessas coleções. Ainda neste elemento, podemos observar que as coleções dos museus Casa Histórica de Alcântara, Casa da Hera, Histórico Nacional – Coleção Museológica, Regional Casa dos Ottoni, Vitor Meirelles, Villa-Lobos, do Diamante e do ouro apresentaram os usos de taxonomia, porém com elementos de metadados com valores vazios. Neste caso, sem valor taxonômico informado. Os demais apresentaram o uso de taxonomia e todos os elementos haviam valores informados.

Para o elemento *Creator*, puderam-se observar valores superiores a 81% com exceção das coleções dos museus Casa Histórica de Alcântara, Museu Histórico Nacional – Moedas de Ouro, Victor Meirelles, Villa-Lobos, da Abolição, de Arqueologia de Itaipu e de Arte Religiosa e Tradicional com taxa de adequação em 0%.

Destacam-se negativamente os elementos *Inscription*, *Measurements* e *Physical Description* com a completa inadequação em todas as coleções.

O elemento *Location* apresentou inadequação em todas as coleções com exceção do Museu da Arqueologia de Itaipu com 100% de adequação.

Destaca-se positivamente também o elemento *Materials and Techniques* com adequação em 93% na coleção do Museu da Abolição e 94% no Museu da Inconfidência, possuindo as demais coleções taxas superiores a 97%, exceto para a coleção do Museu Villa-Lobos com adequação em 0%.

Outro elemento que apresentou alguma taxa de adequação é o *Work Type* com 0% na maioria das coleções, exceto para as coleções do Museu Vitor Meirelles com 77% de adequação e 100% para as coleções dos museus Regional do Caeté, Solar Monjardim – Coleção Museológica, e o da Abolição.

Dentre os elementos discricionais ora elencados, os que apresentaram maior taxa de adequação foram *Class*, *Creation Location*, *Creator* e *Materials and Techniques*. E os elementos com menor taxa de adequação foram *Inscription*, *Location*, *Measurements*, *Physical Description* e *Work Type*.

Por fim, cabe destacar, no caso do elemento *Class*, a sua adequação se deve ao fato ao uso dos 2 (dois) instrumentos de linguagem documentária praticados pelos museus sob gestão do Ibram, o tesauro com versão mais recente (publicado em 2016), e que substituiu o anterior denominado ‘tesauro para acervos museológicos’, e o Tesauro de Objetos do Patrimônio Cultural dos Museus Brasileiros, concebido e disponibilizado pela Secretaria de Cultura do Rio de Janeiro e pelo Museu Histórico

da Cidade, sob a coordenação técnica de Helena Dodd Ferrez, e com padrão ISO 2788:1986. Nesse sentido, os dois tesauros representam terminologias de acervos de caráter histórico e artístico por meio das quais o profissional da informação classifica os itens de coleção por temáticas, assuntos ou contextos de uso. É importante destacar que o uso desses tesauros é apenas recomendado informalmente pelo Ibram, não havendo uma normativa explícita sugerindo seu uso e nem especificando qual das versões deveria ser adotada, se a de 1987 ou a de 2006. Sabe-se que atualmente ambas se encontram em uso pelos museus e possuem significativas diferenças entre si.

5.4 *DataQ Culture:* ferramenta de avaliação de qualidade de dados

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 dos metadados de acervos culturais. Assim, o desenvolvimento da *DataQ Culture* (COELHO JÚNIOR, 2023) surge para preencher essa lacuna.

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, nas Figuras 6 e 7 são apresentadas a interface inicial e a de envio de arquivo.

Figura 6 – Página inicial da ferramenta

Fonte: elaborado pelo autor

Figura 7 – Interface de envio de base de dados para avaliação

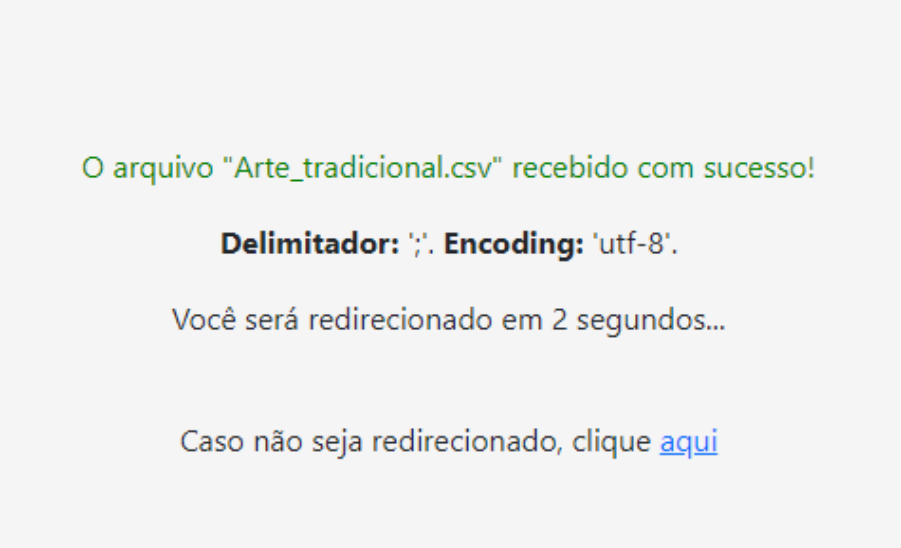
Fonte: elaborado pelo autor

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 Excel 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*[15] (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 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. Desta forma, uma terceira tela foi

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

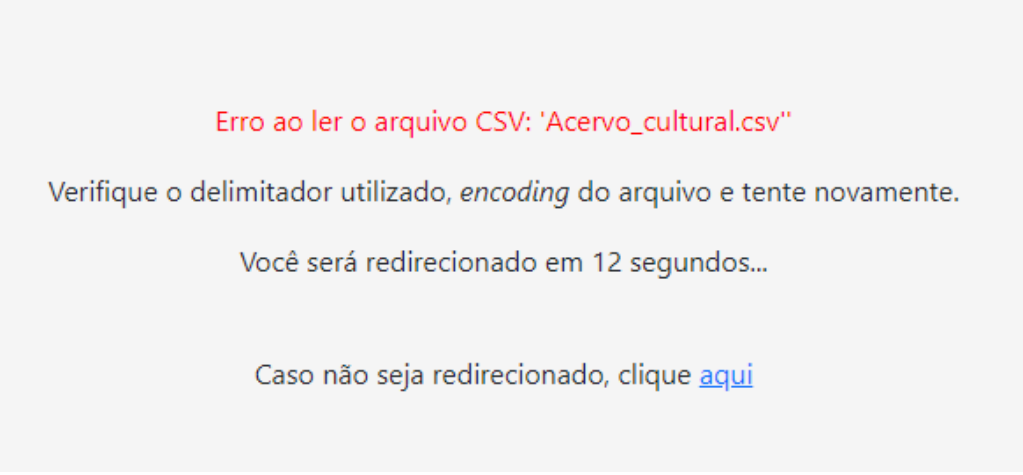
necessária para confirmar e mostrar ao usuário as inferências feitas pela ferramenta, como pode ser observado na Figura 8, quando a identificação é feita com sucesso; e na Figura 9, quando há algo de errado no arquivo do usuário e não é possível realizar a identificação.

Figura 8 – Tela com sinalização de sucesso para identificar *encoding* e delimitador



Fonte: elaborado pelo autor

Figura 9 – Tela com sinalização de erro para identificar *encoding* e delimitador

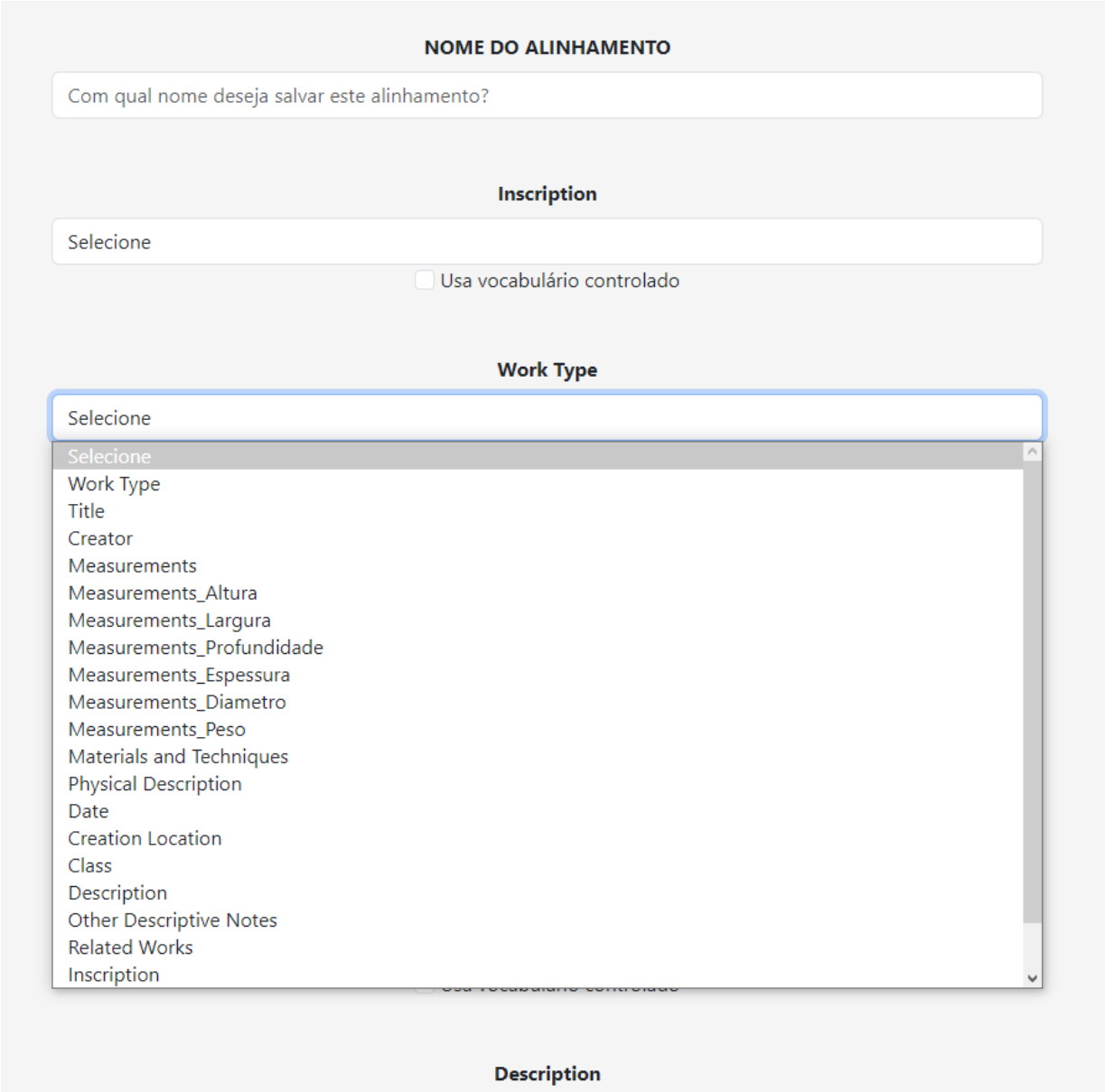


Fonte: elaborado pelo autor

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 10. 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. Essa indicação de uso de vocabulário controlado na exploração semiautomática das bases de dados realizada na seção 4.3.2 foi obtida por meio de metadados administrativos das plataformas Tainacan de cada coleção.

Figura 10 – Tela de alinhamento entre elementos discricionais da base do usuário com os elementos discricionais do CCO



Fonte: elaborado pelo autor

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 11.

Figura 11 – Tela de alinhamento com indicação de alinhamento já existente

Fonte: elaborado pelo autor

Após o alinhamento das bases, uma tela de carregamento aparece enquanto os dados são avaliados, conforme mostrado na Figura 12.

Figura 12 – Tela de espera enquanto os dados são avaliados

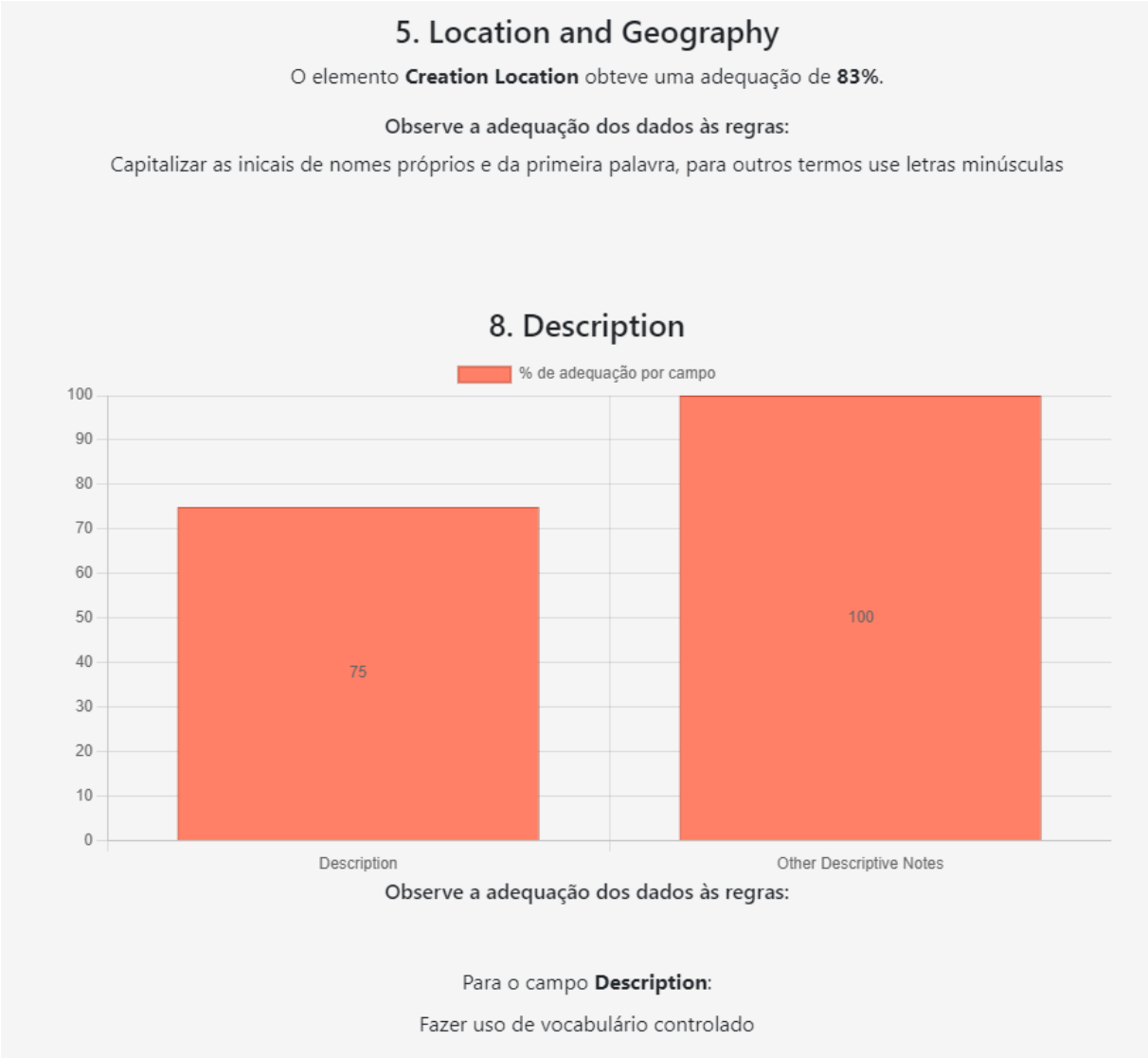
Fonte: elaborado pelo autor

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. Essas características podem ser visualizadas nas Figuras 13 e 14. Por fim, no final da página, é possível baixar uma planilha do Excel com todos os valores avaliados e a indicação se estava adequado à regra ou não, como demonstrado na Figura 15.

Figura 13 – Tela principal com taxa de adequação de coleção avaliada

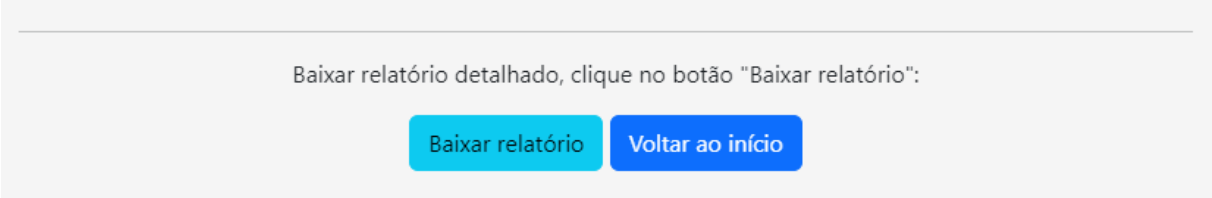
Fonte: elaborado pelo autor

Figura 14 – Regras indicadas para elementos discricionais que não alcançaram 100% de adequação



Fonte: elaborado pelo autor

Figura 15 – Opção de baixar relatório completo em Excel



Fonte: elaborado pelo autor

A ferramenta desenvolvida neste projeto (COELHO JÚNIOR, 2023) 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. Por outro lado, também permite que gestores de acervos e coleções façam a validação de seus dados sem maiores dificuldades.

6 DISCUSSÃO DOS RESULTADOS

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 à qualidade de dados desta ampla região parecem ainda não terem encontrado espaço significativo na pauta de agenda de pesquisa e desenvolvimento acadêmico.

A pesquisa confirmou ainda que padrões de documentação atuais, que promovem qualidade de dados e, por consequência, recuperação da informação (ENGLISH, 1999; BATINI; SCANNAPIECA, 2006; BACA et al., 2006; SIQUEIRA et al., 2021) mais eficiente 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 MARC (*Machine Readable Cataloging*) e o AACR (*Anglo-American Cataloguing Rules*), que são, respectivamente, formatos padronizados para a codificação de dados bibliográficos em formato de máquina, e regras de catalogação para a construção adequada de metadados bibliográficos, incluindo autor, título, editora, ano de publicação, entre outros. O RDA (*Resource Description and Access*) é outro padrão de referência importante, que estabelece diretrizes para a descrição e acesso a recursos culturais analógicos e digitais, 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 VRA Core (*Visual Resources Association Core Categories*) é um conjunto de categorias de metadados para a descrição de recursos visuais, como imagens, vídeos e objetos 3D. O LIDO

(*Lightweight Information Describing Objects*) 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. O EAD (*Encoded Archival Description*) é um padrão para a codificação de guias de arquivos em XML, mantido pela Biblioteca do Congresso em parceria com a Sociedade Americana de Arquivistas. Finalmente, o CCO (*Cataloging Cultural Objects*) fornece diretrizes para a seleção, a organização e a formatação de dados usados para preencher registros de catálogos, com base em categorias principais no CDWA (*Categories for the Description of Works of Art*) e no *VRA Core*. É importante destacar que existem outros padrões de referência, além dos discutidos acima, que também podem ser considerados para a avaliação de conformidade dos dados (BACA et al., 2006, LEMOS; COELHO-JÚNIOR; CARMO, 2021; HARPRING, 2022).

Nesta direção, destaca-se o CCO, padrão que apresenta conceitos genéricos que podem ser empregados a qualquer conjunto de metadados (BACA *et al*., 2006), inclusive, com os elementos descritivos do INBCM, conforme se comprovou na ação de alinhamento (Quadro 11).

Desta forma, os elementos do INBCM podem servir de base para a denominação de um conjunto de categorias que podem ser usadas para criar uma estrutura no formato de campos em um banco de dados ou de propriedades de um recurso em um modelo RDF, por exemplo, o que reforça o aspecto do tratamento para comunicação de dados em ambiente digital (GILLILAND, 2016). Embora uma estrutura de dados seja o primeiro passo lógico no desenvolvimento de esquemas de metadados, uma estrutura por si só não alcançará uma alta taxa de consistência descritiva por parte dos catalogadores, muito menos uma alta taxa de recuperação por parte dos usuários finais (BACA *et al*., 2006), sendo necessário, portanto, outros meios de tratamento sintático e semântico a os dados.

Vale destacar que o INBCM não detalha e nem orienta a respeito dos aspectos sintáticos ou semânticos para os elementos de descrição sugeridos, deixando bastante em aberto a forma como esses elementos devem ser implementados pela instituição. Assim sendo, padrões que regem a sintaxe e a semântica da linguagem (valor de dados) empregada no sistema de informação e sua seleção, organização e formatação (conteúdo de dados) são dois outros tipos de padrões que devem ser usados em conjunto com uma estrutura de dados acordada para a aplicação. Sabe-

se que trabalhos no desenvolvimento de padrões para valores de dados (LEMOS; COELHO-JUNIOR; CARMO, 2021; TRUST, 2022) são muito mais evidentes do que para conteúdo de dados, normalmente na forma de tesauros, vocabulários controlados e ontologias.

Com a avaliação realizada, pode-se perceber que os museus vinculados ao Ibram até possuem a prática de uso de linguagens documentárias (taxonomias, no Tainacan) para preenchimento dos valores de dados para entidades associadas à classificação de temas, assuntos ou contextos de uso (MARTINS *et al*., 2021), conforme se comprovou nos bons índices de adequação da dimensão CCO *Class* (classificação no INBCM) nas coleções avaliadas (todas acima de 70%, conforme Figura 4). Porém, o INBCM não faz menção a qualquer orientação acerca de qual versão do tesauro usar, podendo acarretar significativas diferenças terminológicas no processo de indexação em âmbito geral dos museus, o que poderia acarretar dificuldades numa solução de agregação de dados projetada. Adicionalmente, as taxonomias do Ibram não estão representadas para consumo computacional, isto é, as descrições produzidas por meio desses vocabulários controlados não estão com as suas terminologias configuradas a partir de um identificador único no formato *Uniform Resource Identifier* (URI), como se recomenda no guia CCO. Tal funcionalidade consegue estabelecer interligações e anotações sobre dados sob licença aberta (princípio dos dados abertos ligados), o que os confere possibilidades de reu so e interoperabilidade com outros conjuntos de dados na internet no âmbito do patrimônio cultural (LEMOS; COELHO-JUNIOR; CARMO, 2021).

Com a variedade de vocabulários controlados existentes para diversos domínios e tipos de objetos culturais, há a necessidade de conectar e realizar o intercâmbio entre descrições que usam diferentes vocabulários controlados. Conforme apresentado por Kapidakis (2012) e Coburn et al. (2010), o alinhamento e o uso de padrões em diferentes acervos culturais melhoram a catalogação, o que por sua vez melhora o acesso à informação para o usuário final. No entanto, o grande espectro de acervos culturais possui características diferentes, o que leva ao uso de diferentes vocabulários controlados. Ainda neste cenário de utilização de padrões adequados de catalogação, há desafios na agregação desses diferentes acervos em um único repositório, pois é preciso considerar o alinhamento entre as diferentes percepções que podem existir sobre o conjunto de dados, como demonstrado na Figura 5. Assim, pode-se observar os desafios que o Ibram enfrenta na missão de

agregar e interligar os dados de diferentes museus no Brasil (LEMOS; COELHO JÚNIOR, 2023).

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 Westbrook (2012), Palavitsinis (2013), Moulaison (2015), Stephan, Beat e Angelina (2018) e Martins et al. (2021). 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.

Por outro lado, pode-se destacar os estudos de (CANDELA et al., 2021), (CANDELA, 2023) e de (LEMOS; COELHO JÚNIOR, 2023) com uso de procedimentos metodológicos automáticos ou semiautomáticos para o processamento da avaliação diagnóstica.

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).

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 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.

Porém, como ponto de partida para o desenvolvimento de um modelo avaliativo, é necessário explorar as metodologias de criação, extração e validação de metadados de forma automática ou semiautomática. Algumas dessas metodologias incluem o processamento de texto, como expressões regulares (*regex*), reconhecimento ótico de caracteres (OCR), reconhecimento de entidades nomeadas (NER) e processamento de linguagem natural (NLP). Observou-se ainda o uso de metodologias baseadas em aprendizado de máquina, como aprendizado profundo e redes neurais, bem como ferramentas de limpeza e transformação de dados, como OpenRefine e PDFBox. Essas diversas tecnologias encontram aplicação em múltiplas áreas, desde as Ciências Naturais e Biológicas até a Ciência da Computação e da Informação.

Com destaque a tecnologia *regex*, 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 guia 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. Foi disponibilizada ainda uma aplicação de avaliação de qualidade de dados para que diferentes usuários possam realizar a mesma avaliação em seus acervos, 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.

Finalmente, a presente dissertação cumpre o papel de responder a questão de pesquisa, indicando como melhorar a qualidade de dados em acervos culturais - baseado em guia de referência - bem como o objetivo geral, com a proposição e entrega de uma aplicação para avaliação semiautomatizada que possibilite a otimização da qualidade dos dados em acervos de instituições de patrimônios culturais, viabilizando a melhora da qualidade de dados em acervos culturais.

7 CONCLUSÕES E TRABALHOS FUTUROS

À luz dos resultados foi observado o total de 17 trabalhos, que entre 2012 a 2023 relataram acerca da avaliação da qualidade de dados em instituições no domínio da cultura. Com isso, ficou claro que há pouca evidência de um processo de garantia de qualidade de metadados testado, que comprove sua eficácia em um ou mais repositórios. Ressalta-se ainda a escassez de procedimentos que utilizem um modelo de catalogação de referência na área da cultura para fundamentar uma avaliação de qualidade de dados em bases de dados. Tal resultado respalda a hipótese da pesquisa de 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.

Por outro lado, nesta dissertação, a partir do alinhamento entre os elementos descritivos do INBCM e do CCO, pôde-se realizar a implementação de uma porção de regras de catalogação do guia CCO em *regex* por meio 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.

A dissertação trouxe e reforçou as contribuições do guia CCO conjuntamente com os princípios teórico-metodológicos da Ciência da Informação em relação ao tratamento adequado de bases de dados. O profissional da informação envolvido no processo de modelagem de metadados geralmente utiliza padrões terminológicos para prover um vocabulário comum que descreva uma variedade de estruturas de dados capazes de satisfazer a várias comunidades, e, geralmente, são estruturados

seguindo modelos para tratamento dos dados, o que redunda em normalização, qualidade e intercâmbio de suas descrições.

Com o aporte do CCO, incluindo seus grupos de informação a partir de 9 capítulos, os elementos descritivos do INBCM podem se tornar um esquema formal de metadados em pesquisas futuras. Adicionalmente, sistemas de organização do conhecimento contemporâneos (exs.: *Simple Knowledge Organization System* – SKOS) também são recomendados pelo CCO nessa perspectiva de modelagem, cujas terminologias (padrão de valor de dados) apresentam em suas estruturadas um URI semântico para estabelecer interligações e anotações sobre dados sob licença aberta, o que os confere possibilidades de reuso e interoperabilidade (padrão de comunicação de dados) com outros conjuntos de dados associados ao campo do patrimônio cultural. 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.

A avaliação diagnóstica semiautomática desenvolvida permitiu aferir nas coleções museológicas do Ibram que os dados das coleções 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. Recomenda-se, portanto, que práticas de catalogação maduradas oriundas de modelos de referência sejam incorporadas na modelagem de metadados das bases de dados dos museus sob gestão do Ibram, visando qualificar seus atuais padrões de documentação por meio de instrumentos de organização da informação mais sofisticados e orientados para usuários finais de sistemas de informação.

Sobre o modelo de avaliação desenvolvido, é importante destacar que, embora existam diversas metodologias de extração e validação de informação, como inteligência artificial, aprendizado de máquina, modelos estatísticos e processamento de linguagem natural, entre outras, este trabalho utilizou uma metodologia baseada em regras chamada *regex*. É preciso ressaltar que essa abordagem possui limitações em termos de capacidade e aplicação dessa tecnologia. Portanto, como trabalho futuro, pretende-se agregar modelos de inteligência artificial, como modelos de

processamento de linguagem natural, reconhecimento de entidades nomeadas e *machine leaning* para cobrir mais regras de catalogação do CCO no processo avaliativo, bem como a exploração de metadados de assunto, visto que não é considerado nos elementos de descrição para identificação do bem cultural de caráter museológico do INBCM, sugerindo que esse tipo de representação temática não é relevante para o contexto dos museus vinculados ao Ibram, ou talvez não tenha sido explorado o seu potencial pelos especialistas em documentação visando os SRIs.

Assim essas limitações se mostram propícios a serem explorados em trabalhos futuros.

Conclui-se, portanto, que o modelo de avaliação de qualidade de dados proposto neste trabalho, 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

ABADAL, Ernest; CODINA Luis. **Bases de datos documentales**: características, funciones y método. Madrid: Síntesis, 2005.

ABBAS, June. **Structures for organizing knowledge**: exploring taxonomies, ontologies, and other schema. New York: Neal-Schuman Publishers, 2010.

ACERVO EM REDE – **INSTITUTO BRASILEIRO DE MUSEUS – Ibram**. 2023. Disponível em: https://antigo.museus.gov.br/acessoainformacao/acoes-e-programas/acervo-em-rede/. Acesso em: 12 mar. 2023.

ANSI/NISO Z39.19-2005 (R2010). Guidelines for the construction, format, and management of monolingual controlled vocabularies. Baltimore: NISO Press, p. 184, 2005.

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.

BALLOU, Donald; WANG, Richard; PAZER, Harold; TAYI, Giri Kumar. Modeling Information Manufacturing Systems to Determine Information Product Quality. **Management Science**, *[S. l.]*, v. 44, n. 4, p. 462–484, 1998.

BARBIERI, Carlos. **BI - business intelligence: modelagem e tecnologia**. Rio de Janeiro: Axcel Books, 2001.

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. *Em*: NESI, Paolo; SANTUCCI, Raffaella (org.). **Information Technologies for Performing Arts, Media Access, and Entertainment**. Lecture Notes in Computer Science Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. v. 7990p. 90–103. Disponível em: <http://link.springer.com/10.1007/978-3-642-40050-6_9>. Acesso em: 3 ago. 2022.

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.

BOUGHIDA, Karim. CDWA Lite for Cataloguing Cultural Objects (CCO): A New XML Schema for the Cultural Heritage Community. In: **Proceedings of the XVI International Conference of the Association for History and Computing**. Amsterdam: Royal Netherlands Academy of Arts and Sciences, 2005, p. 49–56. Disponível em: <https://repository.ubn.ru.nl/handle/2066/32358>. Acesso em: 12 mar. 2023.

BRACK, Arthur; HOPPE, Anett; BUSCHERMÖHLE, Pascal; *et al*. Cross-domain multi-task learning for sequential sentence classification in research papers. *In*: **Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries**.

CologneGermany: ACM, 2022, p. 1–13. Disponível em: <https://dl.acm.org/doi/10.1145/3529372.3530922>. Acesso em: 5 mar. 2023.

BRASIL, Ministério da Cultura. **Política Nacional de Museus/organização e textos**, José do Nascimento Júnior, Mário de Souza Chagas. – Brasília: Minc, 2007.

BRASIL. MINISTÉRIO DO TURISMO. INSTITUTO BRASILEIRO DE MUSEUS – Ibram. **Museus Ibram – Instituto Brasileiro de Museus. Brasilia**, 2023. Disponível em: https://antigo.museus.gov.br/museus-ibram/. Acesso em: 12 mar. 2023.

BRYNJOLFSSON, Erik; MCAFEE, Andrew. **The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies**. [s.l.]: W. W. Norton & Company, 2014.

BUSH, Vannevar. **As We May Think**. The Atlantic. Disponível em: <https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881>. Acesso em: 6 mar. 2023.

C, Abhilash; MAHESH, Kavi. Graph Analytics Applied to COVID19 Karnataka State Dataset. *In*: **2021 The 4th International Conference on Information Science and Systems**. Edinburgh United Kingdom: ACM, 2021, p. 74–80. Disponível em: <https://dl.acm.org/doi/10.1145/3459955.3460603>. Acesso em: 5 mar. 2023.

CANDELA, Gustavo; ESCOBAR, Pilar; SÁEZ, María Dolores; MARCO-SUCH, Manuel. A Shape Expression approach for assessing the quality of Linked Open Data in libraries. **Semantic Web**, *[S. l.]*, p. 1–21, 2021.

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**. Copenhagen, 2005. Disponível em: <https://www.gbif.org/document/80509>. Acesso em: 28 jul. 2022.

CHARDONNENS, Anne; RIZZA, Ettore; COECKELBERGS, Mathias; *et al*. Mining user queries with information extraction methods and linked data. **Journal of Documentation**, v. 74, n. 5, p. 936–950, 2018.

CHARLES, Valentine; CLAYPHAN, Robina; ISAAC, Antoine. **Definition of the Europeana Data Model v5.2.8**. 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 jul. 2022.

CHARLES, Valentine; CLAYPHAN, Robina; ISAAC, Antoine. **Definition of the Europeana Data Model v5.2.8**. 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 jul. 2022.

CHEN, Min; MAO, Shiwen; LIU, Yunhao. Big Data: A Survey. **Mobile Networks and Applications**, v. 19, n. 2, p. 171–209, 2014.

CLEMENT, Tanya E.; CARTER, Daniel. Connecting theory and practice in digital humanities information work. **Journal of the Association for Information Science and Technology**, [S. l.], v. 68, n. 6, p. 1385–1396, 2017.

COBURN, Erin; LANZI, Elisa; O"KEEFE, Elizabeth; STEIN, Regine; WHITESIDE, Ann. The Cataloging Cultural Objects experience: Codifying practice for the cultural heritage community. **IFLA Journal**, *[S. l.]*, v. 36, n. 1, p. 16–29, 2010.

COELHO JÚNIOR, Abeil. DataQ-Culture. 2023. **GITHUB**. Disponível em: https://github.com/AbeilCoelho/DataQ-Culture. Acesso em: 05 mar. 2023.

CROCHEMORE, Maxime; RYTTER, Wojciech. **Text algorithms**. New York: Oxford University Press, 1994.

DE CLERCQ, Djavan; DIOP, Ndeye-Fatou; JAIN, Devina; *et al*. Multi-label classification and interactive NLP-based visualization of electric vehicle patent data. **World Patent Information**, v. 58, p. 101903, 2019.

DIJKSHOORN, Chris; JONGMA, Lizzy; OSSENBRUGGEN, Jacco Van; SCHREIBER, Guus; WEELE, Wesley Ter; WIELEMAKER, Jan. The Rijksmuseum collection as linked data. **Semantic Web**, v. 9, n. 2, p. 221-230, jan. 2018. Disponível em: <https://research.vu.nl/en/publications/the-rijksmuseum-collection-as-linked-data>. Acesso em: 13 jul. 2022.

DODEBEI, Vera Lúcia Doyle. **Tesauro: linguagem de representação da memória documentária**. Niterói: Intertexto; Rio de Janeiro: Interciência, 2002.

ECKERSON, Wayne W. DATA QUALITY AND THE BOTTOM LINE: Achieving Business Success through a Commitment to High Quality Data. **The Data Wharehouse Institute**, *[S. l.]*, 2002. Disponível em: <http://download.101com.com/pub/tdwi/Files/DQReport.pdf>. Acesso em: 28 jul. 2022.

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

EXPRESSÃO REGULAR. Em: **Wikipédia, a enciclopédia livre**, 2020. Disponível em: <https://pt.wikipedia.org/wiki/Expressão_regular>. Acesso em: 22 jul. 2022.

FELL, Michael; CABRIO, Elena; TIKAT, Maroua; *et al*. The WASABI song corpus and knowledge graph for music lyrics analysis. **Language Resources and Evaluation**, 2022. Disponívelem: <https://link.springer.com/10.1007/s10579-022-09601-8>. Acesso em: 5 mar. 2023.

FENLON, Katrina; EFRON, Miles; ORGANISCIAK, Peter. Tooling the aggregator"s workbench: Metadata visualization through statistical text analysis: 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.

FERRITER, Meghan. **Integrating Wikidata at the Library of Congress | The Signal**. 2019. Disponível em: <www.blogs.loc.gov/thesignal/2019/05/integrating-wikidata-at-the-library-of-congress/>. Acesso em: 21 jul. 2022.

FINK, Eleanor E. **American Art Collaborative (AAC) Linked Open Data (LOD) Initiative Releases** "Overview and Recommendations for Good Practices". 2018. 80 p. Disponível em: <https://repository.si.edu/bitstream/handle/10088/106410/OverviewandRecommendationsAccessible.pdf>. Acesso em: 02 jul. 2022.

FRANCISCO-REVILLA, Luis; TRACE, Ciaran B.; LI, Haoyang; BUCHANAN, Sarah A. Encoded Archival Description: Data Quality and Analysis: 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.

FREY, Carl Benedikt; OSBORNE, Michael A. The future of employment: How susceptible are jobs to computerisation? **Technological Forecasting and Social Change**, v. 114, p. 254–280, 2017.

FRIEDL, Jeffrey E. F. **Mastering Regular Expressions**. [s.l.]: O'Reilly Media, Inc., 2002.

GAONA GARCÍA, Paulo Alonso; FERMOSO GARCÍA, Ana; UNIVERSIDAD PONTIFICIA DE SALAMANCA; SÁNCHEZ ALONSO, Salvador; UNIVERSIDAD DE ALCALÁ. ExploringtheRelevanceof Europeana Digital Resources: Preliminary Ideas on Europeana MetadataQuality. **Revista Interamericana de Bibliotecología**, *[S. l.]*, v. 40, n. 1, p. 59–69, 2017.

GETTY, Getty Research Institute. **Art & Architecture Thesaurus® Online**. 2017. Disponível em: <https://www.getty.edu/research/tools/vocabularies/aat/>. Acesso em: 1 ago. 2022.

GIL, Yolanda; PIERCE, Suzanne A.; BABAIE, Hassan; *et al*. Intelligent systems for geosciences: an essential research agenda. **Communications of the ACM**, v. 62, n. 1, p. 76–84, 2018.

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: 22 jul. 2022.

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

GOV.BR. **Acervo em Rede e Projeto Tainacan. Ministério do Turismo - Instituto Brasileiro de Museus (Ibram)**, 2021. Disponível em: <https://www.gov.br/museus/pt-br/acesso-a-informacao/acoes-e-programas/acervo-em-rede-e-projeto-tainacan>. Acesso em: 18 jul. 2022.

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 ago. 2022.

HARPRING, Patricia. **Metadata Standards Crosswalks**. 2022. Disponível em: <https://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html#endnote1CCO>. Acesso em: 17 jul. 2022.

HENDLER, Jim; ALLEMANG, Dean; GANDON, Fabien. **Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL**. 3. ed. New York, NY, USA: ACM, 2020.

HJØRLAND, Birger. Semantics and Knowledge Organization. **Annual Review of Information Science and Technology**, v. 41, p. 367–405, 2007.

IBM, IBM Corporation. **Operational risk management in the world of big data: Unlocking the value of loss event data and driving the risk-aware enterprise**. 2014. Disponível em: <https://www.ibm.com/downloads/cas/716VKRPO>. Acesso em: 7 ago. 2022.

IBRAM, INSITUTO BRASILEIR DE MUSEUS. **Sobre o órgão – Instituto Brasileiro de Museus – Ibram**. 2023. Disponível em: <https://www.gov.br/museus/pt-br/acesso-a-informacao/institucional/sobre-o-orgao>. Acesso em: 12 mar. 2023.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS (IFLA). **Declaração dos Princípios Internacionais de Catalogação**. Haia, 2016. Disponível em: <https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/icp/icp_2016-pt.pdf>. Acesso em: 22 jul. 2022.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS (IFLA). **Statement on libraries and artificial intelligence**. 2020. Disponível em: <https://repository.ifla.org/handle/ 123456789/1646>. Acesso em: 22 jul.2022.

JIANG, Congfeng; LIU, Junming; OU, Dongyang; *et al*. Implicit Semantics Based Metadata Extraction and Matching of Scholarly Documents: **Journal of Database Management**, v. 29, n. 2, p. 1–22, 2018.

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

KAPIDAKIS, Sarantos. Comparing metadata quality in the Europeana context. In: **Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments**. Heraklion Crete Greece: ACM, 2012, p. 1–8. Disponível em: <https://dl.acm.org/doi/10.1145/2413097.2413129>. Acesso em: 10 fev. 2023.

KENETT, Ron; REDMAN, Thomas C. **The real work of data science: turning data into information, better decisions, and stronger organizations**. Hoboken, NJ: Wiley, 2019.

KIRCHHOFF, Agnes; BÜGEL, Ulrich; SANTAMARIA, Eduard; *et al*. Toward a service-based workflow for automated information extraction from herbarium specimens. **Database**, v. 2018, 2018. Disponível em: <https://academic.oup.com/database/article/doi/10.1093/database/bay103/5122758>. Acesso em: 5 mar. 2023.

KLEENE, Stephen Cole. Representation of Events in Nerve Nets and Finite Automata. *Em*: SHANNON, C. E.; MCCARTHY, J. (org.). **Automata Studies. (AM-34)**. Princiton: Princeton University Press, 1956. p. 3–42. Disponível em: <https://www.degruyter.com/document/doi/10.1515/9781400882618-002/html>. Acesso em: 8 ago. 2022.

KOLTAY, Tibor. Library and information science and the digital humanities: Perceived and real strengths and weaknesses. **Journal of Documentation**, v. 72, n. 4, p. 781–792, 2016.

LAGOZE, Carl; VAN DE SOMPEL, Herbert; NELSON, Michael; WARNER, Simeon. **Open Archives Initiative - Protocol for MetadataHarvesting - v.2.0**. 2002. Disponível em: <http://www.openarchives.org/OAI/openarchivesprotocol.html>. Acesso em: 1 ago. 2022.

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

LANCASTER, Frederick Wilfrid. **Vocabulary control for information retrieval**. 2ª ed. Virgínia: IRP, 1986. 270 p.

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**, v. 28, p. 1–22, 7 fev. 2023.

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

LEMOS, Daniela Lucas da Silva; SOUZA, Renato Rocha. Knowledge Organization Systems for the Representation of Multimedia Resources on the Web: A Comparative Analysis. **KNOWLEDGE ORGANIZATION**, v. 47, n. 4, p. 300–319, 2020.

LI, Yangfan; LI, Kenli; WEI, Wei; *et al*. CoRec: An Efficient Internet Behavior-based Recommendation Framework with Edge-cloud Collaboration on Deep Convolution Neural Networks. **ACM Transactions on Sensor Networks**, v. 19, n. 2, p. 1–28, 2023.

LIAO, Xiaofeng; ZHAO, Zhiming. Unsupervised Approaches for Textual Semantic Annotation, A Survey. **ACM Computing Surveys**, v. 52, n. 4, p. 1–45, 2020.

LIU, Alan. The state of the digital humanities: A report and a critique. **Arts and Humanities in Higher Education**, v. 11, n. 1–2, p. 8–41, 2012.

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.

LYMAN, Peter; VARIAN, Hal R. **How Much Information?.** Disponível em: <https://groups.ischool.berkeley.edu/archive/how-much-info-2003/>. Acesso em: 6 mar. 2023.

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**, v. 10, n. 2, p. 14, 2021.

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.

MARTINS, Dalton Lopes; LEMOS, Daniela Lucas da Silva; ANDRADE, Morgana Carneiro. Tainacan e omeka: proposta de análise comparativa de softwares para gestão de coleções digitais a partir do esforço tecnológico para uso e implantação**. Informação & Informação**, v. 26, n. 2, p. 569-595, 2021.

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.], p. asi.24650, 2022.

MEY, Eliane Serrão A. **Introdução à catalogação**. Brasília: Briquet de Lemos Livros, 1995.

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: 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: 20 jul. 2022.

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.

NAMGUNG, Min; CHIANG, Yao-Yi. Incorporating spatial context for post-OCR in map images. *In*: **Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery**. Seattle Washington: ACM, 2022, p. 14–17. Disponível em: <https://dl.acm.org/doi/10.1145/3557918.3565864>. Acesso em: 5 mar. 2023.

NASAR, Zara; JAFFRY, SyedWaqar; MALIK, Muhammad Kamran. Information extraction from scientific articles: a survey. **Scientometrics**, v. 117, n. 3, p. 1931–1990, 2018.

NISO (National Information Standards Organization). **Understanding metadata**. Bethesda: NISO Press. 2004.

OTMANI, Nassim Abdeldjallal; SI-MOHAMMED, Malik; COMPAROT, Catherine; *et al*. Ontology-based approach to enhance medical web information extraction. **International Journal of Web Information Systems**, v. 15, n. 3, p. 359–382, 2019.

PALAVITSINIS, Nikos. **Metadata Quality Issues in Learning Repositories**. 2013. Universidade de Alcalá, Espanha, 2013. Disponível em: <https://core.ac.uk/download/pdf/58910780.pdf>. Acesso em: 1 ago. 2022.

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.

POOLE, Alex H. The conceptual ecology of digital humanities. **Journal of Documentation**, v. 73, n. 1, p. 91-122, 2017.

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

PURWITASARI, Diana; FATICHAH, Chastine; SUMPENO, Surya; *et al*. Identifying collaboration dynamics of bipartite author-topic networks with the influences of interest changes. **Scientometrics**, v. 122, n. 3, p. 1407–1443, 2020.

ROMERO, Gustavo Candela. **Publicación y enriquecimiento semántico de datos abiertos en bibliotecas digitales**. 2019. UNIVERSIDAD DE ALICANTE, Espanha, 2019. Disponível em: <https://rua.ua.es/dspace/handle/10045/97353>. Acesso em: 1 ago. 2022.

SANTANA, Robson de. **Memória e resistência: do Instituto Brasileiro de Museus aos Museus Comunitários do Coque**. 2020. Dissertação (Mestrado em História) - Universidade Federal de Pernambuco, Recife, 2020.

SARACEVIC, Tefko. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, v. 1, n. 1, 1996. Disponível em: <https://periodicos.ufmg.br/index.php/pci/article/view/22308>. Acesso em: 2 mar. 2023.

SILVA, Daniela Lemos da; SOUZA, Renato Rocha. Representação de documentos multimídia: dos metadados às anotações semânticas. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 9, n.2, p. 1-22, 2014.

SILVA, Daniela Lucas da; SOUZA, Renato Rocha; ALMEIDA, Maurício Barcellos. Ontologias e vocabulários controlados: comparação de metodologias para construção. **Ciência da Informação**, v. 37, n. 3, p. 60–75, 2008.

SIQUEIRA, Joyce; CARMO, Danielle do; MARTINS, Dalton Lopes; LEMOS, Daniela Lucas da Silva; MEDEIROS, Vinícius Nunes; OLIVEIRA, Luis Felipe Rosa. 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. Springer International Publishing, 2021.

ŠLIBAR, Barbara; OREŠKI, Dijana; BEGIČEVIĆ REĐEP, Nina. 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). **EAD: Encoded Archival Description (EAD Official Site, Library of Congress)**. 2022. Disponível em: <https://www.loc.gov/ead/>. Acesso em: 1 ago. 2022.

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

STEPHAN, Haller; BEAT, Estermann; ANGELINA, Dungga Winterleitner. Study in View of the Further Development of DCAT-AP CH. *[S. l.]*, 2018. Disponível em: <https://arbor.bfh.ch/9499/>. Acesso em: 3 ago. 2022.

SVENONIUS, Elaine. The Intellectual Foundation of Information Organization. [s.l.] The MIT Press, 2000.

TANASIJEVIĆ, Ivana; PAVLOVIĆ-LAŽETIĆ, Gordana. HerCulB: content-based information extraction and retrieval for cultural heritage of the Balkans. **The Electronic Library**, v. 38, n. 5/6, p. 905–918, 2020.

TARMIZI, Fatin Amanina Ahmad; TAN, Phan Xuan; SHARIF, Khaironi Yatim; *et al*. Online News Veracity Assessment Using Emotional Weight. *In*: **Proceedings of the 2019 2nd International Conference on Information Science and Systems**. Tokyo Japan: ACM, 2019, p. 60–64. Disponível em: <https://dl.acm.org/doi/10.1145/3322645.3322688>. Acesso em: 5 mar. 2023.

THOMPSON, Ken. Programming Techniques: Regular expression search algorithm. **Communications of the ACM**, *[S. l.]*, v. 11, n. 6, p. 419–422, 1968.

TRUST, Jean Paul Getty; COLLEGE ART ASSOCIATION**. Categories for the Description of Works of Art**. 2022. Disponível em: <https://www.getty.edu/research/publications/electronic_publications/cdwa/introduction.html>. Acesso em: 17 jul. 2022.

TSIFLIDOU, Effie; MANOUSELIS, Nikos. Tools and Techniques for Assessing Metadata Quality. *Em*: GAROUFALLOU, Emmanouel; GREENBERG, Jane (org.). **Metadata and Semantics Research**. Communications in Computer and Information Science Cham: Springer International Publishing, 2013. v. 390p. 99–110. Disponível em: <http://link.springer.com/10.1007/978-3-319-03437-9_11>. Acesso em: 3 ago. 2022.

USAID, U. S. Agency for International Development. TIPS 12: Data Quality Standards. *[S. l.]*, v. 12, n. 2, 2009. Disponível em:

<https://www.fsnnetwork.org/sites/default/files/tips-dataqualitystandards.pdf>. Acesso em: 28 jul. 2022.

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

VISUAL RESOURCES ASSOCIATION. **CCO (Cataloging Cultural Objects): Why CCO?**[S. l.], 2021. Disponível em: <https://pt.slideshare.net/VisResAssoc/cco-cataloging-cultural-objects-why-cco>. Acesso em: 19 jul. 2022.

W3C LINKED DATA. **Data - W3C**. 2021. Disponível em: <https://www.w3.org/standards/semanticweb/data>. Acesso em: 01 ago. 2022.

WANG, Lin. Twinning data science with information science in schools of library and information science. **Journal of Documentation**, v.74, 2018.

WANG, Richard Y.; STRONG, Diane M. Beyond Accuracy: What Data Quality Means to Data Consumers. **Journal of Management Information Systems**, *[S. l.]*, v. 12, n. 4, p. 5–33, 1996.

WESTBROOK, R. Niccole; JOHNSON, Dan; CARTER, Karen; LOCKWOOD, Angela. 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 ago. 2022.

WHITESIDE, Ann. **The Core Categories for Visual Resources – Introduction (VRA Core 1.0)**. 1999. Disponível em: <https://www.loc.gov/standards/vracore/VRACore1_Introduction.pdf>. Acesso em: 2 mar 2023.

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.

WYNAR, Bohdan S. **Introduction to cataloging and classification**. 7ª ed. Colorado: Libraries Unlimited Inc., 1985.

ZENG, Marcia Lei. Interoperability. **Knowledge Organization**, v.46, n.2, p. 122-146, jan. 2019.

ZHANG, Zhe; WANG, Zhangyang; LI, Angela; *et al*. An AI-based Spatial Knowledge Graph for Enhancing Spatial Data and Knowledge Search and Discovery. *In*: **Proceedings of the 1st ACM SIGSPATIAL International Workshop on Searching and Mining Large Collections of Geospatial Data**. Beijing China: ACM, 2021, p. 13–17. Disponível em: <https://dl.acm.org/doi/10.1145/3486640.3491393>. Acesso em: 5 mar. 2023.