UNIVERSIDADE DE BRASÍLIA

PROGRAMA DE PÓS GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

Ata Nº: 36

Aos vinte nove dias do mês de novembro do ano de dois mil 2023, instalou-se a banca examinadora de Tese de Doutorado do(a) aluno(a) **Luis Felipe Rosa de Oliveira**, matrícula **19/0074001**. A banca examinadora foi composta pelos professores Dr(a). Daniela Lucas da Silva Lemos / membro titular externo à UnB / UFES, Dr(a). Laura Vilela Rodrigues Rezende / membro titular externo à UnB / UFG, Dr(a). João de Melo Maricato / membro titular interno à UnB / PPGCINF UnB, Dr(a). Marcio de Carvalho Victorino / suplente / PPGCINF UnB e Dr(a). Dalton Lopes Martins / orientador(a)/presidente / PPGCINF UnB. O(A) discente apresentou o trabalho intitulado **“ASPECTOS TECNOLÓGICOS DA AGREGAÇÃO DE ACERVOS DIGITAIS CULTURAIS: ENTIDADES VINCULADAS AO MINISTÉRIO DA CULTURA DO BRASIL”.**

Concluída a exposição, procedeu-se a arguição do(a) candidato(a), e após as considerações dos examinadores o resultado da avaliação do trabalho foi:

( ) Pela aprovação do trabalho;

(X) Pela aprovação do trabalho, com revisão de forma, indicando o prazo de até 30 dias para apresentação definitiva do trabalho revisado;

( ) Pela reformulação do trabalho, indicando o prazo de (Nº DE MESES) para nova versão;

( ) Pela reprovação do trabalho, conforme as normas vigentes na Universidade de Brasília.

Conforme os Artigos 34, 39 e 40 da Resolução 0080/2021 - CEPE, o(a) candidato(a) não terá o título se não cumprir as exigências acima.

Dr. Dalton Lopes Martins (PPGCINF UnB)
Presidente

Dra. Daniela Lucas da Silva Lemos (UFES)
Membro Titular Externo à UnB

Dra. Laura Vilela Rodrigues Rezende (UFG)
Membro Titular Externo à UnB

Dr. João de Melo Maricato (PPGCINF UnB)
Membro Titular Interno

Dr. Marcio de Carvalho Victorino (PPGCINF UnB)
Suplente

Luis Felipe Rosa de Oliveira
(Doutorando)

Documento assinado eletronicamente por **Dalton Lopes Martins**, **Professor(a) de Magistério Superior da Faculdade de Ciência da Informação**, em 12/01/2024, às 14:36, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Joao de Melo Maricato**, **Professor(a) de Magistério Superior da Faculdade de Ciência da Informação**, em 15/01/2024, às 10:25, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Daniela Lucas da Silva Lemos**, **Usuário Externo**, em 15/01/2024, às 11:40, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **LAURA VILELA RODRIGUES REZENDE**, **Usuário Externo**, em 17/01/2024, às 09:30, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **LUIS ROSA**, **Usuário Externo**, em 22/01/2024, às 13:37, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

Documento assinado eletronicamente por **Clovis Carvalho Britto**, **Coordenador(a) da Pós-Graduação da Faculdade de Ciência da Informação**, em 31/01/2024, às 10:21, conforme horário oficial de Brasília, com fundamento na Instrução da Reitoria 0003/2016 da Universidade de Brasília.

A autenticidade deste documento pode ser conferida no site http://sei.unb.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **10514269** e o código CRC **4F53C46C**.

Referência: Processo nº 23106.128241/2023-34 SEI nº 10514269

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE CIÊNCIA DA INFORMAÇÃO

LUIS FELIPE ROSA DE OLIVEIRA

ASPECTOS TECNOLÓGICOS DA AGREGAÇÃO DE ACERVOS DIGITAIS CULTURAIS: ENTIDADES VINCULADAS AO MINISTÉRIO DA CULTURA DO BRASIL

BRASÍLIA
2023

LUIS FELIPE ROSA DE OLIVEIRA

ASPECTOS TECNOLÓGICOS DA AGREGAÇÃO DE ACERVOS DIGITAIS CULTURAIS: ENTIDADES VINCULADAS AO MINISTÉRIO DA CULTURA DO BRASIL

Tese apresentada à Faculdade de Ciência da Informação da Universidade de Brasília para obtenção do título de Doutor no Programa de Pós-Graduação em Ciência da Informação, na área de concentração de Gestão da Informação.

Orientador: Prof. Dr. Dalton Lopes Martins

BRASÍLIA
2023

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Oliveira, Luis Felipe Rosa de
OO48a ASPECTOS TECNOLÓGICOS DA AGREGAÇÃO DE ACERVOS DIGITAIS CULTURAIS: ENTIDADES VINCULADAS AO MINISTÉRIO DA CULTURA DO BRASIL / Luis Felipe Rosa de Oliveira; orientador Dalton Lopes Martins. -- Brasília, 2023.
151 p.

Tese(Doutorado em Ciência da Informação) -- Universidade de Brasília, 2023.

1. agregação de acervos digitais. 2. organização do conhecimento. 3. representação da informação. 4. mapeamento de metadados. 5. instituições do patrimônio cultural. I. Lopes Martins, Dalton, orient. II. Título.

DEDICATÓRIA

Dedico esta pesquisa às pessoas envolvidas com as iniciativas culturais brasileiras que trabalham para a preservação e democratização da riqueza cultural deste país, pessoas que, mesmo com as dificuldades estruturais impostas sobre elas, lutam e batalham para manter a memória e o conhecimento preservados e acessíveis à sociedade.

Estendo esta dedicatória aos estudantes de pós-graduação que passaram ou passam por dificuldades no desenvolvimento de seus estudos. O processo de produção de uma pesquisa científica é um trabalho árduo, que exige um grande esforço mental e psicológico, e que, por vezes, pode levar à exaustão. Buscar ajuda profissional e evitar uma autocobrança excessiva são atitudes importantes nesse momento.

Por fim, deixo uma dedicatória especial ao meu pai, Luis Batista de Oliveira, falecido em 25 de março de 2021, em decorrência de complicações da COVID-19. Homem de pouca noção cultural, mas de um conhecimento de vida admirável.

AGRADECIMENTOS

Agradeço, primeiramente, à minha companheira de vida, Calíope Victor Spíndola de Miranda Dias, por sofrer comigo, por lutar comigo, por insistir em mim e estar ao meu lado durante todo o processo de desenvolvimento desta tese, sem ela, talvez não tivesse conseguido. Agradeço ao professor e meu orientador, Dalton Lopes Martins, que me orientou desde a graduação e me proporcionou condições de viver experiências extraordinárias dentro do contexto acadêmico, agradeço pelo conhecimento compartilhado e pela paciência em me orientar. Agradeço também aos professores Daniela Lucas e João Maricato, pelas orientações que resultaram na produção de uma tese muito melhor. Agradeço à equipe do Laboratório de Inteligência de Redes (UnB), anteriormente Laboratório de Políticas Públicas Participativas (L3P), pelos bons e longos anos de vivência e perseverança juntos. Agradeço aos colaboradores das instituições do patrimônio cultural: Funarte, Ibram, Iphan, e a todos os demais incentivadores e apoiadores da cultura brasileira, os quais tive o prazer de conhecer e vivenciar boas experiências. Agradeço, ainda, ao Programa de Pós-graduação em Ciência da Informação da Universidade de Brasília, mais especificamente aos professores que me lecionaram essa incrível ciência, e às demais pessoas que trabalham na secretaria e nas comissões do programa, que sempre estiveram a postos para auxiliar quando preciso. Agradeço, por fim, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo financiamento da pesquisa através da bolsa de doutorado que recebi, e à Fundação de Amparo à Pesquisa do Estado de São Paulo, por financiar o projeto de pesquisa que deu origem ao meu doutorado.

RESUMO

A disseminação da documentação sobre o patrimônio cultural na internet é um fator crescente na realidade das instituições de memória que buscam digitalizar e publicar seus acervos. O aumento dessa presença digital dos acervos revela um fenômeno de duas faces: por um lado, aumentam as condições de acesso aos bens culturais pelos usuários com acesso à internet; por outro, flerta com o problema da diversidade de locais de acesso aos acervos, revelando a lacuna da dificuldade de se recuperar informações de maneira agregada sobre uma mesma temática, que neste caso é a cultural. É sobre esta lacuna identificada que a presente pesquisa é proposta. A respeito da temática da agregação de acervos culturais digitais, busca-se responder como estão disponibilizados os acervos digitais publicados na internet pelas instituições vinculadas ao Ministério da Cultura do Brasil, quanto às suas características tecnológicas e de organização do conhecimento e representação da informação, diante da possibilidade de implementação de um serviço de agregação destes acervos. Através do referencial teórico composto por uma revisão de literatura sobre a organização do conhecimento e representação da informação em acervos digitais, por um levantamento conceitual sobre o mapeamento de metadados para a agregação desses acervos, e uma revisão das características tecnológicas das iniciativas de agregação de várias nações, foi possível fundamentar uma estratégia de pesquisa em busca da caracterização dos acervos digitais das entidades vinculadas ao Ministério da Cultura e o desenvolvimento de um protótipo de agregação desses acervos. Para alcançar esses resultados, foi utilizada a metodologia de estudo de casos múltiplos, juntamente com a aplicação das técnicas de análise categorial e estatística descritiva para processar os resultados encontrados. O método de desenvolvimento do protótipo foi constituído de cinco etapas, baseadas no referencial teórico desta pesquisa. A partir dos resultados da caracterização dos acervos digitais das entidades culturais analisadas, foi identificada a condição da heterogeneidade em vários aspectos da documentação dos acervos, o que ajudou a prospectar a dificuldade de implementação de um serviço de agregação. Quanto ao protótipo de agregação, foi desenvolvido, em sua maioria, a partir de programas de raspagem de dados, o que indica uma alta complexidade e capacidade mínima de sustentabilidade de um serviço de agregação, além do uso de tecnologias eficientes de agregação como o Elastic Stack, e o uso do Dublin Core como padrão

de agregação de metadados. Dessa forma, conclui-se que a pesquisa desenvolvida nesta tese atingiu as propostas de caracterizar os acervos digitais das entidades vinculadas ao Ministério da Cultura e a proposta de desenvolvimento de um protótipo de agregação desses acervos digitais. Entende-se, ainda, que o conteúdo deste estudo pode servir como insumo científico para demais pesquisas na temática de agregação dos acervos digitais, bem como na tomada de decisão no âmbito das instituições do patrimônio cultural brasileiro, além de apontar para a necessidade de uma articulação entre essas instituições em busca de promover mais iniciativas brasileiras de agregação de acervos culturais.

Palavras-chave: agregação de acervos digitais; organização do conhecimento; representação da informação; mapeamento de metadados; instituições do patrimônio cultural.

ABSTRACT

The dissemination of documentation about cultural heritage on the internet is a growing factor in the reality of memory institutions that seek to digitize and publish their collections. The increase in this digital presence of collections reveals a two-sided phenomenon: on the one hand, it improves access conditions to cultural assets for users with internet access; on the other hand, it poses the problem of the diversity of access points to collections, revealing the gap in the difficulty of retrieving information in an aggregated manner on the same topic, in this case, culture. It is this identified gap that the present research aims to address. Concerning the theme of aggregating digital cultural collections, we aim to answer how digital collections published on the internet by institutions linked to the Ministry of Culture of Brazil are made available, in terms of their technological characteristics, organization of knowledge, and representation of information, considering the possibility of implementing an aggregation service for these collections. Through the theoretical framework, which includes a literature review on the organization of knowledge and representation of information in digital collections, a conceptual survey on metadata mapping for aggregating these collections, and a review of the technological characteristics of aggregation initiatives from different nations, we were able to formulate a research strategy to characterize the digital collections of entities linked to the Ministry of Culture and develop a prototype for aggregating these collections. To achieve these results, we employed the methodology of multiple case studies, coupled with the application of categorical analysis and descriptive statistics techniques to process the results obtained. The prototype development method comprised five stages, based on the theoretical framework of this research. Based on the results of the characterization of the digital collections of the cultural entities analyzed, we identified heterogeneity in several aspects of the collections' documentation, which helped to anticipate the difficulty of implementing an aggregation service. As for the aggregation prototype, it was primarily developed using data scraping programs, indicating high complexity and minimal sustainability capacity for an aggregation service, in addition to the use of efficient aggregation technologies such as the Elastic Stack, and the use of Dublin Core as a metadata aggregation standard. Therefore, we conclude that the research conducted in this thesis fulfilled the initial proposal of characterizing the digital collections of entities linked to the Ministry of Culture and developing a prototype for

aggregating these digital collections. It is also understood that the content of this study can serve as scientific input for other research on the topic of aggregating digital collections, as well as in decision-making within the scope of Brazilian cultural heritage institutions. Additionally, it underscores the need for coordination between these institutions to promote more Brazilian initiatives for aggregating cultural collections.

Keywords: aggregation of digital collections; knowledge organization; information representation; metadata mapping; cultural heritage institutions.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURA E SIGLAS

Ancine	Agência Nacional do Cinema
API	*Application Programming Interface*
CGI	Comitê Gestor da Internet
CSV	*Comma Separeted Values*
DPLA	*Digital Public Library of America*
EDM	Europeana Data Model
FAIR	*Findable, Accessible, Interoperable e Reusable*
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
FBN	Fundação Biblioteca Nacional
FCRB	Fundação Casa de Rui Barbosa
FP	Fundação Palmares
FTP	*File Transfer Protocol*
Funarte	Fundação Nacional de Artes
Ibram	Instituto Brasileiro de Museus
IFLA	*International Federation of Library Associations and Institutions*
INBCM	Inventário Nacional de Bens Culturais Musealizados
IPHAN	Instituto do Patrimônio Histórico e Artístico Nacional
JSON	*JavaScript Objects Notation*
KOS	*Knowledge Organization System*
LOD	*Linked Open Data*
MAP	*Metadata Profile Application*
OAI-PMH	*Open Archive Initiative Protocol for Metadata Harvesting*
OCR	*Optical Character Recognition*
RDF	*Resource Description Framework*
RSS	*Really Simple Syndication*
SPARQL	*SPARQL Protocol and RDF Query Language*
UnB	Universidade de Brasília
XML	*Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

A disseminação da documentação sobre o patrimônio cultural na internet é um fator crescente na realidade das instituições de memória que buscam digitalizar e publicar seus acervos. É comum que essas instituições compartilhem seus acervos online para democratizar o acesso ao conhecimento cultural e o contexto da sociedade em rede, e as tecnologias digitais disponíveis atualmente contribuem para essa finalidade.

Ferramentas avançadas de digitalização de objetos do patrimônio cultural, com foco na internet, promovem o crescimento dos acervos digitais online (Dijkshoorn *et al*., 2018; Potenziani *et al.,* 2018; Scopigno *et al*., 2017), resultando em um conjunto complexo e heterogêneo de objetos digitais disponíveis através de sistemas digitais de informação utilizados para gerenciar e publicar acervos. Dessa forma, a maioria desses objetos está disponível por meio de infraestruturas informacionais, que têm o objetivo de armazenar, dar suporte, preservar e divulgar esses objetos digitais culturais (Doerr *et al*., 2010a; Siqueira; Martins, 2022).

Essas infraestruturas informacionais podem ser entendidas como “um *framework* técnico, social e político que abrange pessoas, tecnologias, ferramentas e serviços utilizados para facilitar o uso distribuído e colaborativo de conteúdo” (Borgman, 2010, p. 19, tradução nossa). Isso destaca o crescente interesse nas pesquisas que relacionam estudos sociais aplicados ao uso das tecnologias digitais, como o campo de estudo das humanidades digitais (Liu, 2012; Koltay, 2016; Poole, 2017; Clement; Carter, 2017).

Vale ressaltar a interdisciplinaridade que existe entre as humanidades digitais e a ciência da informação, uma vez que as duas áreas têm objetos de estudo e práticas metodológicas em comum. Temas de pesquisas relevantes, como produção, preservação, organização, gerenciamento, disseminação, acesso, recuperação, uso e reutilização de dados, metadados, informações, conhecimento, documentos e outros elementos informativos em diferentes contextos e circunstâncias, são compartilhados por essas áreas (Clement; Carter, 2017; Koltay, 2016)

Borko (1968) e Saracevic (1999) validam essa visão com base nos estudos da ciência da informação, ao afirmarem que esse domínio direciona suas investigações para as características e o funcionamento da informação, as influências que governam sua circulação e as abordagens empregadas para manipulá-la com o

objetivo de aprimorar sua disponibilidade e utilidade de maneira coletiva, organizacional ou pessoal, sempre que houver demanda por informação.

Sendo assim, entende-se que está dentro do escopo de pesquisas nessas áreas, compreender o fenômeno da formação de grandes bases de dados sobre objetos digitais do patrimônio cultural, e como isso afeta uma parcela da sociedade imbuída em processos informacionais e de conhecimento. O que pode, por exemplo, gerar impactos até em processos econômicos, quanto ao reúso dos objetos digitais na indústria criativa, na promoção do turismo, e ainda no auxílio a processos educativos (Poort *et al*., 2013; Tessler, 2013).

Os lugares físicos onde esses objetos digitais estão normalmente alocados incluem bibliotecas, museus, arquivos e repositórios de pesquisa usados para estudos científicos. No entanto, com a digitalização e publicação desses acervos online, tais objetos assumem uma existência virtual com características específicas de acesso e reúso, mesmo quando os objetos físicos não estão presentes (Araripe, 2004).

Em tempo, com a característica heterogênea dos acervos digitais do patrimônio cultural, que abrange diferentes origens e formatos, se faz pertinente definir o escopo do conceito utilizado para os propósitos desta pesquisa, que caminha em direção ao exposto por Martins *et al*. (2022):

> os acervos são considerados manifestações físicas ou digitais de patrimônio que fomentam a memória social por meio de fontes de informação, com a perspectiva de compreender melhor o mundo atual, tornando-se fundamental que estejam bem organizados e representados. Assim, os acervos digitais são vistos como sistemas de informação com funções de comunicação social onde as coleções de bibliotecas, arquivos, museus e outros ambientes de patrimônio cultural convergem em uma arena multimodal digital (Martins *et al*., 2022, p. 2).

Dessa forma, é estudando as características tecnológicas e de infraestrutura informacional desse contexto dos acervos digitais que esta pesquisa buscará explorar os desafios e possíveis caminhos para melhorar o acesso e reúso dos acervos digitais.

Existem indícios de que a forma como as instituições publicam seus acervos digitais heterogêneos pode dificultar a recuperação de informações sobre os objetos culturais (Martins *et al*., 2021; Júnior; Lemos, 2023). À medida em que cada instituição de memória torna público um ou mais de seus acervos digitais para acesso via internet, é necessário interagir com diferentes sites, sistemas de repositórios e, consequentemente, diferentes formas de exploração e recuperação da informação.

Para exemplificar, no âmbito do Instituto Brasileiro de Museus (Ibram) existem quase 30 museus com acervos publicados na internet[1]. Eventualmente, surge a necessidade de realizar consultas transversais aos acervos, como buscar obras produzidas pelo artista brasileiro Aleijadinho. Hoje, cada museu disponibiliza seu acervo em site próprio e específico, o que requer que a busca pelas informações seja feita de forma individualizada a partir do acesso a cada um destes repositórios digitais, o que acaba por resultar em um consumo de tempo e esforço considerável. Embora a busca individualizada ainda seja mais eficiente que a busca em acervos físicos, devido à eficácia dos sistemas de informação, o processo destas pesquisas poderia ser aprimorado e se tornar mais eficaz se fosse possível realizar tal consulta a partir de um buscador único.

No exemplo mencionado, existe o fator atenuante de que os museus do Ibram utilizam o sistema Tainacan[2] para publicar seus acervos digitais na internet. No entanto, se essa consulta fosse ampliada para abranger a cultura brasileira, incluindo bibliotecas e fundações culturais, por exemplo, o escopo das variáveis aumentaria ainda mais. É possível sistematizar essas variáveis em oito agrupamentos (Martins *et al*., 2022, p. 11), a saber: o tipo de sistema de recuperação da informação, o software utilizado, as licenças de direitos autorais, os modelos de organização do conhecimento e representação da informação do acervo, as formas de visualização do acervo, as possibilidades de extração de dados, os tipos de mídia dos objetos digitais e o número de itens do acervo.

Esse escopo de variáveis situa a diversidade dos componentes envolvidos na existência dos acervos digitais culturais publicados na internet. Cada instituição de memória tem um universo de possibilidades tecnológicas para publicação de seus acervos na internet, e a variedade dessas possibilidades se torna um desafio para o acesso único e agregado às informações culturais.

Cientes desse desafio, algumas nações têm adotado a estratégia de agregar os acervos digitais de suas instituições de memória em um único repositório digital, proporcionando acesso completo às informações sobre a cultura (Siqueira; Martins, 2020). Essas iniciativas de agregação de acervos envolvem diferentes processos de implementação para lidar com a organização do conhecimento e as variáveis

[1] Vide: https://www.gov.br/museus/pt-br/museus-ibram/museus-ibram.
[2] Vide: https://tainacan.org/.

tecnológicas presentes nos acervos publicados na internet. Uma descrição mais detalhada dessas iniciativas será apresentada no tópico 3 desta pesquisa.

Como já mencionado, a estratégia de agregação dos acervos culturais pode ser vista como uma forma de reúso (Freire; Sales; Sayão, 2020). Sob a perspectiva da curadoria digital, os acervos digitais culturais desempenham um papel ativo ao serem publicados na internet, resultando em benefícios como o aumento da visibilidade dos seus objetos, reunindo tanto acervos de amplo quanto aqueles com baixo acesso, e aprimorando a eficiência das consultas entre eles, tanto por órgãos gestores institucionais quanto por usuários que desejam explorar os conteúdos ou até mesmo desenvolver pesquisas que têm os acervos como objeto de estudo.

Portanto, existe uma dinâmica no processo de disponibilização de acervos digitais através de serviços de agregação. Essa dinâmica envolve, dentre outras vertentes, a definição da infraestrutura informacional e tecnológica de acordo com modelo de agregação proposto, a fim de permitir que os usuários, como consumidores finais, possam explorar e recuperar as informações de maneira unificada, padronizada e com qualidade.

A presente pesquisa busca compreender as características dessa dinâmica com o escopo delimitado às instituições do patrimônio cultural brasileiro. Assim, o objeto de estudo se constitui a partir dos acervos digitais publicados na internet das sete instituições de memória vinculadas ao Ministério da Cultura: Instituto Brasileiro de Museus (Ibram), Fundação Nacional de Artes (Funarte), Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN), Agência Nacional do Cinema (Ancine), Fundação Biblioteca Nacional (FBN), Fundação Casa de Rui Barbosa (FCRB) e Fundação Palmares (FP).

1.1 Problema de pesquisa

No âmbito dos conceitos apresentados até então, bem como dentro da delimitação de escopo do objeto em estudo, busca-se compreender como estão disponibilizados os acervos digitais publicados na internet pelas instituições vinculadas ao Ministério da Cultura do Brasil, quanto às suas características tecnológicas e de organização do conhecimento e representação da informação, frente à possibilidade de implementação de um serviço de agregação destes acervos?

1.2 Hipótese de pesquisa

A partir do problema da pesquisa, que busca compreender as características da infraestrutura informacional e tecnológica dos acervos digitais brasileiros conforme definidos no escopo, nossa hipótese é de que promover a implementação de um serviço de agregação desses acervos heterogêneos do modo em que estão publicados atualmente, do ponto de vista da manutenção e reprodução, de mecanismos tecnológicos e de modelos de organização do conhecimento para garantir a integração dos dados dos objetos digitais, exigirá um conjunto complexo de ações e poderá ter como resultado um modelo e formato pouco sustentável e escalável.

Essa hipótese se fundamenta nos estudos já realizados sobre a qualidade dos dados e sobre as características desses acervos (Martins *et al*., 2022; Martins *et al*., 2021; Júnior; Lemos, 2023), que indicam uma baixa sofisticação da documentação e práticas tecnológicas necessárias para acesso e reúso dos objetos digitais.

Buscaremos validar essa hipótese a partir do desenvolvimento de um protótipo de agregação dos dados dos acervos digitais das instituições do patrimônio cultural vinculadas ao Ministério da Cultura do Brasil, e entende-se que com o processo de desenvolvimento será necessário despender de mecanismos tecnológicos e de modelos de organização do conhecimento para agregar os acervos identificados.

1.3 Objetivo geral

Como objetivo geral se espera identificar e descrever as condições informacionais e tecnológicas necessárias para a agregação de acervos digitais culturais do Brasil, conforme delimitado pelo escopo desta pesquisa.

1.4 Objetivos específicos

- Mapear sistematicamente as diferentes formas de organização do conhecimento e representação da informação existentes nos acervos digitais das instituições vinculadas ao Ministério da Cultura do Brasil;

- Coletar os dados dos acervos das instituições vinculadas ao Ministério da Cultura do Brasil;
- Padronizar os dados coletados dos acervos digitais a partir de um padrão de metadados;
- Desenvolver um protótipo de agregação dos acervos digitais padronizados a partir de um único padrão de metadados.

1.5 Justificativa

Descrever de forma aplicada quais as características da infraestrutura informacional e tecnológica para a agregação dos acervos digitais das instituições vinculadas ao Ministério da Cultura do Brasil é a questão que essa pesquisa se propõe a solucionar. O produto da solução desta questão é o levantamento dessas características de maneira pontual, através das análises dos resultados do estudo, bem como de maneira aplicada, através do desenvolvimento de um protótipo de agregação dos acervos.

Esse produto é composto por elementos de cunho tecnológico, definidos pelos mecanismos digitais utilizados e identificados no processo do estudo, e por elementos de cunho informacionais, compostos pelos modelos de formas de organização do conhecimento necessárias para compreender e solucionar o problema da agregação dos objetos do patrimônio cultural que os acervos armazenam.

Dessa forma, entende-se que a atual pesquisa está delimitada dentro do contexto de estudos da ciência da informação e interdisciplinarmente com as humanidades digitais (Liu, 2012; Koltay, 2016; Poole, 2017; Clement; Carter, 2017), uma vez que se busca comunicar de maneira pertinente as características de uso da tecnologia digital em um objeto de estudo de cunho social, que são os acervos digitais do patrimônio cultural brasileiro, em busca de promover referencial científico sobre a melhoria do acesso e reúso das informações sobre os objetos digitais contidos nesses acervos.

Com os resultados desta pesquisa, é esperado contribuir com outros estudos na área da ciência da informação e das humanidades digitais, que buscam compreender a dinâmica de agregação de acervos digitais culturais no contexto brasileiro. Uma vez que não foram encontrados estudos cujo objeto tenha a abrangência definida pelas instituições de patrimônio cultural vinculadas ao Ministério

da Cultura do Brasil, a descrição de maneira aplicada das características de infraestrutura informacional e tecnológica desse processo de agregação, como produto desta pesquisa, é entendido como um insumo essencial que complementa os estudos técnicos e aplicados dentro deste escopo de investigação científica.

Outra perspectiva que justifica a produção desta pesquisa é a contribuição prática com a área da cultura brasileira. De acordo com a pesquisa TIC Cultura do Comitê Gestor da Internet realizada em 2020, que tem como objetivo levantar dados e gerar análise sobre a situação do uso das Tecnologias de Informação e Comunicação no âmbito das instituições de memória brasileiras, há uma diferença entre a existência de acervos digitalizados e sua disponibilização ao público:

> Os indicadores da pesquisa TIC Cultura sobre a presença de acervos nos equipamentos culturais apontam que, a despeito da existência de uma grande variedade de coleções compostas por diversos tipos de materiais, apenas uma pequena parte desse patrimônio foi digitalizada e disponibilizada à sociedade por meio de plataformas digitais. A criação e a disseminação de acervos digitais é, portanto, uma oportunidade muito pouco explorada pelas instituições culturais brasileiras, que possibilitaria a difusão e a ampliação do acesso a esses conteúdos em sua conexão com a memória e a contemporaneidade (CGI, 2021, p. 74).

Mesmo com a baixa disponibilização dos acervos digitais em plataformas para acesso pela internet, é expressiva a oportunidade da implementação e divulgação de acervos digitais culturais. Entende-se, então, que há uma abertura no campo para a proposta de pesquisas sobre a publicação de acervos digitais e como ampliar as formas de acesso ao conteúdo desses acervos.

Isso também é corroborado por estudos que indicam a situação crescente da necessidade da implementação de políticas e ações voltadas para o contexto dos acervos digitais (Martins; Silva; Carmo, 2018; Dias, Martins, 2020; Dias 2021), bem como projetos de publicação e agregação de acervos museológicos brasileiros (Olivera; Feitosa, 2021; Siqueira; Martins, 2021).

Por fim, vale ressaltar, ainda, que esta pesquisa é complementar e foi realizada em concomitância com o projeto FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo)/UnB (Universidade de Brasília), denominado “Interoperabilidade entre os repositórios digitais do patrimônio cultural brasileiro: da Web Semântica e dados abertos ligados às ferramentas de busca e recuperação da informação”, do qual participei como pesquisador e atuei em todos os processos de pesquisa. Assim, como produto do projeto de pesquisa, e com a proposta de aumentar o escopo teórico e técnico dos resultados do projeto, esta tese também se fundamenta

na necessidade latente de iniciativas de agregação de acervos digitais de instituições de memória, como sugere a aprovação e o fomento do projeto pela FAPESP.

1.6 Estrutura da pesquisa

Esta pesquisa está estruturada em cinco grandes tópicos: 1 – Introdução, composta pela contextualização da pesquisa, e a apresentação do problema de pesquisa, a hipótese de pesquisa, o objetivo geral, os objetivos específicos, e a justificativa; 2 – Referencial teórico, composto pela revisão de literatura sobre a organização do conhecimento e a representação da informação no contexto dos acervos digitais, e um referencial sobre o mapeamento de metadados para agregação de acervos digitais; 3 – Iniciativas de agregação de acervos digitais do patrimônio cultural, um estado da arte das iniciativas de agregação aplicadas em diferentes nações, e um resumo sobre as características técnicas da agregação de acervos digitais; 4 – Metodologia, composto pela descrição do método de estudo de casos múltiplos, bem como a especificação de todas as etapas desta metodologia e seu protocolo de aplicação; 5 – Resultados, compostos pelo cômputo da coleta de dados sobre os acervos digitais das entidades vinculadas ao MinC, e pelo desenvolvimento do protótipo de agregação dos acervos dessas entidades; 6 – Conclusões, composto por uma reflexão geral sobre a pesquisa, e os elementos dos alcances e limitações identificados.

2 REFERENCIAL TEÓRICO

Com o objetivo de melhor contextualizar o objeto de estudo desta pesquisa e fundamentar conceitualmente as variáveis de aplicação utilizadas na metodologia, este capítulo apresenta o referencial teórico sobre a organização do conhecimento e representação da informação, e sobre o mapeamento de metadados para agregação de acervos digitais culturais.

2.1 Acervos digitais do patrimônio cultural: organização do conhecimento e representação da informação

Ao decompor o objeto de estudo desta pesquisa é possível destacar dois elementos: os acervos digitais publicados na internet, e as instituições de memória que detêm tais acervos. Um terceiro elemento, intrínseco aos dois primeiros, é o patrimônio cultural, que define as características dos recursos informacionais envolvidos neste contexto. Assim, entende-se por patrimônio cultural, “o legado de objetos físicos, ambiente, tradições e conhecimentos de uma sociedade que são herdados do passado, mantidos e desenvolvidos no presente, e preservados (conservados) para o benefício das gerações futuras” (Hyvönen, 2022, p. 1, tradução nossa).

Observa-se a intangibilidade deste elemento, uma vez que é composto de fenômenos sócio-históricos. Tais fenômenos são registrados através de diferentes meios (manuscritos, livros, artefatos, obras de arte, documentos, fotografias etc.), que caracterizam uma primeira instância da representação da informação, aquela em que os objetos que constituirão o acervo são concebidos (Alvarenga, 2003).

A concepção do acervo em si envolve outra instância de representação, uma vez que,

> [...] no processo de tratamento ou processamento dos registros de conhecimento para fins de armazenagem nos sistemas de informação, é requerido um novo estágio de representação, desta vez partindo-se não do ser ontológico em si, mas do conhecimento sobre o ser, expresso em documentos. Esta seria uma representação secundária. Nesse sentido, a representação secundária teria por objeto prioritário não o acervo da ontologia, das coisas e seres existentes, mas o acervo de conhecimentos sobre essas coisas e seres, objetos da epistemologia (Alvarenga, 2003, p. 6).

Essa segunda instância de representação da informação acerca dos acervos de patrimônio cultural é permeada por um processo de organização informacional composto por análise, classificação, ordenação e recuperação desses registros (Lima; Alvares, 2012). Assim, um acervo do patrimônio cultural é constituído a partir da organização da informação presente nesta segunda instância de representação da informação do patrimônio cultural. Essa constituição do acervo pode ser interpretada como um fenômeno observado a partir da ótica de estudos sobre a organização do conhecimento e representação da informação (Almeida, 2013; Dahlberg, 1993; Hjorland, 2003; Hjorland, 2007; Hjørland, 2015; Svenonius, 2000; Taylor, 2004; Zeng, 2019).

Quando contextualizado pela digitalização e publicação na internet, esse fenômeno é imbuído pelos adventos das tecnologias da informação e comunicação e pelos mecanismos computacionais envolvidos. Isso situa os acervos digitais em ambientes multimídia com variações em categorias, formatos e complexidade. Além dos tipos comuns de mídia, como texto, áudio, imagem e vídeo, outros meios de disseminação dos acervos de objetos digitais culturais ganham força, como objetos 3D, tour virtual, imagens em extrema qualidade[3], anotações semânticas em imagens[4], por exemplo.

Esses novos elementos que envolvem o fenômeno da digitalização e publicação dos acervos digitais culturais na internet criam um contexto mais complexo e abrangente. Isso exige uma abordagem que implemente diferentes formas de se lidar com o processamento e organização da informação vinculada aos objetos multimídia, em busca de promover o melhoramento das formas de busca e recuperação dessas informações, como abordagens semânticas de agregação. (Lemos; Souza, 2020).

A busca por melhorar as formas de recuperação de informação em acervos é um dos elementos de conexão entre as áreas de estudo de ciência da informação e ciências da computação (Saracevic, 1999), sobretudo quanto aos processos que envolvem a organização de documentos e a característica semântica da informação

[3] O projeto Gigapixel da iniciativa Google Arts & Culture permite interagir com os detalhes de obras de artes através do zoom em várias camadas, possibilitado pela digitalização em camadas das obras. Vide: https://artsandculture.google.com/project/gigapixels.

[4] Processo que permite inserir metadados em elementos de imagens.

(Almeida, 2013; Almeida; Souza; Fonseca, 2011; Barite, 2000; Dahlberg, 1978; Hjorland, 2007; Jacob, 2004; Khoo; Na, 2006; Ranganathan, 1967).

Dentro deste contexto, percebe-se uma evolução na forma como as pesquisas dessa área foram concebidas, principalmente no que tange ao uso de técnicas de análise de conteúdo e de categorização para estudo dos registros de conhecimento sobre os sistemas de informação (Martins *et al*., 2022).

Essa dinâmica que envolve a evolução dos estudos em recuperação da informação (Abadal; Lluis, 2005; Baeza-Yates; Ribeiro-Neto, 2011; Hjørland, 2016; Lancaster, 1986; Luhn, 1953; Machado; Souza; Simões, 2019; Mooers, 1960), e o avanço das tecnologias de informação e comunicação, promove a consolidação de uma intersecção entre áreas de pesquisa que tem como objetos de estudo a informação, a computação e as humanidades. Assim, recursos informacionais do âmbito do patrimônio cultural ganham evidência quando estudados junto a uma realidade de disseminação digital na internet, como indicam as iniciativas da IFLA (2020) e a EUROPEANA TECH (2021).

Os componentes que estão presentes na estrutura da organização do conhecimento dos acervos digitais do patrimônio cultural envolvem várias dessas camadas de recursos informacionais. Os objetos e seu processo de curadoria, carregam a memória e a representação da informação sócio-histórica representando a camada das humanidades. Os meios e formatos digitais, os mecanismos que processam a apresentação e busca no acervo e as plataformas e os softwares utilizados para armazenamento e publicação dos acervos representam a camada computacional, e os sistemas de organização do conhecimento, que fundamentam a estruturação dos registros dos objetos do acervo, representando a camada informacional.

Dentro desta camada informacional, os sistemas de organização do conhecimento (KOS) são “mecanismos de organização da informação” (Hodge, 2000, p. 3), e estão presentes no fundamento de execução dos processos de curadoria dos acervos em instituições do patrimônio cultural. Esses sistemas têm como premissa a estruturação de recursos informacionais, quanto às suas relações, indexação e catalogação, orientação e disposição, e categorização (Hjorland, 2007).

Os KOS podem ser classificados em três categorias: listas de termos, presentes, por exemplo, no controle de autoridades, localidades, línguas, entre outros

(como as listas de código definidas no padrão de metadados MARC 21[5]); classificação e categorização, presente na definição de tipologias de objetos, materiais utilizados, e períodos históricos; e relacionamentos, que incidem sobre uma camada mais complexa da organização do conhecimento, representada por tesauros e taxonomias, que são mecanismos de organização que lidam com definição de hierarquias e associação de termos (como os tesauros voltados para o patrimônio museológico brasileiro (Ferrez; Bianchini, 1987; Ferrez, 2016) e, no caso das ontologias, com uma maior definição dos relacionamento entre termos e de suas regras lógicas que delimitam a aplicação dos termos nos sistemas de informação (Guizzardi, 2020; Lemos; Martins; Carmo, 2022).

À medida que os recursos informacionais são estruturados e organizados pela aplicação dos KOS, a capacidade de recuperação da informação de fontes de representação da informação como os acervos aumenta. Uma vez classificados, os termos que representam as categorias e classificações desses acervos podem se tornar filtros e campos de busca de sistemas computacionais, permitindo uma exploração específica e objetiva do conteúdo dos acervos. Além disso, a definição de KOS comuns entre acervos de uma mesma área ou de áreas distintas pode auxiliar na agregação semântica desses acervos, uma vez que a forma de representar o objeto do patrimônio cultural segue uma mesma forma de organização.

Assim, a aplicação dos KOS é entendida como uma prática essencial no processo de organização do conhecimento e representação da informação em conjuntos de recursos informacionais multimídia na internet (Lemos; Souza, 2020). Essa representação da informação ainda pode ser compreendida como produto do processo de descrição da própria informação, que pode ser a partir do conteúdo ou da mídia do objeto representado (Svenonius, 2000).

Essa abordagem auxilia na compreensão prática dos elementos que compõem a representação de um objeto de um acervo digital. Enquanto documento ou mídia, a informação pode ser descrita através de uma linguagem que representa (Abbas, 2010; Gilliland, 2016; IFLA, 2009; Zeng; Qin, 2016), em um processo de catalogação, os aspectos que descrevem o documento, o que o unifica em um acervo on-line, por exemplo, e permite sua identificação por sistemas de busca e descoberta

[5] Os formatos MARC 21: Antecedentes e Princípios. Vide: https://www.loc.gov/marc/96principl.html#five.

(IFLA, 2016). Já enquanto temática do objeto, a aplicação de linguagens de descrição do conteúdo (Foskett, 1973; Lancaster, 1986; Niso, 2005) busca a representação dos aspectos intangíveis e temáticos, comumente traduzidos utilizando tesauros e/ou ontologias, por exemplo. Isso permite a exploração do acervo através de filtros, e a agregação entre diferentes fontes de informação em um mesmo domínio.

Como exemplos dentro desse contexto das linguagens de descrição da informação e descrição do conteúdo dos objetos informacionais, existem princípios de catalogação destinados ao usuário, como o IFLA (2016), e os princípios FAIR (Encontrável, Acessível, Interoperável e Reutilizável), destinados também às máquinas.

Os princípios FAIR foram pensados para atender a uma lacuna na recuperação de informações a partir das descrições dos objetos digitais do ponto de vista do seu conteúdo. Uma vez que o ser humano, por sua capacidade de interpretação de contextos, tem a capacidade intuitiva de identificar e classificar os significados de um objeto digital, mas essa capacidade encontra uma barreira quando escala, velocidade e escopo dos objetos digitais aumentam consideravelmente. Nesse caso, o auxílio dos mecanismos computacionais é bem-vindo, e para isso é necessária mais atenção à descrição dos objetos digitais, principalmente quanto ao seu conteúdo (Wilkinson *et al*., 2016).

Um elemento fundamental na implementação dessas premissas de organização dos dados, como a catalogação descritiva e os princípios FAIR é o metadado. O metadado, no contexto do patrimônio cultural, são “informações de valor agregado criadas para organizar, descrever, rastrear e melhorar o acesso a objetos de informação e aos acervos físicos relacionados a esses objetos” (Gilliland, 2016, p. 9, tradução nossa). É o elemento responsável por mediar a representação da informação dos objetos digitais, é através dele que são alocados os registros informacionais que dão vida digital ao objeto do acervo. Além dessas características, os metadados são responsáveis também, quando definidos a partir de conjuntos padronizados, por permitir a interoperabilidade entre os acervos (Guizzardi, 2020; Zeng, 2019), uma vez que esses metadados padronizados podem ser comuns entre os acervos de um mesmo domínio.

Os padrões de metadados são essenciais para a manutenção das premissas de organização e representação da informação nos acervos digitais. A definição dos padrões de metadados está associada ao meio em que os metadados são aplicados,

e tem como objetivo garantir que acervos sob uma mesma temática compartilhem do mesmo conjunto de definições e qualidade dos recursos informacionais, promovendo, inclusive, a troca de dados entre si.

Gilliland (2016) categoriza os padrões de metadados em quatro tipos (Tabela 1): os padrões de estrutura de dados, que são conjuntos ou esquemas compostos por metadados; os padrões de valores de dados, que fazem referência aos sistemas de organização do conhecimento descritivos de conteúdo, como taxonomias e tesauros; os padrões de conteúdos de dados, que delimitam as regras de catalogação e sintaxe dos metadados; e os padrões de intercâmbio técnico dos dados, que também são processados por máquina e permitem a comunicação entre os sistemas de informação dos acervos.

Tabela 1 - Uma tipologia de padrões de metadados

Tipo	Exemplos
Padrões de estrutura de dados (conjuntos de elementos de metadados, esquemas). Estas são "categorias" ou "contêineres" de dados que compõem um registro ou outro objeto de informação.	O conjunto de campos MARC (formato de catalogação legível por máquina), Descrição Arquivística Codificada (EAD), Conjunto de Elementos de Metadados Dublin Core (DCMI), Categorias para a Descrição de Obras de Arte (CDWA), Categorias Principais do VRA (Visual Resources Association).
Padrões de valores de dados (vocabulários controlados, tesauros, listas controladas). Estes são os termos, nomes e outros valores que são utilizados para preencher os padrões de estrutura de dados ou conjuntos de elementos de metadados.	Assuntos de Cabeçalhos da Biblioteca do Congresso (LCSH), Arquivo de Autoridade de Nomes da Biblioteca do Congresso (CNAF), Tesauro para Materiais Gráficos da Biblioteca do Congresso (TGM), Descritores de Assuntos Médicos (MeSH), Tesauro de Arte e Arquitetura (AAT), Lista de Nomes de Artistas (ULAN), Tesauro Getty de Nomes Geográficos (TGN), ICONCLASS.
Padrões de conteúdo de dados (regras de catalogação e códigos). Estas são diretrizes para o formato e a sintaxe dos valores de dados que são utilizados para preencher os elementos de metadados.	Regras de Catalogação Anglo-Americanas (AACR), Descrição e Acesso a Recursos (DA), Descrição Bibliográfica Internacional Normalizada (ISBD), Objetos Culturais de Catalogação (CCO), Padrão de Descrição de Arquivos: Um Padrão de Conteúdo (DACS).
Padrões de formato de dados/intercâmbio técnico (padrões de metadados expressos em forma legível por máquina). Esse tipo de padrão é frequentemente uma manifestação de um padrão de estrutura de dados específico (tipo 1 acima), codificado ou marcado para processamento por máquina.	MARC21, MARCXML, DTD EAD XML, METS, MODS, CDWA Lite XML schema, Esquema XML Dublin Core Simples, Esquema XML Dublin Core Qualificado, Esquema XML VRA Core 4.0.

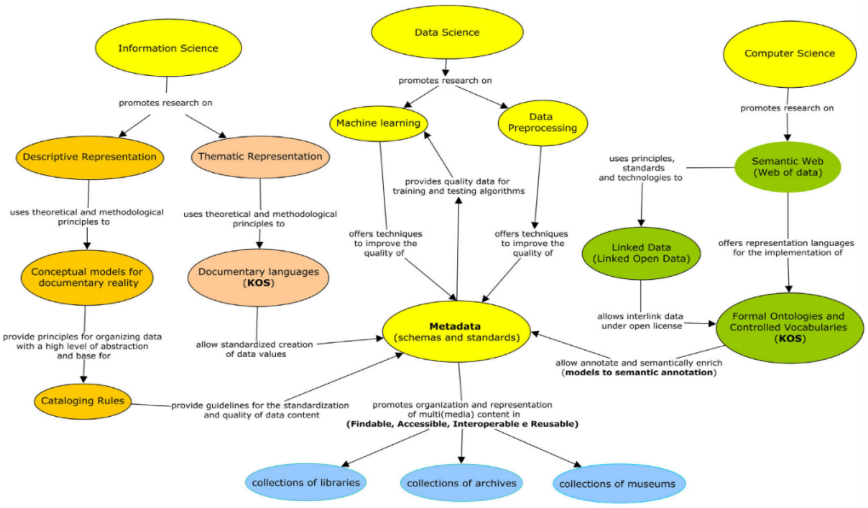
Fonte: Adaptado de Guillain (2016, p. 3, tradução nossa).

Essas categorias de padrões de metadados (Tabela 1) abarcam o contexto de organização e representação da informação que incide sobre acervos digitais. Sendo assim, existe um esforço que busca estruturar o modo como a informação é representada nesses acervos digitais, e que vai ao encontro da necessidade de facilitar a recuperação dessas informações, tanto do ponto de vista da interpretação e busca pelos usuários, quanto pela mediação dos mecanismos computacionais envolvidos na otimização da escala e complexidade desses acervos.

Do ponto de vista dos campos de pesquisa e de sua interconexão, é possível perceber que esse esforço de organização e representação da informação tem como centro os metadados, e como elementos relacionados, os processos teóricos e aplicados provenientes das áreas da ciência da informação, ciência de dados e ciência da computação (Figura 1).

O diagrama da Figura 1 apresenta como a interdisciplinaridade entre essas áreas é concebida através dos estudos e aplicações relacionadas aos metadados. Existe uma gestão da informação representada pelos metadados, o que promove a organização e representação dos objetos digitais, dentro da lógica dos princípios FAIR, em acervos do patrimônio cultural.

Figura 1 - Representação de acervos digitais: dos metadados à anotação semântica.



Fonte: Martins *et al*. (2022, p. 7).

Este diagrama da Figura 1 ainda apresenta um conjunto de elementos relacionados a aplicações semânticas no processamento dos metadados. Essas aplicações semânticas são provenientes da necessidade de lidar com a divergência entre os diferentes formatos de dados (Martins *et al*., 2022), como o JSON (Notação

de objetos JavaScript), muito comum em sistemas que utilizam API (Interface de Programação de Aplicação), e o XML, mais comum em plataformas de acervos digitais. Essas diferenças chegam a incidir nas características semânticas e sintáticas dos metadados, e sistemas de organização do conhecimento simples (SKOS) são utilizados para sanar a interoperabilidade semântica (Guizzardi, 2020) entre esses sistemas heterogêneos.

Uma parcela desses SKOS é voltada para vincular rótulos aos objetos digitais dos acervos online (Catalano *et al*., 2020; Hyvönen *et al*., 2016; Lemos;Souza, 2020; Messaoudi *et al*., 2018; Robledano-Arillo *et al*., 2020). Esse processo de rotulagem vinculada ocorre entre acervos, possibilitando a vinculação da informação de metadados em outras instâncias com contextos sob outras temáticas, promovendo a contextualização dos objetos digitais através da interoperabilidade semântica entre os acervos. Assim, os metadados antes relacionados somente aos objetos em seu acervo digital de origem, agora podem servir de insumo para o enriquecimento de metadados em outras instâncias.

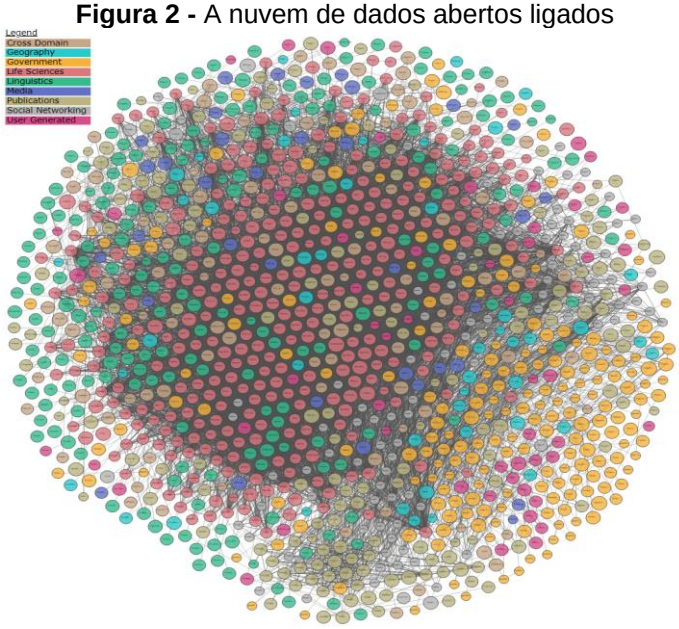
A vinculação de metadados é promovida no âmbito dos dados abertos na web, uma vez que se faz necessário publicar e disponibilizar para reúso os metadados para então vinculá-los a outros acervos. Esse processo compreende um conjunto de princípios e tecnologias chamado de dados abertos ligados (LOD) (Bizer; Heath; Berners-Lee, 2009; Machado; Souza;Simões, 2019).

A premissa da iniciativa dos dados abertos ligados pode ser tipificada em cinco níveis de “abertura” e “ligação” de dados (Berners-Lee, 2009): nível 1, os dados devem estar disponíveis na internet; nível 2, os dados devem estar disponíveis como dados estruturados e legíveis por máquina (permitir processamento computacional automático); nível 3, além das premissas dos níveis anteriores, os dados devem estar disponíveis em formatos não-proprietários; nível 4, todas as premissas dos níveis anteriores, e os dados, devem utilizar padrões abertos, como RDF (*Resource Description Framework*) ou SPARQL, que permitem a identificação e recuperação do objeto e suas relações; e nível 5, incluir as premissas dos níveis anteriores, e ainda ligar os dados a outros provedores;

Todas essas premissas e a busca por abertura de dados, principalmente governamentais, formam uma nuvem de relações entre os provedores de dados (Figura 2). Essa rede de conexões traz um grande potencial de contextualização dos objetos digitais, elevando a capacidade de representação da informação e

organização do conhecimento desses objetos através da interoperabilidade semântica.

Figura 2 - A nuvem de dados abertos ligados



Fonte: lod-cloud.net, 2023.

Para assegurar a manutenção e promoção de sistemas de organização da informação como esse da iniciativa de dados abertos ligados, além de levar em consideração os padrões de metadados e a aplicação dos KOS, o cuidado com a qualidade dos dados é importante. Existem situações, inclusive no contexto dos acervos digitais do patrimônio cultural brasileiro, em que são notadas divergências na aplicação dos padrões de metadados, e por vezes a qualidade dos dados é comprometida, limitando as possibilidades de interoperabilidade e recuperação da informação, uma vez que há incidência de variações na forma como os dados são apresentados (Siqueira; Martins, 2022; Martins *et al*. 2021; Lemos; Martins; Carmo, 2022).

Para esses casos, conforme apresentado na Figura 1, o campo da ciência de dados, a partir das aplicações de pré-processamento, que envolvem análise, normalização e padronização dos dados, bem como o uso de aprendizagem de máquina para auxiliar na normalização e tratamento de dados, é importante para promover a qualidade dos dados (Virkus; Garoufallou, 2020). Assim, entende-se que a ciência de dados se mostra uma área de desenvolvimento profissional importante para o contexto do patrimônio cultural (Almeida, 2013; Guizzardi, 2020; Virkus; Garoufallou, 2019).

As aplicações dos processos desta área vão além do pré-processamento, normalização e tratamento, e podem auxiliar ainda no processo de geração automática e semiautomática de metadados (Colla *et al*., 2022; Park; Brenza *et al*., 2015), através de aplicações que permitem identificar padrões e agrupamentos entre os dados (EUROPEANA TECH, 2021; IFLA, 2020). Ainda, é possível citar a capacidade de complementar a produção de pesquisas e relatórios sobre os acervos por meio de produção de estatísticas e métricas que representam indicadores relacionados a estes contextos (Greenberg, 2017).

Em síntese, é possível situar os acervos digitais do patrimônio cultural brasileiro na contextualização da ciência da informação, de modo a envolver componentes interdisciplinares, que relacionam também as áreas de ciência da computação e ciência de dados, bem como suas aplicações relacionadas aos metadados e aos conceitos latentes de representação da informação e sistemas de organização do conhecimento. Esse conjunto de aplicações e teorias permite estruturar os pilares científicos que sustentam o objeto em estudo e as metodologias aplicadas para estudá-lo.

2.2 Mapeamento de metadados para agregação de acervos digitais culturais

Um dos estágios mais importantes da agregação de acervos digitais é o mapeamento. Este é o processo em que o modelo semântico dos provedores de dados é reconciliado com o modelo semântico escolhido para agregação. Essa reconciliação entre o modelo de origem e o modelo de agregação nem sempre é equivalente, por vezes o conjunto de metadados do acervo de origem diverge do conjunto de metadados selecionados para agregação, bem como os vocabulários controlados utilizados podem ser diferentes, por exemplo. Por isso o mapeamento se faz necessário, para unificar essas diferenças em um único modelo semântico, e padronizar a forma de organização do conhecimento desses objetos digitais culturais.

O conceito de modelo semântico utilizado na presente pesquisa é baseado no referencial conceitual sobre organização do conhecimento e representação da informação, conforme apresentado no tópico 2.1. Parte-se do entendimento de que os padrões de metadados (Gilliland, 2016) e os sistemas de organização do conhecimento (KOS) (Hodge, 2000; Hjorland, 2007) são elementos que compõem a estrutura de um modelo semântico, e tais elementos são responsáveis pela

representação do significado dos objetos digitais culturais dos acervos em estudo (Abbas, 2010; Gilliland, 2016; IFLA, 2009; Zeng; Qin, 2016).

2.2.1 Interoperabilidade e mapeamento de metadados

Com base no exposto acima, destaca-se que um dos produtos desse processo de mapeamento é a interoperabilidade dos metadados (Chan; Zeng, 2006a; CHAN; Zeng, 2006b; Haslhofer; Klas, 2010). Neste processo, uma vez estabelecida a reconciliação entre os esquemas de metadados dos provedores e do modelo de agregação, os metadados se tornam interoperáveis, passíveis de processamentos automáticos por máquina, permitindo a sustentabilidade do processo de agregação dos acervos.

Haslhofer e Klas (2010) definem a interoperabilidade de metadados em três níveis:

> [...] em um nível técnico mais baixo, as máquinas devem ser capazes de se comunicar umas com as outras para acessar e trocar metadados. Em um nível técnico mais elevado, uma máquina deve ser capaz de processar as informações de objetos de metadados recebidos de outra máquina. E em um nível semântico muito alto, devemos garantir que máquinas e seres humanos interpretem corretamente os significados pretendidos dos metadados (Haslhofer; Klas, 2010, p. 2, tradução nossa).

Dessa forma, a interoperabilidade de metadados é fundamental para o processo de agregação de acervos digitais como um todo, garantindo que os dados coletados passarão pelo processo de reconciliação como definido na etapa de mapeamento. O resultado dessa interoperabilidade afeta diretamente como os dados serão disponibilizados ao público. A depender da forma como a interoperabilidade é realizada, o contexto dos metadados pode ser reduzido (caso o modelo semântico de agregação seja menos abrangente que o modelo do provedor, por exemplo).

Chan e Zeng (2006a, 2006b) , ao elaborar uma revisão sobre esforços para interoperabilidade de metadados, identificaram que esses podem se concentrar em três níveis de execução: 1 – Nível de esquema, quando os elementos de um esquema de metadados são os principais componentes do processo de interoperabilidade. Nesse caso um ou mais elementos são analisados e se chega a um esquema de metadados que represente todos, sem necessariamente o uso em uma aplicação; 2 – Nível de registros, que leva em consideração os registros relacionados em cada metadado, em que a semântica é importante. Nesse caso, o produto normalmente é

gerado pela combinação dos registros para produção de registros normalizados; 3 – Nível de repositório, que está relacionado a iniciativas que integram várias fontes de dados, e o foco está na reconciliação entre valores e elementos normalizados (metadados na origem com valores categóricos mapeados para o metadado de agregação de tipo, por exemplo.).

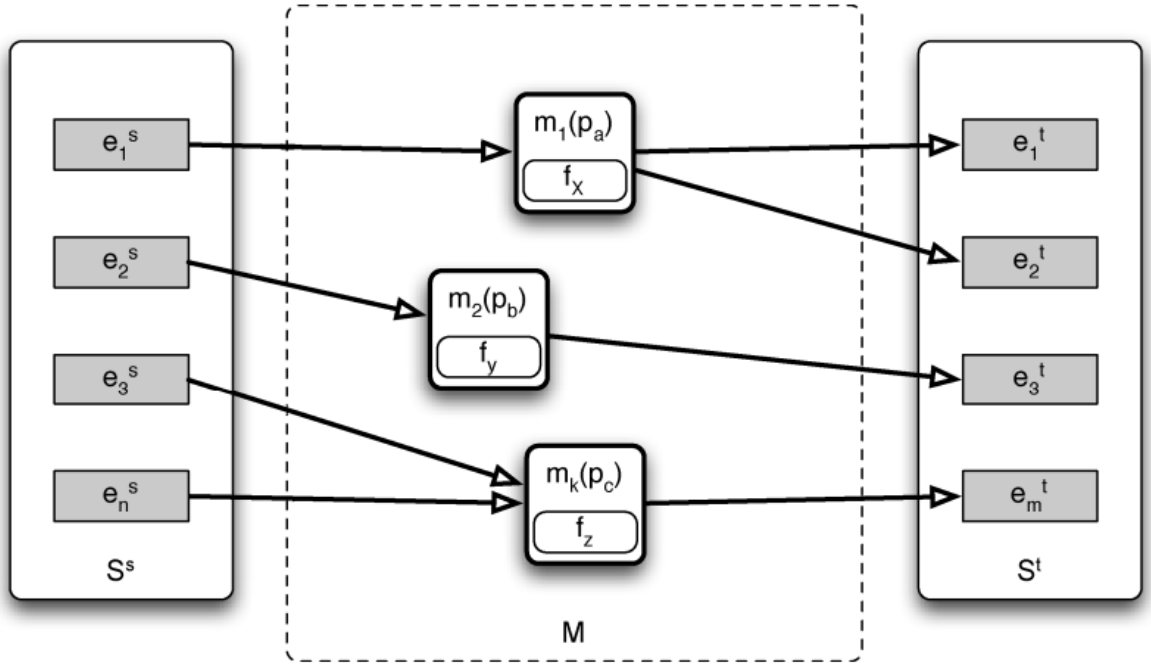
Dentre os níveis identificados por Chan e Zeng (2006a, 2006b), aquele que mais se aproxima do levantamento de iniciativas de diferentes nações (tópico 3), e a proposta desta pesquisa é o nível 3, já que se busca contexto de integração de várias fontes de dados diferentes. Segundo as autoras, neste nível foram identificados seis tipos de processos de interoperabilidade. Dentre eles, o processo de agregação é o que mais se aproxima em relação ao que foi levantado na presente pesquisa.

O processo de interoperabilidade por agregação, segundo a revisão Chan e Zeng (2006a, 2006b), apresentou a necessidade de lidar com a existência problemas nos metadados coletados dos provedores, tais como dados faltando, dados incorretos, dados confusos ou/e, ainda, dados insuficientes. Nesse caso, como o serviço de agregação coleta dados de várias fontes, é realizada uma verificação dos valores de metadados e identificadores do objeto digital coletado, bem como a busca destes dados em outras fontes, com o objetivo de enriquecer esse objeto digital com informação de outros provedores (por exemplo, um objeto digital vem com a informação de nome de autor, com isso é possível coletar outros dados sobre esse autor em uma fonte sobre autoridades, de modo a complementar suas informações).

Além desse tipo de mapeamento de metadados, outros processos básicos compõem o esforço de mapear modelos semânticos, como o processo de *crosswalk,* que, dentre outras abordagens, utiliza funções de transformação, levando em consideração a possibilidade de cardinalidade do tipo 1:1 ou n:1 (muitos elementos da fonte para um elemento do modelo de agregação) (Haslhofer; Klas, 2010).

A Figura 3 mostra como os metadados da fonte de dados (esquerda) são mapeados para uma expressão de mapeamento que relaciona o metadado do provedor com o metadado do modelo de agregação, quando necessário, utilizando uma função de transformação e derivando para o metadado do modelo de agregação (direita).

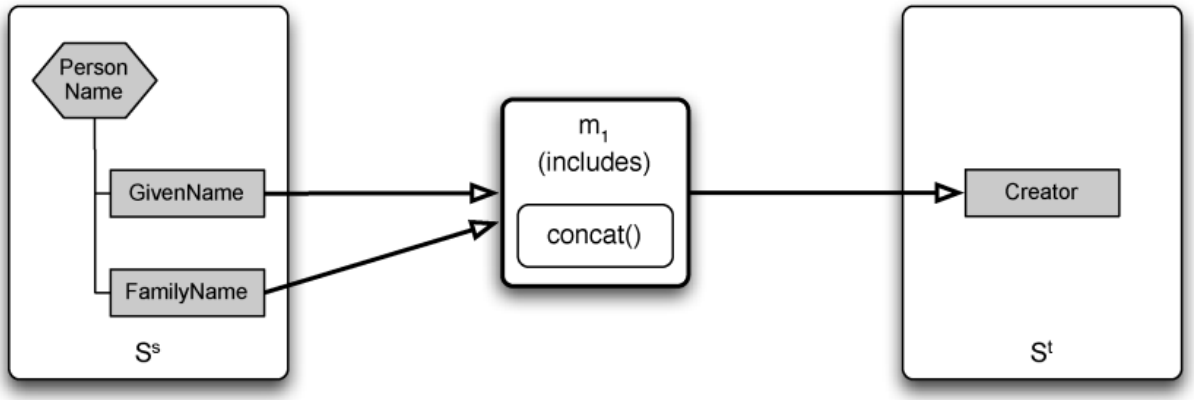
Figura 3 - Elementos principais de uma especificação de mapeamento de metadados



Fonte: Haslhofer; Klas, 2010, p. 26.

Essa função de transformação citada é responsável por normalizar o valor do metadado no processo de mapeamento, de modo que ele vá se adequar ao modelo de agregação objetivo. A Figura 4 dá o exemplo de um metadado *Person Name* (Nome pessoal), que contém dois valores, o *Given Name* (Prenome) e o *Family Name* (Sobrenome). Esses valores passam pela expressão de mapeamento de inclusão no metadado de agregação *Creator* (Autor), utilizando a função de concatenação, para juntar os dois valores em um só.

Figura 4 - Alcançando a interoperabilidade de metadados por meio da transformação de instâncias



Fonte: Haslhofer; Klas, 2010, p. 27.

Esses são alguns exemplos dos possíveis processos envolvidos no mapeamento de metadados entre provedores de dados. Com esses tipos de processos, necessários para garantir a interoperabilidade entre os metadados, é possível efetivar um mapeamento entre os provedores de dados e o serviço de agregação. No caso das instituições do patrimônio cultural (bibliotecas, museus, arquivos e galerias), que são entidades com distintos modelos semânticos aplicados aos seus acervos, executar um mapeamento de metadados para agregação desses acervos se apresenta como um desafio.

Uma das possibilidades práticas de implementação do crosswalk se dá a partir de modelos prontos de crosswalk entre diferentes padrões de metadados. Esses tipos de modelo são uma indicação da relação de correspondência entre os metadados de um padrão com outro, e para isso é necessário que os acervos utilizem de maneira correta um padrão de metadados formal.

O Dublin Core, por exemplo, tem como parte da sua documentação uma tabela com a indicação do crosswalk com outros padrões (HARPRING, 2022), que pode ser utilizada para guiar iniciativas de agregação cujos acervos apresentem o uso de padrões heterogêneos.

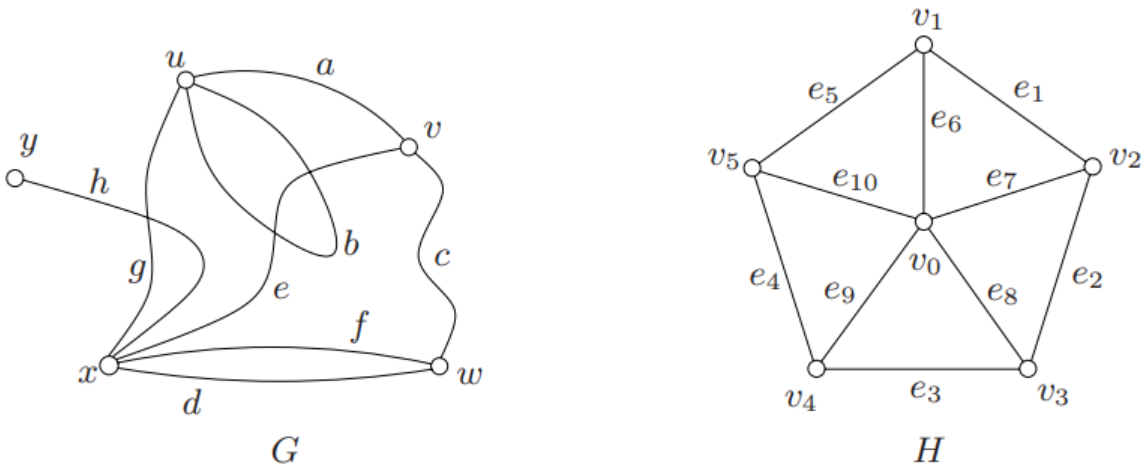
2.2.2 Visualização da interoperabilidade e mapeamento de metadados através de grafos

A interoperabilidade e mapeamento de metadados tem como processo principal a reconciliação dos metadados utilizados pelos provedores e os metadados definidos no modelo semântico de agregação (Chan; Zeng, 2006a; Chan; Zeng, 2006b; Haslhofer; Klas, 2010). Entende-se que a busca por descrever como esse processo se dá pode se tornar mais clara a partir de uma representação visual dessa reconciliação. Dessa forma, como este processo de reconciliação é um relacionamento entre um ou mais metadados do provedor para um metadado do modelo de agregação, a visualização através de grafos pode atender à necessidade de enxergar essas conexões entre os metadados, tornando o processo visualmente mais claro e facilitando sua compreensão.

Um grafo é representado por um conjunto de vértices e arestas. Os vértices representam os indivíduos e as arestas as relações, de modo que a visualização dos

vértices é geralmente indicada por pontos, e as arestas são indicadas por linhas que ligam esses pontos (Figura 5).

Figura 5 - Diagrama dos grafos G e H



Fonte: Bondy; Murty, 2008, p. 3.

Um grafo é construído por uma matriz de relações, em que é necessário indicar qual indivíduo se relaciona com qual outro indivíduo, quantas vezes e como, se é de forma direcional (quando a ligação parte de um indivíduo para o outro), ou não (Bondy; Murty, 2008). Essa matriz pode ser desenvolvida a partir da análise de relação entre os metadados dos provedores e os metadados do modelo de agregação, bem como a quantidade de relações pode ser definida pela co-ocorrência dos metadados entre provedores.

Esse tipo de abordagem, que envolve o uso de visualização de grafos no contexto informacional, não é novidade, como indica Cherven (2015):

> Grafos que examinam o fluxo de informações pela Web são típicos dessa categoria de análise de rede. Essas conexões podem fazer referência a qualquer coisa, desde conexões entre blogueiros, páginas na Wikipedia, ou entre redes de citações de artigos científicos. Este é um tipo muito popular de gráfico, dada a acessibilidade das informações via Web e suas diversas aplicações (Cherven, 2015, p. 8, tradução nossa).

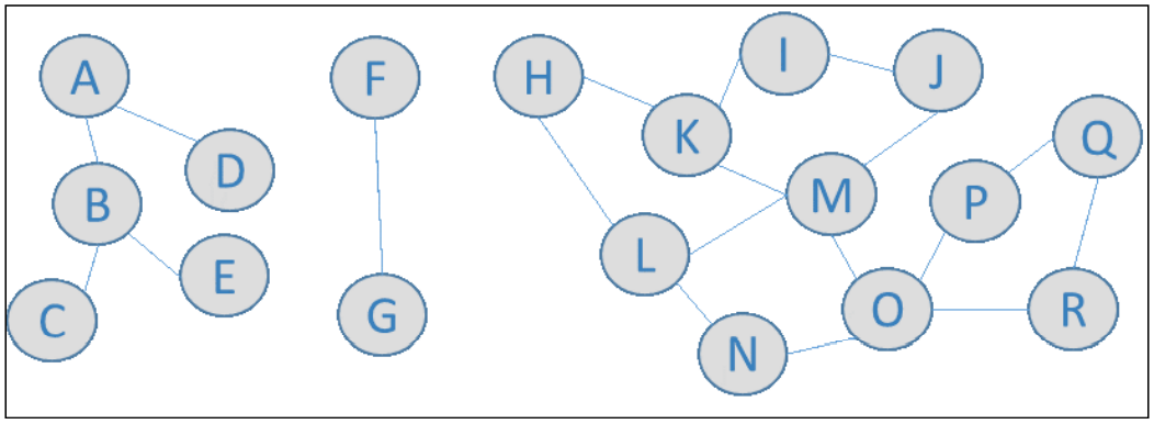
Além da capacidade de visualização das relações entre os metadados através do elemento visual do grafo, a teoria por trás desses tipos de elemento ainda é composta por métricas de relacionamento. Essas métricas definem como o grafo é exibido a partir das medidas entre os vértices e suas relações.

Um exemplo de métrica importante para o caso dos grafos de mapeamento de metadados é a medida de centralidade, que busca “compreender a influência

relativa de nós individuais dentro da rede” (Cherven, 2015, p.14). Nesse contexto, destaca-se a métrica de centralidade de grau, que mensura a quantidade de vértices ligados a um vértice específico.

Outro componente presente na visualização por grafos é o agrupamento, que é composto por componentes da rede que são formados através de vértices que têm relação entre si. No caso da agregação de acervos, por exemplo, espera-se que bibliotecas compartilhem um modelo de metadados semelhantes entre si, e forme um grande componente de agrupamento na visualização do grafo (Figura 6).

Figura 6 - Um grande componente composto pelos nós H a R



Fonte: Cherven, 2015, p.18

Esses tipos de métricas e características auxiliam a mostrar quais os metadados do modelo semântico de agregação são mais representativos em relação aos metadados utilizados pelos modelos aplicados pelos provedores, bem como indicar quais provedores de dados utilizam modelos semânticos semelhantes.

Dessa forma, diante do exposto, entende-se que o uso desse tipo de abordagem visual no contexto da presente pesquisa, é de cunho descritivo, e fundamental para complementar a caracterização dos processos de padronização dos dados coletados dos acervos digitais a partir de um padrão de metadados, que é um dos objetivos aqui aplicados.

3 INICIATIVAS DE AGREGAÇÃO DE ACERVOS DIGITAIS DO PATRIMÔNIO CULTURAL

Cientes dos desafios e complexidade existentes no processo de representação da informação e organização do conhecimento de acervos digitais, algumas instituições de memória buscam facilitar a difusão e recuperação

informacional sobre a cultura de seus países a partir da ótica e aplicação da interoperabilidade entre seus acervos digitais, de modo a promover serviços de agregação que possibilitem a busca e reúso dos objetos digitais culturais por meio do acesso a uma plataforma online única.

Visto que a presente pesquisa se situa no escopo de agregação de acervos digitais do patrimônio cultural em contexto brasileiro, este tópico elenca algumas iniciativas de agregação de acervos no âmbito de outros países, em busca de revisar as características tecnológicas de implementação e os mecanismos informacionais e computacionais envolvidos nesse processo.

3.1 Europeana

A Europeana (Figura 7) é um dos exemplos de agregação de acervos digitais mais expressivos. A iniciativa reúne acervos bibliográficos, arquivísticos e museológicos, de aproximadamente 4.000 instituições do patrimônio cultural de ao menos 24 países europeus. O conjunto total é composto por mais de 51 milhões de arquivos de mídia, entre imagem, áudio, texto, vídeo e objetos 3D.

Figura 7 - Página inicial do repositório Europeana

Fonte: Europeana, 2023.

3.1.1 Síntese sobre a iniciativa da Europeana

A idealização da Europeana surgiu formalmente em meados de 2005, por meio de uma carta conjunta de representantes governamentais de França, Espanha, Itália, Polônia e Hungria, enviada ao então presidente da comissão europeia recomendando a criação de um ambiente digital para democratização do patrimônio cultural europeu (Purday, 2009).

Inicialmente, tal ambiente foi concebido no contexto das bibliotecas europeias, através de uma prova de conceito que envolveu bibliotecas de Portugal, Hungria e França, e a agregação de aproximadamente 12.000 livros digitalizados dessas bibliotecas. Posteriormente, o desenvolvimento do serviço digital da Biblioteca Europeia, que efetivamente agregou objetos digitais de bibliotecas de mais de 46 países europeus, entre outros projetos semelhantes, fundamentou as perspectivas de implementação do projeto da Europeana (Purday, 2009).

A partir da comoção de diferentes entidades do patrimônio cultural europeu, inclusive serviços de agregação já consolidados em alguns países, a iniciativa da Europeana ganhou a perspectiva de atuação no formato de uma rede de agregadores, que asseguram que os acervos das instituições integradas estejam conforme os padrões exigidos para agregação, cujo principal elemento é o padrão de metadados denominado EDM (Europeana Data Model), que também contempla um conjunto de metadados provenientes do Dublin Core, e é utilizado para mapear em um único padrão de agregação, os metadados utilizados nos acervos digitais das instituições de patrimônio cultural vinculadas ao agregador (Doerr *et al*., 2010b).

O repositório que agrega os acervos, além de facilitar o acesso e recuperação da informação dos objetos digitais culturais, aumenta e promove o reúso desses objetos, por meio da disponibilização de consulta por 6 tipos de APIs, que permitem o desenvolvimento de aplicações que consomem os dados do repositório. Além disso, os dados ainda estão disponíveis no formato de dados abertos ligados, disponíveis para consulta via SPARQL (SPARQL Protocol and RDF Query Language).

3.1.2 Aspectos técnicos da Europeana

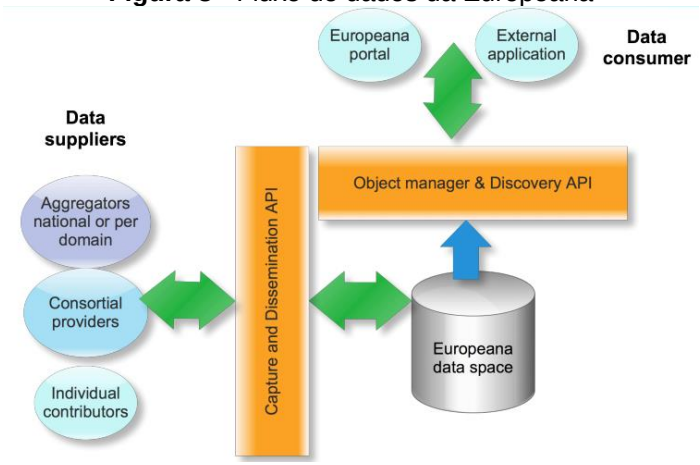
O primeiro ponto a ser esclarecido quanto aos aspectos técnicos da Europeana é a percepção do conjunto de estruturas da iniciativa. Como apresentado no tópico anterior, a Europeana também é composta por uma estrutura de articulações

em rede entre instituições do patrimônio cultural, além das articulações políticas entre os países europeus.

No aspecto da estrutura tecnológica, a Europeana pode ser compreendida como um sistema de biblioteca digital, baseado em uma plataforma prestadora de serviços através de APIs, que por sua vez permitem o acesso aos objetos culturais digitais e fornece esse acesso através de um portal web (Concordia, 2009).

Como apresenta a Figura 8 abaixo, essa concepção tecnológica da Europeana é fundamentada em um espaço de dados para armazenamento dos objetos digitais culturais, e algumas APIs gerenciam os dados desse espaço em duas vertentes: uma para os provedores, e outra para os consumidores, que podem ser usuários do portal web, ou outras aplicações que podem consumir o conteúdo da API.

Figura 8 - Fluxo de dados da Europeana



Fonte: Concordia, 2009.

No caso do acesso ao sistema de biblioteca digital pelos provedores, são considerados os repositórios de objetos digitais culturais, sejam eles sistemas que contenham um único acervo (como o acervo de um museu), ou sistemas que já fazem o papel de agregação nacional ou por domínio, e ainda sistemas de instituições que fazem parte de um consórcio entre si. A API nesse caso é utilizada para capturar os dados dos sistemas desses fornecedores e para disseminar os dados de volta, em um processo de enriquecimento do acervo do fornecedor com os dados agregados pela Europeana.

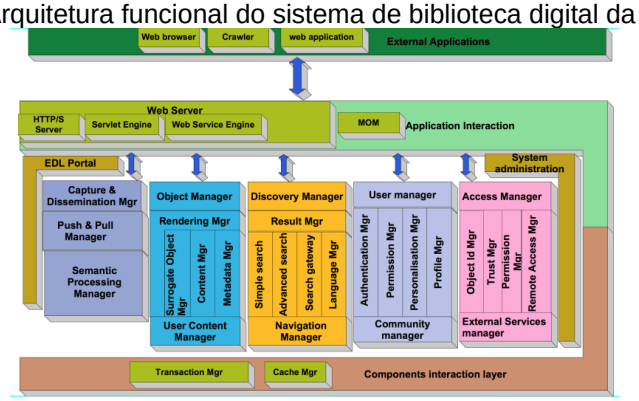
Esse processo de coleta dos dados dos provedores se dá em duas instâncias, sendo a primeira referente à adequação do provedor, que envolve 5 passos (EUROPEANA , 2023): 1 – entender os benefícios de se fornecer dados para a

Europeana, 2 – atender aos requisitos básicos de compartilhamento de dados com a Europeana, que prevê que o conteúdo deve estar digitalizado, o acervo deve ser sobre a Europa, e o escopo do acervo deve estar alinhado com a estratégia de conteúdo da Europeana; 3 – atender aos critérios técnicos, uso do formado EDM na submissão dos metadados, e alinhamento com as camadas de conteúdo e de metadados da Europeana; 4 – dados licenciados para os objetos digitais culturais a serem compartilhados; 5 – compartilhar os dados dos objetos com agregadores locais que já têm uma conexão com a Europeana. A segunda refere-se à instância tecnológica, que é dividida em dois modos, um para lidar com jornais, e outro para lidar com os demais tipos de conteúdo. Nos dois modos é comum um procedimento de processamento dos dados e mídia coletados, chamado de gestor unificado e ingestão, que carrega os dados e os transforma no modelo especificado para a base de dados, além de transformar as imagens para um tipo unificado de arquivo e processar os dados para indexação e relacionamentos com outros dados. No caso dos objetos, além dos jornais, existe um processo de enriquecimento que relaciona os objetos com outros já existentes na base, e enriquece os dados sobre ele no momento da ingestão. Após esse processo de carga e transformação, os dados são armazenados em bancos de dados específicos, assim como as imagens, que também têm um banco para armazenamento (Lefferts *et al*. 2015).

Em relação ao acesso ao sistema de biblioteca digital pelos consumidores, os dados coletados dos fornecedores são consumidos por dois processos, o gerenciador de objetos e uma API para descoberta. Estes dois meios contêm diferentes tecnologias para suportar as necessidades de acesso através do portal web e pelo consumo direto por aplicações externas através da API de descoberta.

Esse fluxo de dados entre as partes interessadas e a base de dados sobre os objetos digitais culturais é mantido por um conjunto de gerenciadores do conteúdo e dos dados, além de tecnologias específicas das camadas da arquitetura funcional da Europeana (Figura 9).

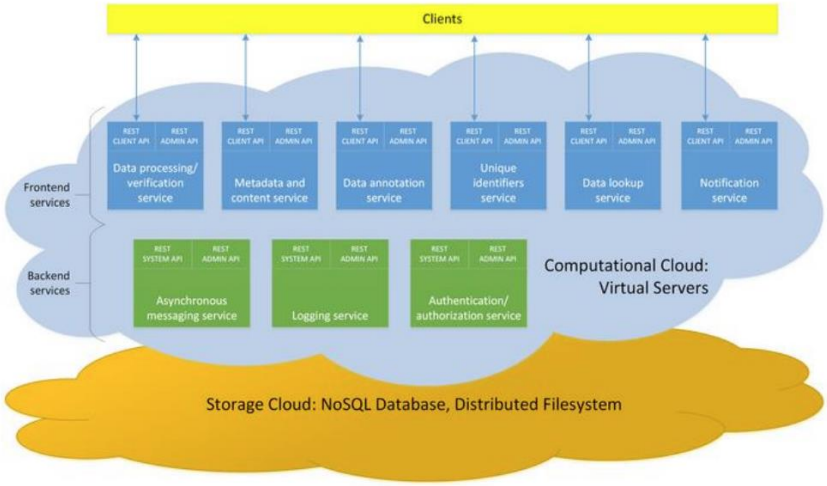
Figura 9 - Arquitetura funcional do sistema de biblioteca digital da Europeana



Fonte: Concordia, 2009.

Tais gerenciadores são programados através das APIs, disponíveis em servidores virtuais em nuvem, separados por serviços de APIs que funcionam para o *frontend* (camadas visuais de interação, como o portal web), e *backend* (camadas internas de processamento como a autenticação de usuários). Já o armazenamento utiliza tecnologias como o NoSQL (forma de armazenamento de dados sem o padrão de relacionamento entre tabelas, e que permite maior escalabilidade e velocidade de processamento), e sistema de arquivos distribuídos (que permite armazenar objetos digitais em diferentes lugares, aumentando também a escalabilidade) (Figura 10).

Figura 10 - Arquitetura do sistema de nuvem da Europeana

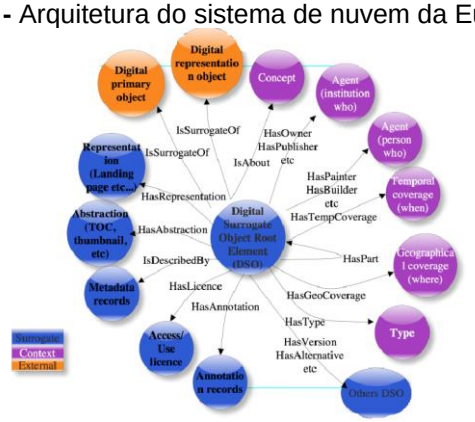


Fonte: Adamski *et al.*, 2017.

Em síntese, as características técnicas da Europeana funcionam sob uma arquitetura tecnológica fundamentada no uso de um espaço de dados, gerenciado por um conjunto de APIs, e exposto ao público na internet através dos serviços de descoberta e portal web.

Vale ainda salientar, que o espaço de dados da Europeana existe a partir de uma estrutura de organização do conhecimento e representação da informação para garantir a padronização dos dados sobre os objetos digitais. Essa estrutura é chamada de objeto substituto, formado por três camadas: a camada semântica do próprio objeto, proveniente dos metadados dos objetos dos provedores; uma camada de contextualização que situa as características do objeto, como instituição de origem, conceito e tipo; e uma camada externa que se relaciona com o objeto digital de origem (Figura 11).

Figura 11 - Arquitetura do sistema de nuvem da Europeana



Fonte: Concordia, 2009.

Os objetos coletados dos repositórios digitais de outras instituições devem se adequar a um mesmo formato de representação, que no caso é o modelo semântico EDM. Para isso, a arquitetura dos dados leva em conta a criação de um objeto substituto na biblioteca digital da Europeana, e esse objeto é constituído a partir de um mapeamento dos padrões de metadados utilizados pelos provedores de dados, para o modelo EDM, e na representação da mídia original do objeto (Dekkers; Gradmann; Meghini, 2009).

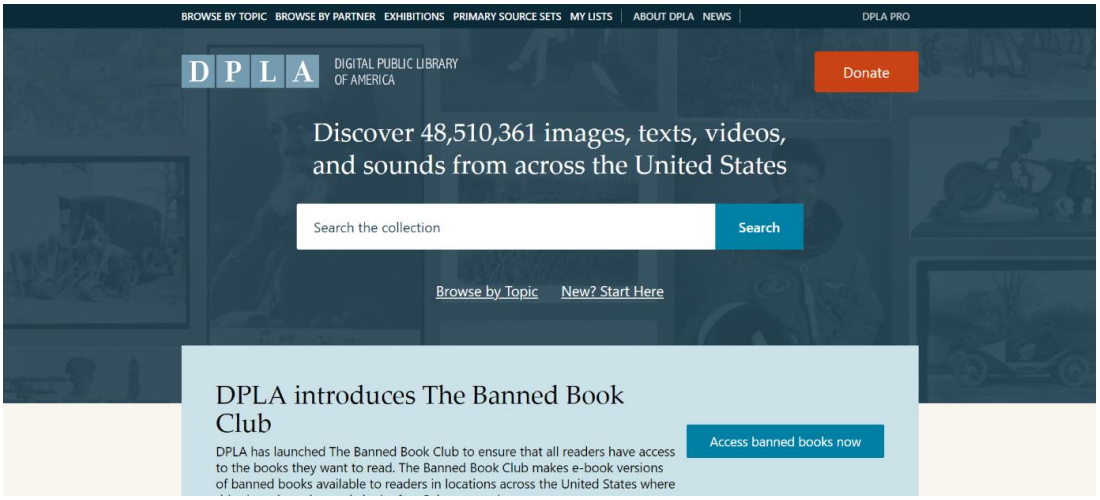
Assim, a arquitetura tecnológica da Europeana prevê uma etapa de coleta dos dados dos provedores, que envolve uma instância de adequação do provedor, e uma instância de processamento tecnológica. Esses dados coletados são armazenados e processados por gerenciadores baseados em APIs para disponibilizar os dados para acesso pelo portal web e acesso via API para demais aplicações. Essa etapa de coleta é importante, pois os dados são heterogêneos, estão sob os subdomínios do setor cultural (artes visuais, artes plásticas, música, manuscritos, jornais, etc.), e são provenientes de diferentes instituições (museus, galerias, bibliotecas, arquivos). Agregar esses tipos de objetos digitais demanda um processo de organização do

conhecimento e representação da informação em torno da proposição de um modelo unificado de metadados (no caso o EDM), e o mapeamento dos metadados dos provedores para esse modelo. Ainda, vale ressaltar a representação desse processo na estrutura do objeto digital substituto e o enriquecimento dos dados com o relacionamento com outros bancos de dados.

3.2 Digital Public Library of America (DPLA)

Outro exemplo importante de agregação de acervos digitais culturais internacional é a Digital Public Library of America (DPLA), que reúne mais de 46 milhões de registros entre arquivos de som, áudio, vídeo, imagens e texto de acervos de bibliotecas, arquivos, museus e outras instituições culturais de mais de 76% dos estados dos Estados Unidos (Clink, 2014).

Figura 12 - Página inicial do repositório DPLA



Fonte: Digital Public Library of America, 2023.

3.2.1 Síntese sobre a iniciativa da DPLA

A agregação de acervos digitais na DPLA ocorre através de integração de acervos provenientes de fontes de informação (Hub) de serviço, que são repositórios que também reúnem objetos digitais de forma temática/contextual, e fontes de informação de conteúdo, que são repositórios com acervos digitais institucionais (Triques; Gonçales; Albuquerque, 2023).

Para fazer parte da DPLA como uma fonte de informação de serviço, de forma geral, é necessário representar um grupo de repositórios digitais regionais através de um único ponto de acesso e disponibilizar ao menos 50 mil registros únicos. Já para fontes de informação de conteúdo, é preciso disponibilizar ao menos 150 mil registros únicos, e se comprometer com a manutenção e enriquecimento dos registros.

Os dados das fontes de informação são obtidos em sua maioria através do protocolo OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting), que permite a comunicação entre sistemas para coleta de dados. Já o processo de integração dos metadados dos diferentes acervos culturais agregados pela DPLA é executado através de um perfil de aplicação de metadados MAP (Metadata Profile Application), baseado no EDM da Europeana, flexível o suficiente para atender ao mapeamento de diferentes tipos de padrões de metadados utilizados pelas fontes de informação (DPLA, 2017).

A DPLA também disponibiliza o acesso a uma API, mediante solicitação de chave de acesso, para consumir os dados do repositório agregado e promover diferentes tipos de aplicações de reúso destes dados.

3.2.2 Aspectos técnicos da DPLA

O levantamento dos aspectos técnicos que envolvem o serviço de agregação da DPLA foi baseado nas informações disponíveis em seu portal web, como algumas orientações sobre qualidade e mapeamento de metadados para os Hubs, e referências para o repositório de código.

De acordo com as informações encontradas no repositório de códigos (DPLA, 2023a), é possível abstrair um conjunto de quatro grandes processos tecnológicos: a ingestão/coleta de dados; o desenvolvimento de um painel analítico dos hubs da DPLA; uma API para gerenciar o consumo e exposição dos dados; e uma aplicação para o portal web da DPLA, que garante o acesso à busca e navegação do usuário.

O processo de ingestão/coleta de dados da DPLA é o responsável por “coletar, mapear e enriquecer metadados do patrimônio cultural da rede de parceiros da DPLA para o perfil de aplicação de metadados da DPLA.” (DPLA, 2023b, p.1). Esse processo envolve as seguintes etapas: 1 – Coleta; 2 – Mapeamento e validação; 3 – Enriquecimento e normalização; 4 – Conexão com a Wikimedia.

A etapa de coleta é realizada a partir dos Hubs da DPLA, em que a conexão para coleta dos objetos digitais é realizada principalmente através de três categorias: API (15%), OAI (53%) e arquivos (32%) (DPLA, 2023c). Essas categorias são as formas de comunicação entre o repositório de agregação da DPLA e os repositórios dos Hubs que compartilham seus objetos digitais.

A etapa de mapeamento e validação é representada pela necessidade de cada provedor de dados (Hub) ter um modelo de mapeamento que indica quais metadados de origem são correspondentes aos metadados do perfil de aplicação da DPLA (DPLA, 2023d). Por exemplo, se o provedor de origem é um agregador de bibliotecas digitais e utiliza como modelo de metadados o Dublin Core, os elementos desse modelo devem ter seus correspondentes indicados no perfil de aplicação da DPLA. O perfil de aplicação de metadados da DPLA está em sua versão 5.0 (DPLA, 2017), e é baseado principalmente no modelo utilizado pela Europeana, o EDM, com algumas premissas do Dublin Core. Esse perfil de aplicação pode ser interpretado como um mecanismo de organização do conhecimento e representação da informação, uma vez que define algumas premissas, tais como os elementos de representação do objeto digital cultural, os sistemas de organização do conhecimento (KOS), como vocabulários e listas de termos utilizadas para classificar e categorizar o objeto, e um modelo de representação das relações entre as informações e a mídia do objeto (Abbas, 2010; Gilliland, 2016; IFLA, 2009; Zeng; Qin, 2016, Hjorland, 2007).

Ainda dentro da etapa de mapeamento e validação existe o processo de filtragem, que basicamente trata ocorrências comuns de erro no mapeamento, principalmente nos valores dos metadados de extensão (*extent*), tipo (*type*) e formato (*format*). No caso do metadado de extensão, são procurados todos os valores que são semelhantes a medidas existentes em outros metadados e inseridos no metadado de extensão. Já no caso do metadado tipo e formato, são buscados todos os valores de tipos dentro do metadado de formato, que é recorrente, e os valores são realocados no metadado correto. Continuando na etapa de mapeamento e validação, ocorre o processo de validação, que assegura o cumprimento das regras de mapeamento para o perfil de aplicação da DPLA. Nessa etapa ocorrem dois processos comuns, a validação de preenchimento dos metadados obrigatórios, e a normalização e validação do metadado de licença *edmRights*, utilizado para indicar sob qual licença determinado objeto digital agregado se encontra.

A etapa de enriquecimento e normalização (DPLA, 2023f) tem como objetivo padronizar o texto dos valores de alguns metadados, tais como a remoção de caracteres indesejados, remoção de termos duplicados, relacionamento dos dados com outras bases de dados para garantir uma padronização entre fontes, como a padronização de nomes dos provedores com os nomes utilizados na Wikimedia, ou ainda o relacionamento de termos dispersos no campo de tipo para a listagem de termos padronizada indicada pelo modelo do Dublin Core.

A etapa de conexão com a Wikimedia (DPLA, 2023g) é responsável por publicar os metadados e a mídia do objeto digital agregado na plataforma Wikimedia. A partir da identificação de elegibilidade do objeto para essa etapa, que envolve principalmente o uso de licenças de direitos autorais abertas, o objeto que é publicado na Wikimedia pode ser enriquecido com contribuições de toda a comunidade Wiki, e assim democratizar ainda mais o acesso e a apropriação do objeto cultural pela sociedade.

Os demais processos tecnológicos expostos no repositório de códigos da DPLA indicam como os dados coletados no processo de ingestão são geridos e disponibilizados. A frente que disponibiliza os dados para o portal é desenvolvida em uma aplicação utilizando a tecnologia React (DPLA, 2023h); para o consumo automatizado, os dados do repositório agregado também estão disponíveis para consulta através de uma API (DPLA, 2023i). Já para controle interno do processo de coleta dos objetos dos provedores, a DPLA utiliza um dashboard analítico que apresenta quantitativos e a necessidade de re-ingestão de dados (DPLA, 2023j).

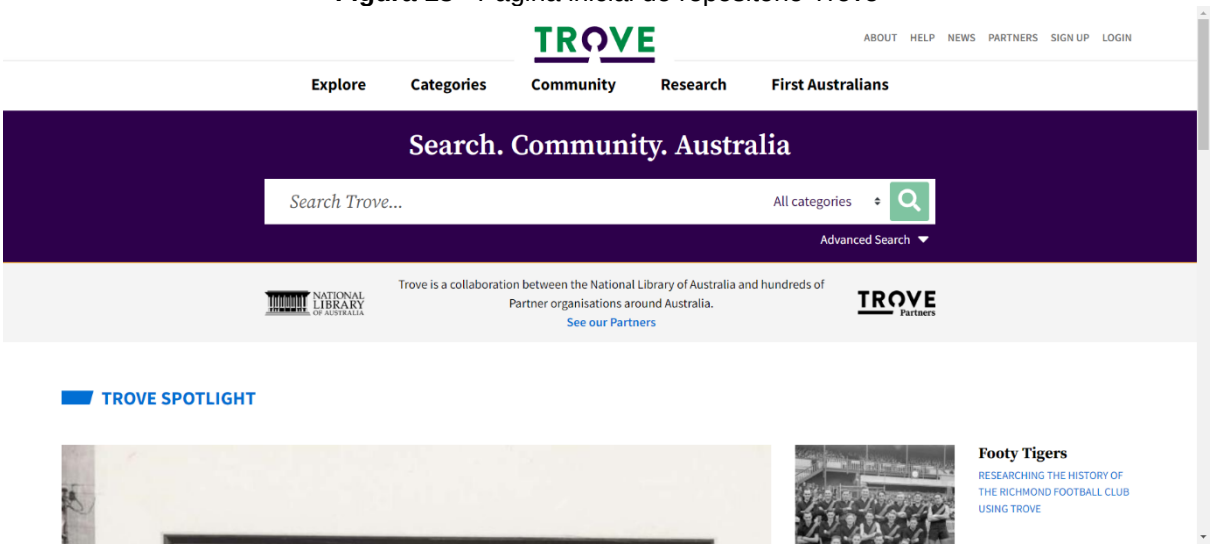
Em síntese, quando se trata de comunicação entre o repositório de agregação da DPLA e o repositório dos Hubs, contendo os acervos digitais do patrimônio cultural, e na disponibilização de um portal web e API para consulta e recuperação dos dados, o processo de agregação envolve etapas estritamente tecnológicas. Mas, assim como observado no contexto da Europeana, também há um processo de organização do conhecimento e representação da informação, que se molda como um dos pilares para a manutenção deste tipo de serviço, que é parte da etapa de mapeamento e validação. Tal etapa, além de garantir uma correspondência entre modelos semânticos diferentes, tanto no quesito de elementos dos padrões de metadados, quanto no uso dos sistemas de organização do conhecimento, também se preocupa com a normalização através do relacionamento com outras bases de conhecimento,

promovendo um ambiente propício para o contexto dos dados abertos ligados (Bizer; Heath; Berners-Lee, 2009; Machado; Souza; Simões, 2019).

3.3 Portal Trove

O portal Trove da Biblioteca Nacional da Austrália reúne mais de 6 bilhões de objetos digitais de centenas de parceiros, tais como veículos midiáticos, bibliotecas, museus, galerias, governo e organizações comunitárias da Austrália (Holley, 2010).

Figura 13 - Página inicial do repositório Trove



Fonte: Repositório Trove, 2023.

3.3.1 Síntese sobre a iniciativa do portal Trove

O processo de agregação executado pelo Trove parte de uma etapa de coleta mais abrangente, permitindo coletar acervos digitais através de *Sitemaps*, OAI-PMH e APIs. Para a coleta dos dados dos acervos, é recomendada a adequação a um conjunto de metadados indicados, baseado também no Dublin Core. Alguns softwares ainda são utilizados para garantir a padronização do repositório, como o ArchivesSpace e o Preservica.

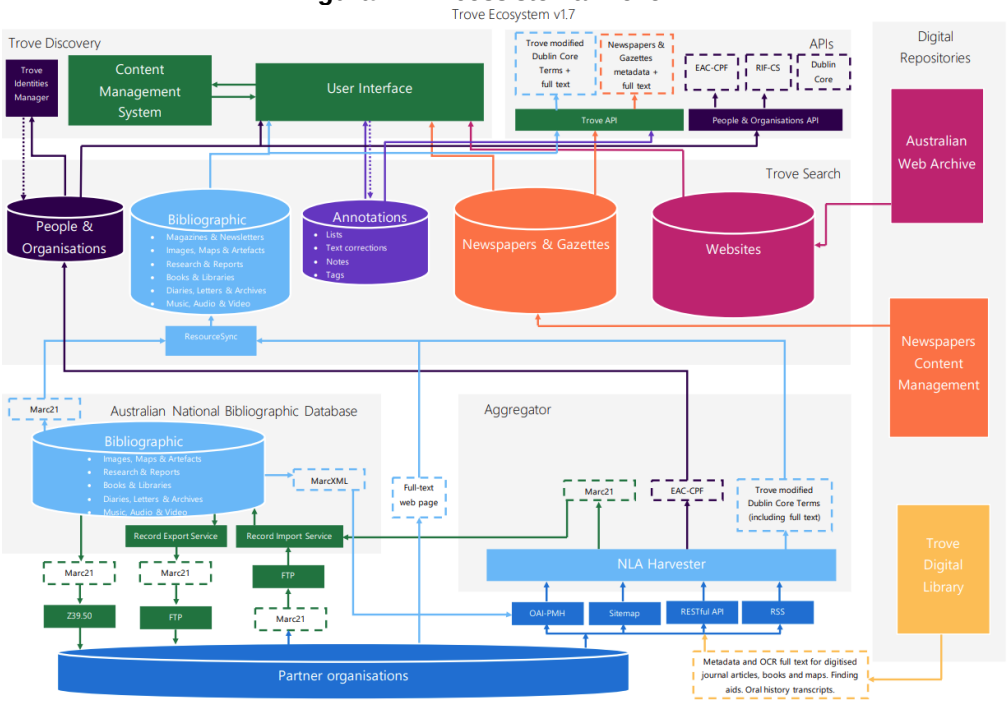
Apesar do acesso aos dados dos objetos digitais dos acervos agregados ser possibilitado por um único repositório de acesso, e API, o Trove é constituído por três plataformas diferentes; uma bibliográfica, uma arquivística, e outra para jornais de notícia, que permitem a gestão dos objetos digitais de maneira específica para cada caso.

3.3.1 Aspectos técnicos do portal Trove

O conjunto de elementos tecnológicos e de organização do conhecimento utilizados na agregação dos acervos do patrimônio cultural australiano é chamado de ecossistema técnico (TROVE, 2023). Esse ecossistema é composto por sete peças que trabalham juntas para a agregação, a saber: o serviço de descoberta, repositórios digitais, o agregador, uma base de dados cooperativa de catalogação, um diretório de instituições com acervos digitais, um suporte de serviço de busca, e uma API.

Como apresenta a Figura 14, a coleta dos objetos digitais das instituições parceiras é realizada em duas instâncias, uma específica para dados provenientes da Biblioteca Nacional da Austrália, que publica seus metadados no modelo MARC21, e outra forma de coleta de dados mais variada, que utiliza o protocolo OAI-PMH e arquivos no formato XML, bem como o *Sitemap*, texto completo de páginas da web, APIs RESTful e feed RSS.

Figura 14 - Ecossistema Trove v1.7



Fonte: Trove, 2020.

Os dados coletados passam por duas etapas de processamento. No caso da base de dados da biblioteca nacional, os protocolos de comunicação FTP e Z39.50 são utilizados para importar os dados e mídias. Já para os demais tipos de objetos

digitais, uma etapa de agregação é responsável por padronizar e mapear para os modelos de metadados utilizados no repositório, tais como Marc21, EAC-CPF e o perfil de aplicação do Dublin Core.

Além da coleta de dados dos repositórios de parceiros, o portal Trove utiliza três repositórios digitais para agregar objetos das instituições do patrimônio cultural da Austrália, e esses repositórios são disponibilizados para a comunidade australiana com o apoio de bibliotecários e arquivistas. São eles, a Biblioteca Digital Trove, que indexa objetos digitais como livros, periódicos, artigos, fotografias, mapas, histórias orais e diários; o Gerenciador de Conteúdo de Jornais, que indexa jornais australianos digitalizados; e o Arquivo Web Australiano, que é um repositório histórico de páginas da web australiana.

Todos esses dados e mídias sobre os objetos coletados dos parceiros e dos repositórios digitais do Trove são disponibilizados em uma camada de busca, que organiza os dados em cinco bases: a base de autoridades, que organiza os dados sobre pessoas e organização; a base bibliográfica, que organiza os dados relativos aos objetos digitais bibliográficos agregados; a base de anotações, que organiza dados sobre listas, correção de textos, notas e *tags*; a base de jornais e revistas, que organiza dados sobre objetos digitais desse contexto; e a base de web sites, que organiza os dados provenientes do repositório de Arquivo Web Australiano.

O consumo desses dados organizados em bases diferentes pode ser realizado de duas maneiras, sendo a primeira é pelo contexto de Descoberta Trove, que reflete os componentes do portal web, como o sistema gerenciador de conteúdo, a interface de usuário e o gerenciador de identidades Trove. O outro modo de consumo é pelas APIs disponíveis, a Trove API, que disponibiliza os metadados no modelo de aplicação do Dublin Core da Trove, e o conteúdo dos Jornais e Revistas, ambas com texto completo, e a outra opção é a API de Pessoas e Organizações, que permite coletar dados sobre autoridades nos formatos EAC-CPF, RIF-CS e Dublin Core.

Em síntese, o portal Trove tem características gerais semelhantes ao já apresentado no caso da Europeana e da DPLA, uma etapa de coleta/ingestão de dados, que coleta dados de Hubs, Agregadores de vários países e/ou parceiros, uma etapa de normalização e mapeamento, uma etapa de organização e gerenciamento dos objetos digitais armazenados, e uma etapa de publicação desses objetos digitais através do portal web e de APIs.

Uma característica diferente dos demais é a disponibilização de três repositórios digitais anexos aos agregados que permitem a indexação de objetos digitais relacionados ao patrimônio cultural digital do país, com a participação da comunidade e curadoria de profissionais. Esses repositórios já estão configurados no modelo semântico utilizado pelo agregador, e têm o potencial de democratizar o acesso e uso dos objetos digitais pela comunidade australiana.

3.4 DigitalNZ

Mais um exemplo de agregador de acervos digitais culturais, é o DigitalNZ, que agrega mais de 30 milhões de objetos digitais, provenientes de mais de 200 instituições da Nova Zelândia, dentre bibliotecas, museus, galerias, departamentos governamentais, veículos midiáticos, e grupos comunitários (Rollit, 2009).

Figura 15 - Página inicial do repositório DigitalNZ

Fonte: DigitalNZ, 2023.

3.4.1 Síntese sobre o DigitalNZ

Se comparado aos demais agregadores apresentados acima, o processo de agregação de acervos digitais pelo DigitalNZ é mais flexibilizado. Desde a submissão manual, até a agregação através de *Sitemaps*, APIs, feed RSS, ou OAI-PMH, são métodos de coleta de dados contemplados. Também não há indicativo de padronização dos metadados coletados, uma vez que a própria instituição menciona,

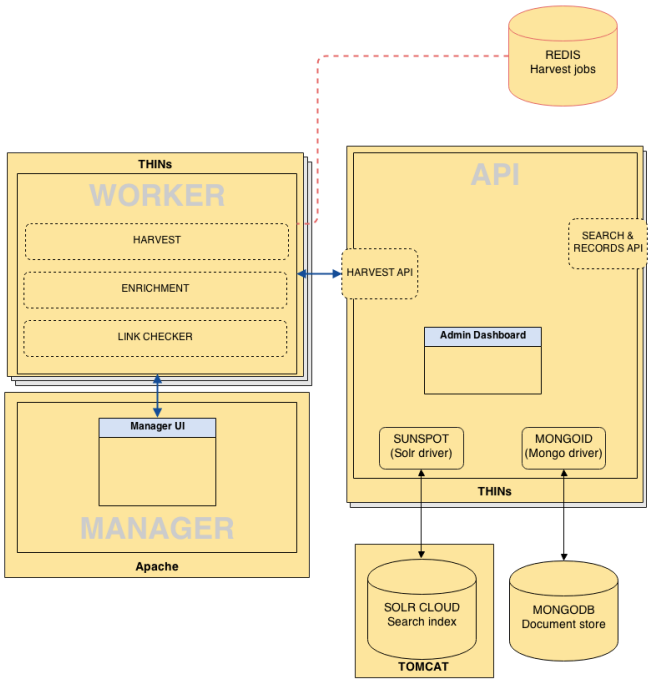
na área de API de acesso aos dados, que a qualidade e disponibilidade dos metadados varia consideravelmente.

3.4.2 Aspectos técnicos do DigitalNZ

A arquitetura técnica do portal DigitalNZ atualmente é fundamentada pelo desenvolvimento do software denominado Supplejack, que foi desenvolvido na linguagem de programação Ruby on Rails pela equipe DigitalNZ da Biblioteca Nacional da Nova Zelândia e pelo Departamento de Assuntos Internos.

O Supplejack tem em seu núcleo de arquitetura alguns componentes denominados *Worker* (trabalhador em português), usados para gerenciar os processos de coleta, enriquecimento e checagem de links dos objetos digitais agregados (Figura 16), são eles: *Manager* (Gerenciador em português), que é uma interface gráfica para usuários internos gerenciarem os processos e o conteúdo agregado, e *API* (Interface de programação de aplicações), que é composta por um conjunto de APIs para coleta, armazenamento, indexação e acesso à busca e registros do repositório agregado. Vale ressaltar que para garantir a escalabilidade o armazenamento utiliza o MongoDB, que é um banco de dados NoSQL, e usa também a solução Solr para indexação, ambas as tecnologias são características de um contexto de *big data*, em que a quantidade de dados é muito grande (Supplejack, 2023).

Figura 16 - Arquitetura técnica Supplejack



Fonte: Supplejack, 2023.

Além deste núcleo tecnológico, o Supplejack fornece uma interface web através de um sistema de gerenciamento de conteúdo para publicação e acesso ao repositório agregado pelos usuários da web. Juntamente com o acesso pelo site, ainda é possível configurar o acesso da API através de uma biblioteca denominada Supplejack Client, que permite a conexão e interação com as funcionalidades da API de descoberta do DigitalNZ.

Outro ponto importante desse serviço de agregação é o mapeamento de metadados, no qual se utiliza o Dublin Core como modelo comum e os metadados dos provedores coletados são mapeados para esse padrão. Como a maioria das coletas é feita por meio do protocolo OAI-PMH, esse processo utilizando o Dublin Core é facilitado, uma vez que o protocolo já implementou o modelo em sua estrutura (Rollit, 2009).

Em síntese, a característica marcante da iniciativa DigitalNZ é o desenvolvimento de um software à parte para o processo de agregação. A disponibilização do Supplejack, juntamente com sua documentação, abre espaço para sua apropriação por outras instituições com o interesse em agregar acervos digitais na web. Os procedimentos em si são semelhantes aos demais já apresentados. Existem processos de coleta, mapeamento, validação e enriquecimento que são auxiliados por APIs, e os dados coletados são armazenados em bancos de dados, e

indexados para disponibilização através de um portal web, e de uma API de consulta e descoberta. Vale chamar a atenção pelo uso dos sistemas Solr e MongoDB, que almejam lidar com a grande quantidade de dados acumulados na agregação, de maneira que não onere a experiência de consulta do usuário final.

3.5 Hispaña - Acceso en Línea al Patrimonio Cultural

O Hispaña - Acceso en Línea al Patrimonio Cultural agrega mais de 10 milhões de objetos digitais provenientes de 233 repositórios digitais, e é um dos principais provedores de dados para a Europeana, evidenciando uma prática interessante de reúso de repositórios digitais agregados em outras iniciativas maiores de agregação.

Figura 17 - Página inicial do repositório Hispaña

Fonte: Hispaña, 2023.

3.5.1 Síntese sobre o Hispaña

Para que as instituições do patrimônio cultural da Espanha tenham seus acervos agregados pelo Hispaña, é necessário preencher um formulário de participação e ter o acervo publicado através de um repositório que permita coleta por meio do OAI-PMH, utilizando o padrão de metadados Dublin Core (Xavier; Hernandez, 2020) e, posteriormente, um termo de consenso para agregação pela Europeana também será necessário.

3.5.2 Aspectos técnicos do Hispaña

O caso da iniciativa Hispaña é um exemplo de serviço de agregação nacional que serve como provedor de dados para a Europeana. Como já mencionado, além de uma iniciativa tecnológica, a Europeana envolve uma articulação política para garantir que as instituições do patrimônio cultural europeu estejam alinhadas no propósito de disponibilizar seus dados para agregação, como é o caso do repositório espanhol Hispaña (Hispaña, 2023).

Toda a infraestrutura do Hispaña é fundamentada no protocolo OAI-PMH. Mesmo que o modelo almejado seja o EDM da Europeana, a comunicação entre sistemas através do OAI-PMH facilita a troca de dados via padrão de metadados Dublin Core e assim, utilizando mapeadores disponibilizados pela própria Europeana, é possível mapear o Dublin Core para o EDM (Xavier; Hernandez, 2020).

Os provedores de dados sobre acervos digitais espanhóis também disponibilizam seus dados através do protocolo OAI-PMH, e o Hispaña utiliza o software de repositório digital DIGIHUB para agregar todo o conteúdo em um único lugar (Xavier; Hernandez, 2020) e, nesse caso, o processo fica mais simples, uma vez que os dados já estão em um mesmo padrão, seja ele o Dublin Core ou EDM. Esse serviço de agregação faz o papel de coletar, armazenar e disponibilizar, sem um aprofundamento na etapa de mapeamento e enriquecimento, que será realizado pela Europeana, e então disponibilizado de volta para consumo do Hispaña.

3.6 Mexicana – Repositório del Patrimonio Cultural de México

Como exemplo de agregador de acervos digitais culturais no contexto latino, a Mexicana – Repositório del Patrimonio Cultural de México, agrega mais de 44 mil objetos, provenientes de 17 instituições que estão vinculadas ao Ministério da Cultura do México, como museus, bibliotecas, arquivos, televisão, e estações de rádio. Os objetos digitais estão disponíveis através dos formatos de imagem, áudio, vídeo, texto e 3D.

Figura 18 - Página inicial do repositório Mexicana

Fonte: Mexicana, 2023.

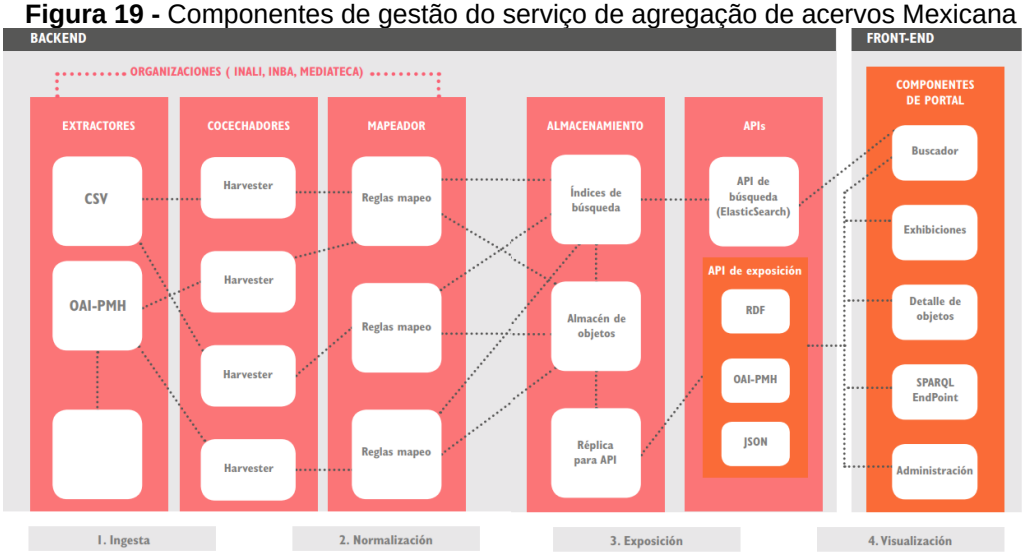
3.6.1 Síntese sobre a Mexicana

O processo de agregação dos acervos digitais executado pela Mexicana se dá pela coleta através de OAI-PMH ou RDF (Resource Description Framework) para os sistemas que permitem esse tipo de coleta, e coleta periódica dos registros para sistemas que não permitem protocolos de coleta automática. Após o processo, é realizado o mapeamento para um modelo de dados abrangente, baseado no CIDOC-CRM (CIDOC Conceptual Reference Model), visando à interoperabilidade de metadados entre os acervos (Salinas, 2023).

3.6.2 Aspectos técnicos da Mexicana

A arquitetura técnica do serviço de agregação da Mexicana pode ser entendida em duas camadas do sistema, o *back-end*, composto pela parte de ingestão/coleta, normalização e exposição do agregador, e o *front-end*, responsável pela visualização dos dados e mídias dos objetos digitais agregados (Figura 19).

Figura 19 - Componentes de gestão do serviço de agregação de acervos Mexicana



Fonte: Secretaria de Cultura, 2018.

Iniciando pelo *back-end*, dentro do processo de ingestão dos dados fornecidos pelos provedores, o primeiro componente a entrar em ação são os extratores, responsáveis por conectar com o repositório do acervo de origem e organizar os dados em um sistema local, preparando-os para as demais etapas. São previstas as existências de vários extratores, de acordo com o tipo de conexão com o provedor de dados, como CSV (tabulação em planilhas), ou OAI-PMH (Secretaria de Cultura, 2018).

O próximo conjunto de componentes são os coletadores, que atuam como gerenciadores intermediários entre os extratores e o mapeador (próximo elemento). Os coletadores configuram a demanda e necessidade de extração, bem como validam a correspondência com as regras previamente estabelecidas de coleta (Secretaria de Cultura, 2018).

Já o mapeador é o componente responsável pela organização do conhecimento dos dados obtidos na etapa de ingestão, e busca unificar como os objetos serão representados através de um único modelo semântico. A Mexicana utiliza como referência o Modelo de Dados do México, este modelo é baseado no modelo mais amplo para objetos culturais CIDOC-CRM. Assim, os diferentes modelos adotados pelas instituições do patrimônio cultural que são agregadas pela Mexicana são admitidos na etapa de ingestão, e na etapa de mapeamento é realizada uma correspondência entre os elementos dos modelos de origem para o modelo de dados do México (Secretaria de Cultura, 2018).

Já na etapa de armazenamento, ocorrem três processos: 1- a indexação utilizando a ferramenta ElasticSearch, para formação de um índice de busca que alimentará a API que gerencia o buscador da plataforma web; 2 - o armazenamento dos objetos em bancos de dados e de arquivos; 3 - a replicação desses bancos de dados para a API de exposição, que alimenta outros componentes do *front-end* (Secretaria de Cultura, 2018).

A última etapa de visualização é composta dos elementos que mediam a interação dos usuários pelo portal web: o buscador do repositório agregado, que é gerenciado pela API de busca; as exibições, detalhes dos objetos, administração do conteúdo, e o *endpoint* SPARQL (Secretaria de Cultura, 2018). Este último é responsável por oferecer um ponto de consulta relacional, que está dentro do contexto dos dados abertos ligados.

Em síntese, a Mexicana também segue um processo semelhante de agregação em 4 etapas, ingestão, normalização, exposição e visualização. Na etapa de coleta é interessante observar que o formato CSV foi citado na Figura 19 e esse formato não permite atualização automática. Dessa forma, nesse processo de agregação é necessária a atualização manual dessas planilhas. Outro ponto interessante é o uso do ElasticSearch, que é uma ferramenta de indexação de objetos digitais para busca e que melhora a eficiência do retorno das consultas, tornando o processo mais veloz e objetivo.

3.7 Brasiliana Museus

No âmbito brasileiro, a Brasiliana Museus é uma iniciativa de repositório de agregação dos acervos museológicos brasileiros atualmente vinculados ao Instituto Brasileiro de Museus. São 20 acervos digitais museológicos agregados, totalizando aproximadamente 17 mil itens com serviço de busca integrada disponível.

Figura 20 – Página inicial do repositório Brasiliana Museus

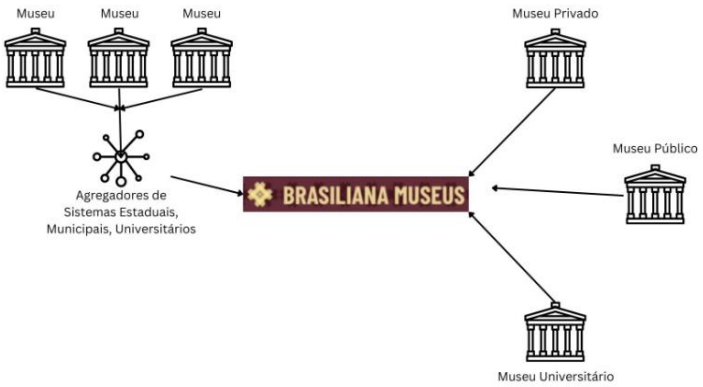
Fonte: Brasiliana Museus, 2023a.

3.7.1 Síntese sobre a Brasiliana Museus

Como esses acervos digitais vinculados ao Ibram estão publicados através da plataforma Tainacan, a API da plataforma foi utilizada para a comunicação de dados entre os acervos, e as tecnologias do conjunto Elastic Stack são a estrutura necessária para manter o serviço de agregação e busca integrada operante (Siqueira; Martins; Lemos, 2022; Siqueira; Martins, 2021).

Além dos museus vinculados ao Ibram, a proposta da Brasiliana Museus é também agregar outras instituições museológicas que não estão vinculadas ao instituto, sejam elas privadas, públicas ou universitárias (Figura 21), semelhante ao processo desenvolvido na Europeana com os serviços de agregação dos países europeus.

Figura 21 - Arquitetura de informação da rede Brasiliana Museus



Fonte: Brasiliana Museus, 2023b.

3.7.2 Aspectos técnicos da Brasiliana Museus

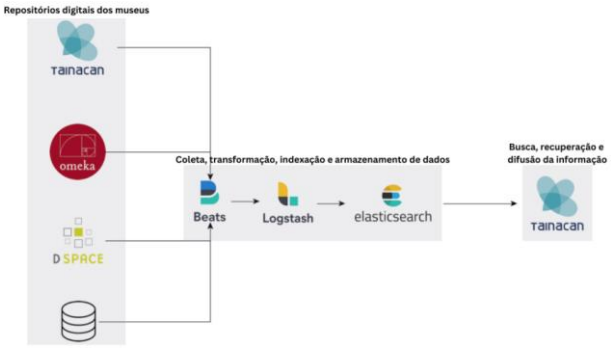
A Brasiliana Museus, é um serviço de agregação baseado na integração entre o Tainacan e um conjunto de softwares denominado Elastic Stack (Figura 22), que são utilizados em contextos com grande fluxo e volume de dados em tempo real (Siqueira; Martins, 2021). O Tainacan entra como o software de repositório que irá armazenar os dados e arquivos obtidos dos provedores, e através do WordPress como sistema de gerenciamento de conteúdo, é responsável pela interface de interação do usuário com o repositório de agregação (Martins; Lemos; Andrade, 2021).

Figura 22 - Arquitetura simplificada do Brasiliana Museus

Fonte: Siqueira; Martins; Medeiros, 2022.

Os acervos museológicos dos museus vinculados ao Ibram, por estarem publicados através do Tainacan, permitem o acesso automático aos dados e arquivos através de sua API de exposição do acervo. Todo o processo de coleta a partir da API, até o armazenamento na instância de agregação de acervos do Tainacan é realizado pelo conjunto de softwares Elastic Stack (Figura 22 e Figura 23). Os softwares desse conjunto são: o Elasticsearch, o mesmo utilizado pela Mexicana para indexação e gerenciamento do mecanismo de busca; o Beats, utilizado para conexão com os provedores e coleta dos dados; o Logstash utilizado para gerenciar e processar os dados coletados, aplicando filtros, e transformações (Siqueira; Martins; Lemos, 2022).

Figura 23 - Arquitetura de tecnologias da rede da Brasiliana Museus



Fonte: Brasiliana Museus, 2023b.

Mesmo que provenientes de museus vinculados a uma mesma instituição do patrimônio cultural, os museus Ibram não seguem necessariamente um modelo semântico de metadados, caracterizando variação no conjunto de campos utilizados para descrever o objeto digital. Assim como nas demais iniciativas apresentadas anteriormente, se fez necessário processo de mapeamento dos metadados para um único modelo de agregação.

No caso do Brasiliana Museus, não se tem um modelo semântico definido. Os metadados foram mapeados para um conjunto de campos definidos pelo Ibram no Inventário Nacional de Bens Culturais Musealizados (INBCM), que elenca 15 elementos descritivos e uma descrição semântica de cada um. Os metadados dos museus vinculados foram mapeados para este modelo em busca de promover a agregação dos acervos (Siqueira; Martins; Lemos, 2022).

Em síntese, a Brasiliana Museus integra dois conjuntos de tecnologias para fornecer um serviço de agregação de acervos, o primeiro conjunto é o Tainacan + WordPress, que estão situados tanto como sistema utilizado pelos provedores de dados, quanto sistema utilizado para armazenar e disponibilizar o acervo agregado, e o segundo é o conjunto de softwares da Elastic Stack, que é responsável por gerenciar o processo de coleta, normalização, mapeamento e mecanismo de busca. Como bem informado em seu portal web, a iniciativa ainda tem grande potencial para se tornar um serviço de agregação de outros tipos de provedores do patrimônio cultural.

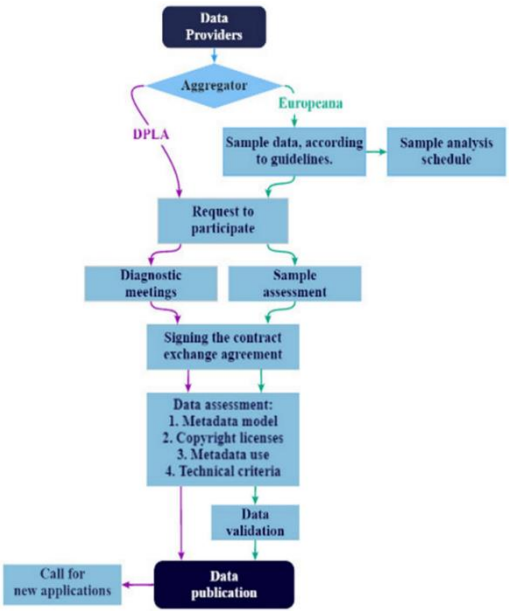
3.8 Outras revisões sobre aspectos técnicos de agregação de acervos digitais culturais

Além da revisão sobre as características da arquitetura tecnológica das iniciativas apresentadas nos últimos tópicos, alguns estudos no campo da ciência da informação também revisaram as características técnicas desse tipo de iniciativa a partir de outros recortes (Siqueira *et al*., 2021; Siqueira; Martins, 2020; Siqueira; Martins, 2022). Nesse sentido, em busca de complementar a revisão realizada na presente pesquisa, nos parágrafos abaixo serão apresentados resultados de demais estudos.

Assim como a Europeana e a DPLA foram iniciativas citadas neste trabalho como caso de referência em serviços de agregação de acervos digitais culturais, Siqueira *et al*. (2021) produziu uma pesquisa sobre os elementos necessários para a construção de uma política de qualidade de dados nesse contexto de agregação.

Mesmo que com outros objetivos, os autores realizaram o levantamento de um modelo dos estágios de agregação das duas iniciativas (Figura 24). Os estágios são referentes ao processo de avaliação de qualidade de agregação de dados, que estão mais relacionados às etapas gerais de coleta e mapeamento dos metadados das iniciativas que foram apresentadas.

Figura 24 - Estágios de Agregação de Dados, com foco na qualidade de Dados



Fonte: Siqueira *et al*., 2021.

Como observado acima, há um conjunto de etapas que descrevem o acordo e as atividades necessárias que um provedor de dados (instituição que possui o acervo digital cultural) deve seguir para fornecer os dados para agregação. É interessante observar que mesmo que exista a possibilidade de coleta automática, a instituição que fornece o acesso ao serviço de agregação tem que se comprometer a participar de todo o processo de modo a garantir a integridade de elementos, tais como o modelo de metadados, os direitos de uso dos objetos, o tipo de uso dos metadados e os critérios técnicos de fornecimento dos dados.

Em síntese, estas etapas ocorrem em paralelo com a implementação técnica de um novo acervo no agregador. Antes de realizar a coleta em si, é necessário passar por um processo de requisição para participação, que envolve diagnósticos e análises do acervo a ser agregado, a assinatura de contrato entre as partes, validação dessa integridade mencionada no parágrafo anterior, para então ser possível prosseguir com mapeamento, armazenamento e publicação no portal do serviço de agregação.

Em outra pesquisa, Siqueira e Martins (2022) fizeram o levantamento de fluxos de agregação de dados do patrimônio cultural. O estudo é uma revisão sistemática de literatura e identificou 12 modelos de fluxos de agregação para o contexto dos acervos digitais culturais.

Um dos primeiros pontos do estudo dos autores que complementa a presente pesquisa é o levantamento de passos para agregação de dados realizado a partir da análise dos fluxos identificados na revisão. Foram identificados 9 passos, a saber: 1 – Coleta; 2 – Ingestão; 3 – Mapeamento; 4 – Indexação; 5 – Armazenamento; 6 – Monitoramento; 7 – Enriquecimento; 8 Exibição; e 9 – Publicação LOD (Siqueira; Martins, 2022).

Ao fazer uma reflexão a partir dos elementos encontrados na presente pesquisa, das etapas identificadas, a coleta e a ingestão aparecem por vezes em uma mesma etapa, que executa a conexão com o provedor de dados e insere os dados no sistema de agregação. Outro ponto que não foi evidenciado nos estágios de agregação das iniciativas de diferentes nações foi a etapa de monitoramento, que pode estar implícita no processo de validação ou de gerenciamento de conteúdo.

Ainda no sentido de compreender os elementos de contribuição para esta pesquisa, o estudo de Siqueira e Martins (2022) apresenta a identificação de quais tecnologias são utilizadas nos modelos de fluxo de agregação. Os autores identificaram 28 tipos de tecnologias com diferentes licenças e distribuições aplicadas,

sendo que algumas delas vão ao encontro das identificadas nas iniciativas de diferentes nações aqui apresentadas, tais como o uso da linguagem SPARQL pela Mexicana, e sistemas que trabalham com bancos de dados não-relacional, como por exemplo o MongoDB, Solr, e Elastic Search, que foram observados na Europeana, Mexicana, Brasiliana Museus e DigitalNZ.

Entende-se que esses tipos de estudos complementam as características tecnológicas levantadas sobre as iniciativas de agregação de acervos do patrimônio cultural identificadas na presente pesquisa. O desenvolvimento destes estudos traz a contribuição de reunir informações técnicas e reflexões sobre como os repositórios e agregadores foram implementados e vêm se mantendo, além de viabilizar material científico de base e com boa qualidade para uso como referência e embasamento para possível reprodução das soluções.

3.9 Resumos sobre as características técnicas da agregação de acervos digitais culturais

A partir da descrição das iniciativas apresentadas anteriormente, é possível identificar alguns padrões na forma como o processo de agregação ocorre, que são algumas etapas que se assemelham, ou elementos técnicos utilizados em mais de uma ocasião. Esses padrões auxiliam na busca por identificar as condições informacionais e tecnológicas necessárias para a agregação de acervos digitais culturais.

Nesse sentido é possível indicar, de maneira geral, as características de três aspectos comuns entre essas iniciativas:

Quanto aos acervos agregados: 1 - a quantidade de objetos digitais agregados pode variar desde milhares a bilhões; 2 - os acervos digitais que são agregados são em maioria providos por instituições do tipo GLAM, mas também há ocorrência de veículos midiáticos (TROVE, 2023; Secretaria de Cultura, 2018); 3 - Os tipos de mídia dos objetos digitais podem variar, mas em suma são imagens, vídeos, áudios, texto, e em alguns casos, objetos 3D;

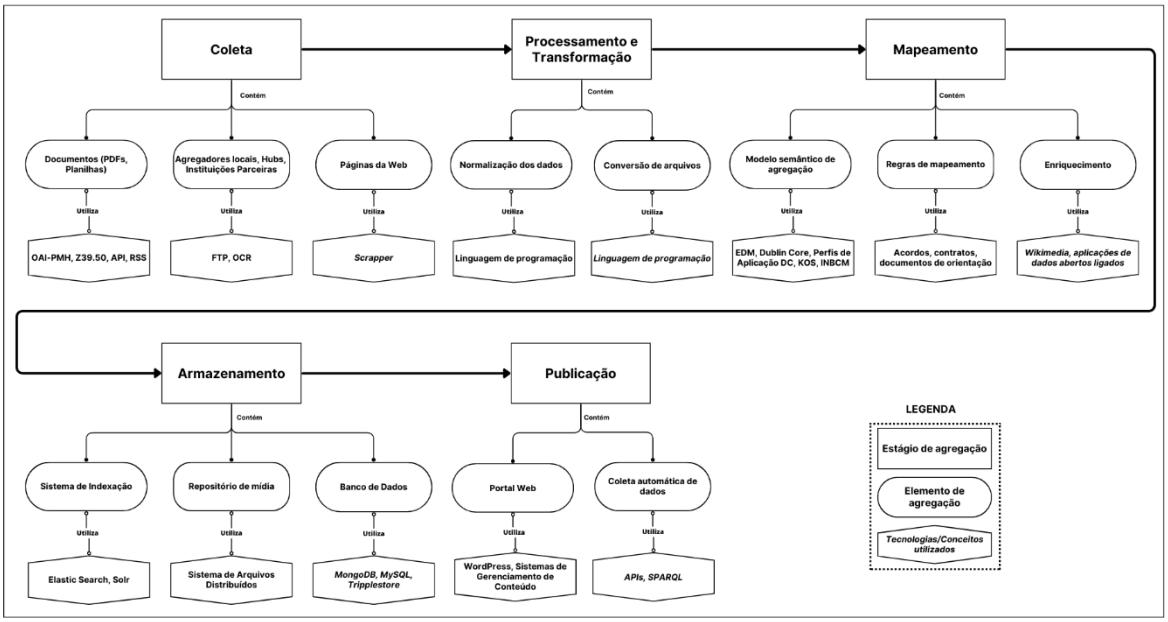
Quanto ao processo de coleta: 1 - todas as iniciativas evidenciam a importância da existência de protocolos de coleta de dados entre os repositórios dos acervos, como OAI-PMH, APIs ou XML, que são utilizados para manter a alimentação automática do agregador de acervos. Em alguns casos é indicada a coleta manual de

arquivos exportados de maneira periódica (Secretaria de Cultura, 2018); 2 - é mandatório um processo de mapeamento dos metadados utilizados nas fontes de dados (acervos coletados) para um modelo de metadados único, como é o caso do EDM da Europeana, para garantir a interoperabilidade entre os acervos através de um padrão de representação dos objetos digitais por meio do modelo único de metadados implementado;

Quanto à disponibilização dos acervos agregados: 1 - Todas as iniciativas apresentadas exploram formas de comunicação de partes do acervo agregado através do portal web, principalmente na criação de páginas expositivas de conteúdo temático, com referência a objetos digitais de diferentes acervos; 2 - A maioria das iniciativas permitem o reúso dos dados agregados através de APIs, que permitem tanto a coleta automática dos dados, quanto o desenvolvimento de aplicações que consumam os dados, como jogos, sistemas de análise de dados e etc. 3 - Em alguns casos há a possibilidade de consultar o acervo agregado através da linguagem de consulta SPARQL, que está dentro do contexto de dados abertos ligados (Secretaria de Cultura, 2018; EUROPEANA PRO, 2023;). No caso da DPLA, essa função pode ser explorada através do acervo integrado com a Wikimedia, que permite consultas SPARQL no universo de dados abertos ligados da Wikimedia Foundation.

Assim, em busca de uma forma de representar os resultados encontrados nessa revisão das características técnicas das iniciativas de agregação de acervos digitais culturais, é proposto um diagrama que apresenta a identificação de cinco estágios em comum de agregação, os elementos contidos nestes estágios, bem como as tecnologias utilizadas em cada elemento (Figura 25).

Figura 25 – Diagrama dos estágios em comum das iniciativas de agregação de acervos digitais culturais de diferentes nações.



Fonte: Elaboração própria, 2023.

O primeiro estágio, a coleta, é definido pelo processo de conexão com o provedor de dados através de uma tecnologia de comunicação entre sistemas ou coleta direta, e obtenção de dados e mídia sobre o objeto digital. Esse estágio é composto por três elementos identificados:

1. Agregadores locais, Hubs, ou Instituições Parceiras, que são provedores de dados com sistemas de repositório digital para seus acervos, que podem se configurar como um serviço de agregação ou não, e costumam permitir a coleta através de meios automáticos como OAI-PMH, API ou RSS;
2. Páginas da Web, que são uma forma de exibir um ou mais objetos digitais, e geralmente é coletada utilizando uma ferramenta de *scrapping* (raspagem de dados);
3. Documentos, que são as formas de se obter um ou mais objetos digitais através de arquivos, que podem ser transmitidos pelo protocolo FTP, e/ou são processados por OCR, para obtenção direta dos dados.

O segundo estágio, processamento e transformação de dados, é o momento em que ocorrem as adequações dos dados para garantir um primeiro nível de padronização, e é composto por dois elementos identificados:

1. Normalização dos dados, que envolve todos os procedimentos executados para garantir que os dados estejam em uma mesma estrutura, como é o caso das normalizações executadas pela DPLA;
2. Conversão de arquivos, que envolve a padronização do formato de mídias e documentos relacionados ao objeto digital.

O terceiro estágio, mapeamento, é caracterizado pela aplicação das práticas de organização do conhecimento e representação da informação, uma vez que o modelo semântico que organiza os acervos dos provedores deve ser mapeado para um modelo semântico de agregação. Esse estágio contém três elementos identificados:

1. Modelo semântico de agregação, que é caracterizado pelo sistema de organização do conhecimento que irá representar e organizar de forma agregada todos os objetos coletados. Esse modelo semântico pode ser ele uma ontologia, um perfil de aplicação de metadados, um conjunto de metadados, e/ou tesauros, vocabulários controlados, e taxonomias;
2. Regras do mapeamento, que são as diretrizes pelas quais o mapeamento se configura, e normalmente são definidas por meio de acordos, contratos, e/ou documentação de orientação entre os provedores e o serviço de agregação;
3. Enriquecimento, que é a etapa caracterizada pelo complemento dos dados obtidos dos provedores. Esse complemento é um processo de relacionamento entre os valores obtidos do modelo semântico do acervo com os valores presentes em outras fontes de referência, que podem ser o próprio repositório agregado, ou a Wikimedia, repositórios de dados de autoridade, georreferenciamento, entre outros.

O quarto estágio, o armazenamento, é caracterizado pelo processo de alocação dos dados e mídias coletados em bancos locais. Esse armazenamento pode ocorrer logo após o processo de coleta, mas neste diagrama ele é apresentado após o mapeamento, para promover o entendimento de que os dados armazenados já são os dados agregados. Esse estágio é composto por três elementos identificados:

1. Sistema de indexação, que é caracterizado pela organização dos objetos digitais armazenados de modo que sua recuperação seja facilitada. Nesse caso de agregação de vários acervos, como o volume de dados é maior, tecnologias de indexação como Elastic Search e Solr aumentam a velocidade dos mecanismos de busca dos objetos agregados;

2. Repositório de mídia, normalmente é gerenciado por um sistema de arquivos na hospedagem do serviço de agregação, mas também pode ocorrer através de sistemas distribuídos, que armazenam os arquivos em vários locais diferentes a fim de aumentar a performance;
3. Banco de dados, é o modo como os dados serão armazenados de acordo com sua estrutura. Se o modelo for relacional, normalmente se usa um sistema gerenciador, como por exemplo o MySQL. Não sendo relacional, é utilizado outro sistema para esse tipo de estrutura, como o MongoDB. Já em caso do armazenamento de dados em triplas, que geralmente são no formato RDF, necessita-se de um banco de triplas (*Tripplestore*).

O quinto estágio, publicação, é o processo de disponibilização do acesso dos acervos agregados ao público de interesse, seja o usuário comum da internet, através do portal web, seja uma aplicação ou usuário específico, através de uma forma de acesso e coleta automática. Esse estágio tem dois elementos identificados:

1. Portal web, caracterizado por disponibilizar ao público o conteúdo dos acervos agregados através de acesso direto aos objetos digitais, exposições digitais e ferramenta de busca. A tecnologia utilizada nesses portais envolve algum sistema de gerenciamento de conteúdo, como o WordPress que é um serviço terceiro, ou embutido na própria ferramenta de agregação;
2. Coleta automática de dados, que é uma opção de conexão com o serviço de agregação que permite a consulta e reutilização dos dados e mídia dos objetos digitais por outras aplicações ou pessoas. Normalmente são disponibilizadas APIs de descoberta e coleta de dados, e no caso de existência de dados em RDF, pode ser disponibilizado um ponto de consulta relacional em SPARQL.

Ainda, vale ressaltar que em alguns casos todo o gerenciamento de processos entre os estágios de agregação pode ser realizado através de APIs, como é o caso, mais explicitamente, da Europeana. Dessa forma, esse tipo de abordagem de desenvolvimento do sistema baseado em APIs não foi caracterizada pelo diagrama, pois se refere à forma como os estágios são gerenciados.

4 METODOLOGIA

Esta pesquisa busca responder como estão disponibilizados os acervos digitais publicados na internet pelas instituições vinculadas ao Ministério da Cultura do Brasil, quanto às suas características tecnológicas e de organização do conhecimento e representação da informação, diante da possibilidade de implementação de um serviço de agregação destes acervos.

Nesse sentido, de acordo com a questão de pesquisa proposta, a não exigência de controle sobre os eventos relacionados aos objetos em estudo e o foco nos acontecimentos contemporâneos, fazem com que a pesquisa se enquadre como estudo de caso (Yin, 2001). Ainda, é de caráter aplicado e descritivo, não participante, e utiliza abordagens qualitativas, como a técnica de análise categorial (Bardin, 1977), e quantitativas, como a estatística descritiva (Farias, 2020; Sampaio; Assumpção; Fonseca, 2018).

4.1 Estudo de casos múltiplos

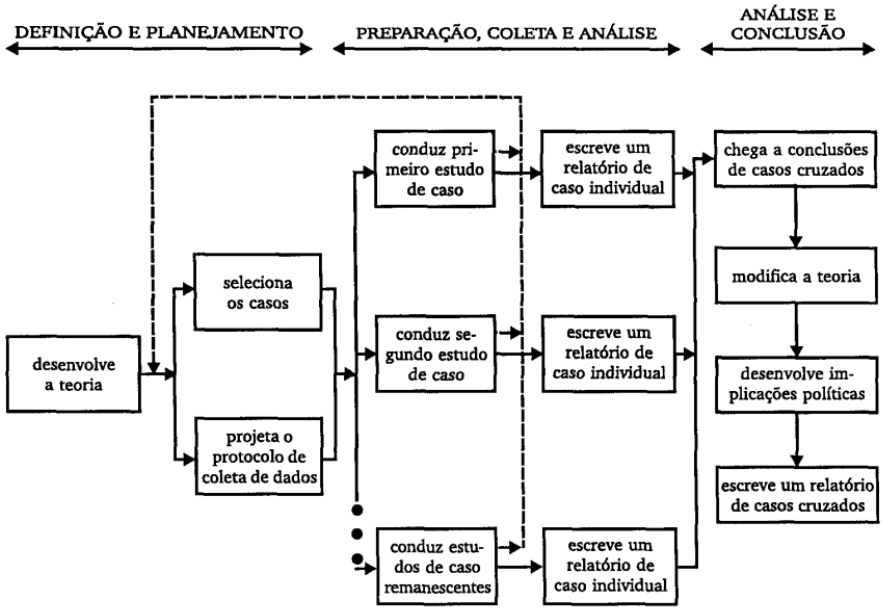
Segundo Yin (2001), existem quatro tipos de estudos de casos, a saber: estudos de casos únicos e estudos de casos múltiplos, que se subdividem em outros dois tipos: o holístico, quando o projeto é de natureza global e investiga o caso ou os casos de maneira geral; e o integrado, quando ocorre a existência de subunidades de análise nos estudos de casos únicos ou múltiplos.

Nas circunstâncias em que a pesquisa realizada envolve mais de um caso único, como por exemplo os casos das sete entidades vinculadas ao Ministério da Cultura do Brasil, o método de pesquisa se configura como estudo de casos múltiplos (Yin, 2001). Dessa forma, no contexto da presente pesquisa, em que serão estudados os acervos digitais destas entidades culturais, onde é possível classificar cada entidade como um estudo de caso único, e os acervos digitais como subunidades de análise, entende-se, então, a aplicação da metodologia de estudo de casos múltiplos integrados.

4.2 Etapas do estudo de casos múltiplos

De forma geral, de acordo com Yin (2001), um estudo de caso segue um modelo metodológico com três fases de execução: 1 - a definição e o planejamento; 2 - a preparação, a coleta e a análise; 3 - a análise e a conclusão (Figura 26).

Figura 26 – Método de estudo de caso



Fonte: Yin, 2001, p. 73.

Para a fase de definição e planejamento, são definidas três etapas: 1 - o desenvolvimento da teoria, que para esta pesquisa envolveu a revisão sobre organização do conhecimento e representação da informação no contexto dos acervos digitais do patrimônio cultural (tópico 2.1), bem como o levantamento das características das iniciativas de agregação de diferentes nações (tópico 3); e reflexão sobre o mapeamento de metadados para agregação de acervos digitais culturais (tópico 2.2); 2 – a seleção dos casos, que foi definida pelo escopo da pesquisa, e diz respeito às sete entidades do patrimônio cultural vinculadas ao Ministério da Cultura do Brasil (Tabela 2); 3 – projeto do protocolo de coleta de dados, que será apresentado no próximo parágrafo.

Tabela 2 - Informações gerais sobre as entidades vinculadas ao MinC

Entidade vinculada	Data de criação	Proveniência	Propósito
Iphan	1937	Lei 378	Preservar o patrimônio cultural brasileiro e garantir sua longevidade e fruição pelas gerações presentes e futuras (IPHAN, 2014)
Ibram	2009	Lei 11.906	Promover e garantir a implementação de políticas públicas no setor museológico, contribuindo para a organização, gestão e desenvolvimento das instituições museológicas e seus acervos (BRASIL, 2009)
Ancine	2001	Medida provisória 2228-1	Regular o setor cinematográfico em benefício da sociedade brasileira e promover o desenvolvimento de uma indústria forte, competitiva e autossustentável (ANCINE, 2018)
FCRB	1928	-	Promover a cultura, a investigação, a aprendizagem, a divulgação e a valorização da vida e obra de Rui Barbosa (FCRB, 2018)
FCP	1988	Lei 7.668	Promover a igualdade racial e valorizar as manifestações de raízes africanas do Brasil, além de implementar políticas públicas que fomentem a participação da população afro-brasileira nos processos de desenvolvimento do país (FCP, 2016)
Funarte	1975	Lei 6.312	Promover, estimular e desenvolver atividades culturais e incentivar a pesquisa, a produção e a formação de artistas, a preservação da memória e a formação de públicos para as artes no Brasil (FUNARTE, 2010)
FBN	1808	Itens trazidos ao Rio de Janeiro por D. João VI de Portugal e sua corte real	Coletar, gerenciar e conservar o patrimônio documental do país em língua portuguesa e sobre o Brasil, além de garantir seu estudo e divulgação (FBN, 2018)

Fonte: Adaptado de Martins *et al*., 2022, p. 9.

4.2.1 Protocolo de coleta de dados

O protocolo de coleta de dados proposto por Yin (2001, p.89 a 90) é definido por 4 seções:

- Visão geral do projeto do estudo de caso: indicação dos objetivos, estrutura, leituras e patrocínios do projeto;
- Procedimentos de campo: descrição das fontes de informação, formas e regras de acesso;
- Questões do estudo de caso: relação das questões de pesquisa que devem ser respondidas, e uma forma de acompanhamento de como as fontes de informação coletadas responderão a essas questões;

- Guia para o relatório do estudo de caso: estrutura de interpretação dos resultados encontrados, como resumo, referencial bibliográfico, documentações etc.

4.2.1.1 Visão geral

Para o contexto deste estudo, como visão geral, temos primeiramente a circunstância de que esta pesquisa ocorreu de forma concomitante com o projeto FAPESP/UnB, denominado “Interoperabilidade entre os repositórios digitais do patrimônio cultural brasileiro: da Web Semântica e dados abertos ligados às ferramentas de busca e recuperação da informação”, o que demonstra o interesse de uma fundação de apoio à pesquisa em estudos como este.

Além disso, os objetivos deste estudo, conforme indicados nos tópicos 1.3 e 1.4, são correlacionados com o referencial teórico apresentada no tópico 2. O referencial sobre a organização do conhecimento e representação da informação, situado no contexto dos acervos digitais culturais, é o elemento que contextualiza este trabalho no campo da ciência da informação. Já o levantamento de iniciativas, realizados no tópico 3, permitiu a síntese das estratégias de agregação através do modelo apresentado no tópico 3.9, e destaca-se que este modelo de agregação servirá como referência para buscar respostas ao problema proposto nesta pesquisa (tópico 1.1). Ainda, uma das etapas mais importantes desse modelo, o mapeamento dos metadados, é relacionada com o referencial teórico indicado no tópico 2.1, que é também descrita com mais detalhes no tópico 2.2.

4.2.1.2 Procedimentos do campo

Como procedimentos do campo, as fontes de informação utilizadas nesta pesquisa se dão em três níveis consecutivos de acesso aos dados: 1 – a identificação dos acervos digitais e de suas características através dos portais web das entidades do escopo da pesquisa; 2 – Coleta dos dados dos acervos digitais identificados no nível anterior; 3 – Coleta dos metadados utilizados nos acervos digitais através dos dados coletados no nível anterior;

4.2.1.2.1 Identificação dos acervos digitais e de suas características através dos portais web das entidades do escopo da pesquisa

Cada um destes níveis de coleta indicados no tópico anterior exige regras específicas de acesso e processamento. O nível 1, no caso desta pesquisa, pode ser interpretado como alusão ao processo de pactuação entre as instituições do patrimônio cultural e os serviços de agregação da Europeana e da DPLA (Siqueira *et al.,* 2021)

Como no contexto deste estudo a busca pelos provedores de dados para agregação parte do processo de pesquisa, é necessária etapa de caracterização desses provedores, que no caso são as entidades vinculadas ao MinC.

Assim, através do site do Ministério da Cultura (MINC, 2023), foram identificados os links dos portais web das entidades vinculadas (Tabela 3).

Tabela 3 - Links para os portais web das entidades vinculadas ao Ministério da Cultura

Entidade Vinculada ao MinC	Link para o portal Web
Agência Nacional do Cinema	https://www.gov.br/ancine/pt-br
Instituto do Patrimônio Histórico e Artístico Nacional	https://www.gov.br/iphan/pt-br
Instituto Brasileiro de Museus	https://www.gov.br/museus/pt-br
Fundação Biblioteca Nacional	https://www.gov.br/bn/pt-br
Fundação Casa de Rui Barbosa	https://www.gov.br/casaruibarbosa/pt-br
Fundação Cultural Palmares	https://www.gov.br/palmares/pt-br
Fundação Nacional de Artes	https://www.gov.br/funarte/pt-br

Fonte: MinC, 2023.

Cada um destes links relacionados às entidades vinculadas ao MinC foi acessado e explorado. A exploração se deu com o objetivo de levantar insumos descritivos sobre acervos digitais para identificação da forma como esses acervos são disponibilizados em relação às suas características tecnológicas e de organização do conhecimento e representação da informação.

Tal exploração foi guiada pela metodologia de análise de conteúdo da Bardin (1977), que em síntese envolve três etapas de execução: 1 - a pré-análise, em que se organiza o material a ser explorado, bem como são executadas leituras complementares para contextualizá-lo; 2 - a exploração do material, que consiste basicamente na análise do material proposta com base na pré-análise, identificando

categorias e codificando o conteúdo de interesse; 3 - o tratamento dos resultados e interpretações, que envolve a análise de frequências das categorias identificadas, bem como a interpretação crítica desses resultados.

Dentro da análise de conteúdo foi aplicada a técnica de análise categorial, que consiste em desmembrar o conteúdo analisado em “unidades, em categorias segundo reagrupamentos analógicos” (Bardin, 1997, p.153). Essas categorias (Tabela 4) foram definidas com base nas etapas da metodologia de análise de conteúdo, utilizando como referência o material levantado nos tópicos 2 e 3, bem como uma pré-análise dos portais web listados.

Tabela 4 - Categorias de análise da fase de caracterização dos acervos digitais das entidades vinculadas ao MinC

Categoria de análise	Descrição da categoria
Tipo de sistema de recuperação da informação	Especificação do sistema de recuperação da informação usado para acessar e manipular o objeto digital.
Software utilizado	Recurso tecnológico subjacente ao sistema de recuperação de informações.
Licenças de uso e Direitos autorais	Especificações do conjunto de licenças necessárias para publicação de objetos digitais.
Organização e representação da informação	Modelos utilizados para organizar e representar informações nas coleções digitais.
Visualização dos acervos	Formato usado para apresentar o conjunto de itens nos sites das instituições filiadas.
Extração de dados	Formas de extração de dados no sistema de recuperação de informação.
Tipo de mídia	Tipo de conteúdo de mídia (por exemplo, imagem, áudio, vídeo, texto).
Número de itens no acervo	Número total de itens identificados em cada site.

Fonte: Adaptado de Martins *et al*., 2022, p. 11.

A partir da exploração dos portais web das entidades vinculadas ao Minc, foram buscadas as tecnologias e conteúdos representativos para cada categoria de análise indicada na Tabela 4. Resultante dessa busca, o material encontrado foi utilizado para levantar as características dos acervos digitais presentes nos portais analisados.

4.2.1.2.2 Coleta dos dados dos acervos digitais identificados

O nível 2 de coleta, referente aos dados dos acervos digitais identificados no nível anterior, é análogo ao estágio de coleta do modelo de agregação apresentado no tópico 3.9. Uma vez definidos os provedores de dados e as características dos

acervos digitais a serem agregados, é o momento de acessar o sistema de recuperação da informação desses acervos e buscar meios de coletar os dados sobre os objetos digitais culturais armazenados nesses ambientes.

Devido à grande quantidade de potenciais acervos identificados na etapa de caracterização (tópico 4.2.1.2.1), e a não totalidade dos acervos com dados coletados até o momento de produção desta etapa da pesquisa, foi realizado um recorte para análise do processo de coleta, definido de acordo com a representatividade dos acervos, conforme os seguintes critérios:

- **Quantidade de objetos:** seleção dos links para os acervos encontrados que representam a maior parte dos objetos publicados pela instituição em seu site;
- **Forma de coleta de dados**: seleção dos links de acervos encontrados cujas formas de coleta de dados contemplem a maior variação possível, possibilitando que os testes aqui realizados permitam generalizar as características de coleta para os demais acervos identificados;
- **Ferramenta utilizada**: seleção dos links de acervos encontrados cujos *softwares* utilizados para publicação do acervo contemplem a maior variação possível, dado que ferramentas como SophiA, DSpace e Tainacan são, por exemplo, bastante difundidas entre as entidades analisadas, ao mesmo tempo em que o Flickr e o Fotoweb 7 aparecem em casos específicos. Assim, esse recorte buscou representar essa variabilidade.

Dessa forma, levando em conta a representatividade da estrutura geral apresentada pelos resultados do diagnóstico levantado na etapa anterior do projeto, e a partir dos critérios apresentados no parágrafo anterior, foi construído um recorte com 16 links para acervos das entidades vinculadas, como apresenta a Tabela 5 abaixo.

Como observação, ressalta-se que, no momento de execução desta pesquisa, não foi encontrado nenhum acervo digital no portal web da Agência Nacional de Cinema (ANCINE), por isso não há a presença de acervos dessa entidade no recorte realizado.

Tabela 5 - Amostra de links de acervos selecionados para coleta dos dados

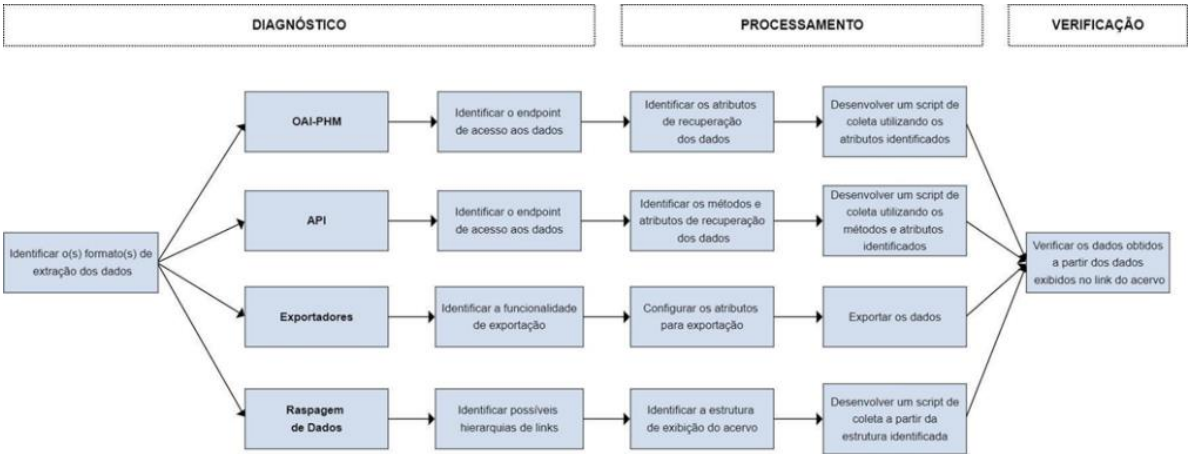
Entidade Vinculada	URL	Ferramenta	Extração dos Dados
FBN	http://bndigital.bn.gov.br/acervodigital/	SophiA	Raspagem de Dados
FBN	http://acervo.bn.gov.br/sophia_web/	SophiA	Raspagem de Dados
FUNARTE	http://sbrittod.funarte.gov.br/sophia_acervo/	SophiA	Raspagem de Dados
FUNARTE	http://cedoc.funarte.gov.br/sophia_web/	SophiA	Raspagem de Dados
FUNARTE	http://www.funarte.gov.br/colecoes-cedoc/	WordPress + Tainacan	API
FCRB	http://rubi.casaruibarbosa.gov.br/	DSpace	OAI-PHM
FCRB	http://iconografia.casaruibarbosa.gov.br/foto web/default.fwx	Fotoweb 7	Raspagem de Dados
FCRB	http://acervos.casaruibarbosa.gov.br/index.ht ml	SophiA	Raspagem de Dados
FCRB	https://www.flickr.com/photos/culturanegra/	Flickr	Raspagem de Dados
FCRB	http://www.palmares.gov.br/?page_id=50190	Wordpress	Raspagem de Dados
IBRAM	http://museudainconfidencia.acervos.museus .gov.br/	WordPress + Tainacan	API
IBRAM	http://museudearqueologiadeitaipu.museus.g ov.br/	WordPress + Tainacan	API
IPHAN	http://acervodigital.iphan.gov.br/xmlui/	DSpace	Raspagem de Dados
IPHAN	https://pergamum.iphan.gov.br/biblioteca/ind ex.php	Pergamum	Raspagem de Dados
IPHAN	http://portal.iphan.gov.br/videos	Página Estática HTML)	Raspagem de Dados
IPHAN	https://sicg.iphan.gov.br/sicg/pesquisarBem	SICG	Raspagem de Dados

Fonte: Elaboração própria, 2020.

A metodologia para coleta dos dados dos acervos identificados anteriormente envolve de maneira generalizada três etapas: o diagnóstico, o processamento e a verificação (Figura 27). Essas etapas foram fundamentadas nas formas de coleta

utilizadas pelas iniciativas de agregação de diferentes nações identificadas (tópico 3), e no processo de caracterização dos acervos digitais das entidades vinculadas (tópico 4.2.1.2.1).

Figura 27 - Níveis gerais de coleta dos dados



Fonte: Elaboração própria, 2020.

A etapa de **Diagnóstico** é composta pelas atividades de identificação tanto da forma de coleta de dados quanto dos pontos de coleta. A identificação da forma de coleta de dados foi executada na etapa de categorização dos acervos, através da categoria “Extração de dados”. Já a identificação dos pontos de coleta de dados foi executada de modo específico para cada tipo de coleta identificado. No caso da coleta através de OAI-PMH ou API, faz-se a identificação do(s) ponto(s) de acesso (*endpoints*); o caso da coleta através de exportadores, envolve a identificação da funcionalidade de exportação; e no caso da coleta através de raspagem de dados, dá-se a identificação das possíveis hierarquias de links.

Em relação à etapa de **processamento,** esta é composta pela identificação dos metadados, valores e objetos, bem como as funções e atributos disponíveis nos diferentes formatos de coleta dos dados. Em específico para a raspagem de dados, essa etapa contempla a identificação da estrutura de exibição do acervo, uma vez que essa técnica de coleta utiliza tal estrutura para acessar e obter os dados.

A etapa de **verificação**, é composta por processo de validação dos dados para cada saída. Para isso, 20 objetos coletados foram escolhidos de forma aleatória e comparados com os dados apresentados no acervo on-line, entendendo este ser um número adequado para identificar potenciais problemas técnicos, não sendo

necessária representação estatística do acervo, dado que o processo de coleta foi realizado utilizando amostra representativa dos acervos das entidades vinculadas.

As tecnologias aplicadas na coleta desses dados dos acervos digitais culturais serão apresentadas como resultado deste trabalho, tendo em vista que a pesquisa se propõe a desenvolver um protótipo de agregação desses acervos, e os elementos que compõem esse desenvolvimento são considerados como resultados.

4.2.1.2.3 Coleta dos metadados utilizados nos acervos digitais

Uma vez coletados os dados dos acervos digitais, eles foram armazenados no formato de tabelas para execução do estágio de mapeamento de metadados (Chan; Zeng, 2006a; Chan; Zeng, 2006b; Haslhofer; Klas, 2010). Os metadados são representados pelas colunas dessas tabelas, e a partir dessas colunas foram obtidos os metadados para prosseguir com o desenvolvimento do estágio de mapeamento.

Dessa forma, o processo de coleta dos metadados ocorreu a partir da seleção das colunas de cada uma das tabelas provenientes dos acervos digitais coletados para cada uma das entidades vinculadas ao MinC estudadas. Essas informações foram estruturadas de modo que fosse possível a análise da relação entre os metadados, seus acervos e as entidades a que pertencem, permitindo assim, sua visualização através de grafos (Bondy; Murty, 2008; Cherven, 2015).

4.2.1.3 Questões do estudo de casos múltiplos

As questões que delimitam a busca por respostas desta pesquisa estão implícitas, primeiramente, no problema de pesquisa: “Como estão disponibilizados os acervos digitais publicados na internet pelas instituições vinculadas ao Ministério da Cultura do Brasil, quanto às suas características tecnológicas e de organização do conhecimento e representação da informação, frente à possibilidade de implementação de um serviço de agregação destes acervos?”, e através dos objetivos específicos:

- Quais as diferentes formas de organização e representação da informação existentes nos acervos digitais das instituições vinculadas ao Ministério da Cultura do Brasil?

- Como coletar os dados dos acervos digitais das instituições vinculadas ao Ministério da Cultura do Brasil?
- Como padronizar os dados coletados dos acervos digitais a partir de um padrão de metadados?
- Como desenvolver um protótipo de agregação dos acervos digitais padronizados a partir de um único padrão de metadados?

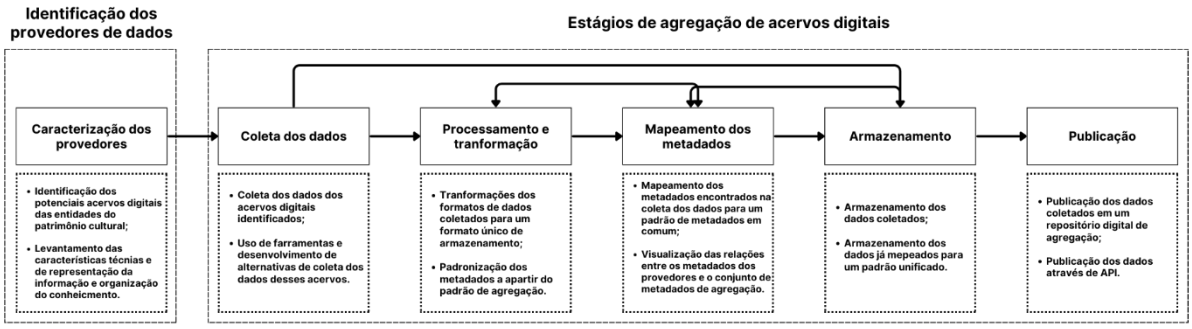
Dessa forma, reforça-se que a presente pesquisa busca caracterizar os acervos digitais do patrimônio cultural brasileiro a partir de suas características tecnológicas e de organização do conhecimento e representação da informação e, usando essa caracterização como insumo, propor um protótipo de agregação desses acervos digitais, em busca de descrever os eventos resultantes dessa experiência, de modo a produzir material para o embasamento de novas pesquisas que avancem na direção da implementação de um serviço de agregação digital da cultura nacional do Brasil.

4.2.1.4 Guia para o relatório do estudo de casos múltiplos

Assim, como guia para interpretação dos resultados encontrados neste estudo, será utilizado o referencial teórico apresentado no tópico 2, principalmente o referencial sobre a organização do conhecimento e representação da informação em acervos digitais do patrimônio cultural (tópico 2.1), e o levantamento das características técnicas das iniciativas de agregação de acervos digitais do patrimônio cultural de diferentes nações (tópico 3).

Em síntese, esta pesquisa percorre um conjunto de seis etapas (Figura 28) para o estudo de cada um dos casos referentes às entidades vinculadas ao MinC. A primeira etapa é a caracterização dos provedores, descrita pelo nível 1 de coleta no tópico (4.2.1.2). Essa etapa é essencial para entender o perfil tecnológico de disponibilização dos acervos digitais pelas entidades culturais pesquisadas. As demais etapas fazem referência ao modelo de estágios de agregação apresentado no tópico 3.9.

Figura 28 - Guia de etapas a serem percorridas no estudo de casos múltiplos



Fonte: Elaboração própria, 2023.

Este conjunto de etapas será realizado para cada caso estudado nesta pesquisa, iniciando pelo estágio de categorização dos provedores (tópico 4.2.1.2.1) a partir da aplicação da técnica de análise categorial (Bardin, 1977). As demais etapas, que englobam os estágios de agregação dos acervos, foram aplicadas conforme indicado na Figura 28 e seguem descritas nos tópicos abaixo:

- A etapa de **coleta dos dados** foi executada conforme indicado no tópico 4.2.1.2.2, com o acréscimo do uso da linguagem de programação Python para desenvolver os programas endereçados à forma de coleta identificada. O desenvolvimento desses programas de coleta é considerado como um resultado da desenvolução do protótipo de agregação, e será descrito no tópico 5.2.1.
- A etapa de **processamento e transformação** também se deu com o uso da linguagem de programação Python para o desenvolvimento dos programas de processamento e transformação. Parte das ações executadas nesta etapa está incluída no processo de coleta, e assim, será apresentada como resultado de pesquisa. As demais ações foram executadas na etapa de mapeamento dos metadados, no processo de automatização da unificação dos dados a partir de um padrão de metadados único, e será descrita juntamente com os resultados do mapeamento de metadados. Em busca de entender as limitações e os alcances desses programas desenvolvidos frente à possibilidade de implementação de um serviço de agregação, sua apresentação será interpretada a partir dos resultados do levantamento das iniciativas de agregação de acervos digitais de diferentes nações (tópico 3).
- A etapa de **mapeamento dos metadados**, envolveu a coleta dos metadados a partir dos dados coletados, como descrito no tópico 4.2.1.2.3, e a análise das

relações entre os metadados dos acervos coletados, utilizando a visualização de grafos (Bondy; Murty, 2008; Cherven, 2015, p.16). A partir dos metadados mais recorrentes utilizados, foi definido um padrão para a agregação e, a partir disso, um conjunto de pesquisadores vinculados ao Laboratório de Inteligência de Redes (UnB), executou o mapeamento realizado para o padrão de agregação. Esse mapeamento foi executado de forma sintática, os nomes dos metadados foram o principal fator de reconciliação, e os valores desses metadados somente foram consultados em caso de dúvida. Vale salientar que esta abordagem de mapeamento não foi baseada nas formas de mapeamento utilizadas pelas iniciativas de agregação de acervos de diferentes nações identificadas (tópico 3), uma vez que no caso dessas iniciativas, esse estágio geralmente se dá a partir de um conjunto de regras e escolhas da instituição que proverá o serviço de agregação.

- A etapa de **armazenamento** está presente em todo o processo de agregação, desde a forma de alocação dos dados coletados, até a forma como os dados foram publicados. Até a etapa de mapeamento, os dados foram armazenados utilizando planilhas no formato aberto CSV, já que são relacionadas com a forma com que os programas de coleta, processamento e mapeamento foram desenvolvidos. Na etapa de publicação, devido à escolha do sistema, os dados foram armazenados em duas instâncias: um banco de dados relacional para exibição dos dados no portal web, e um banco de dados NoSQL, para aumentar a eficiência da busca.
- Por fim, a etapa de **publicação** desses acervos agregados envolveu a escolha de um repositório digital e a sua disponibilização através de um portal web. O *software* de repositório digital escolhido foi o Tainacan, por três motivos: 1 – é um *software* difundido entre os museus brasileiros, e utilizado no caso da iniciativa nacional da Brasiliana Museus; 2 – é um *software* que fornece o meio de acesso ao repositório através de um portal web, já que foi desenvolvido na mesma estrutura do sistema de gerenciamento de conteúdo WordPress; 3 – é o *software* idealizado, desenvolvido e atualmente mantido pela equipe do Laboratório de Inteligência de Redes (UnB), da qual faço parte.

Ainda, menciona-se que que todos os elementos quantitativos destes resultados serão apresentados e interpretados com o auxílio da técnica de estatística descritiva (Farias, 2020; Sampaio; Assumpção; Fonseca, 2018), e os demais

elementos serão analisados conforme o material de referência indicado na revisão bibliográfica desta pesquisa (tópico 2.1).

4.2.2 Preparação, coleta, análise e conclusão

As duas fases restantes da metodologia de estudo de casos múltiplos de Yin (2001) são: Preparação, coleta e análise; e Análise e conclusão. No âmbito desta pesquisa, essas duas fases serão executadas em um mesmo contexto, visto que a aplicação do estudo de casos múltiplos em questão envolve a produção de um protótipo de agregação.

Essas fases serão representadas pela aplicação das etapas descritas na Figura 28 (tópico 4.2.1.4), em cada um dos casos estudados. Os resultados dessa aplicação serão apresentados conforme é indicado no modelo metodológico do estudo de casos múltiplos, primeiramente os resultados individuais de cada caso, e posteriormente os resultados cruzados entre os casos. Dessa forma, como produtos da pesquisa, serão desenvolvidos a caracterização dos acervos digitais das entidades vinculadas ao MinC, bem como o protótipo de agregação desses acervos. A interpretação dos elementos que compõem esses resultados será apresentada no tópico de conclusão (tópico 6).

5 RESULTADOS

Os resultados desta pesquisa serão apresentados de acordo com a proposta indicada na metodologia. Serão dois blocos, o primeiro referente aos processos de caracterização e coleta amostral dos metadados dos acervos: caracterização dos acervos digitais das entidades vinculadas ao MinC; coleta dos dados dos acervos digitais identificados nos portais web das entidades.

O segundo bloco de resultados já é referente ao desenvolvimento do protótipo de agregação dos acervos caracterizados no bloco de resultados anterior, e será composto pelo mapeamento dos metadados para um modelo de agregação, pela descrição das ferramentas tecnológicas utilizadas para desenvolvimento, e pela publicação do protótipo.

5.1 Dados coletados sobre os acervos digitais das entidades vinculadas ao MinC

Neste tópico serão apresentados os resultados referentes à execução das duas primeiras etapas da pesquisa, a caracterização dos acervos das entidades vinculadas ao MinC, e a coleta dos dados dos acervos identificados na etapa de caracterização.

Primeiramente serão apresentados os resultados gerais sobre os acervos digitais encontrados nos portais web das entidades vinculadas; posteriormente os resultados da análise categorial dos acervos identificados nos portais web das entidades vinculadas ao MinC; e, por fim, uma discussão analítica sobre os resultados encontrados.

5.1.1 Caracterização dos acervos digitais das entidades vinculadas ao MinC

No processo de pré-análise, através da exploração dos portais web das entidades vinculadas ao MinC, foram encontradas 217 formas de acesso aos acervos digitais nos sites oficiais das entidades vinculadas, que representam um total de 2.537.921 objetos. Vale destacar que para 28 formas de acesso não foi possível mensurar a quantidade de objetos, por falta de recurso técnico disponível para acesso a essa informação; portanto, o número de objetos disponíveis nos sites pode ser maior do que o mencionado acima.

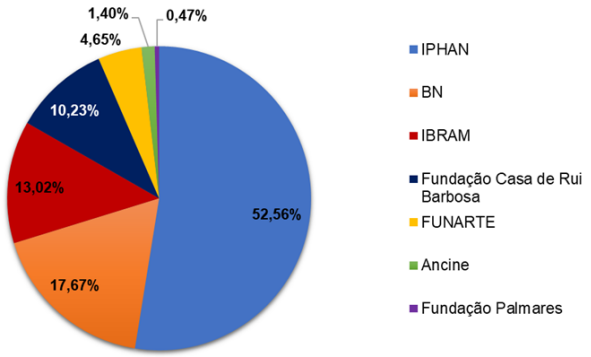
Essas formas de acesso ao acervo são os links encontrados que derivam para acervos digitais nos portais web das entidades. Importante reafirmar que a definição de acervos digitais que guia essa análise tem aspecto mais amplo (Martins *et al*., 2022), como indicado na introdução desta pesquisa.

Ao observar a proporção de formas de acesso por entidade, (Figura 29) o IPHAN demonstra maior evidência (52,56%) em relação às demais entidades com mais da metade das formas de acesso encontradas. Já a Biblioteca Nacional, o Ibram e a Fundação Casa de Rui Barbosa demonstram entre 10% e 18% das formas de acesso encontradas. Com menos expressividade Funarte, Ancine e Fundação Palmares apresentam uma proporção de formas de acesso menor que 5% dentre todas as vinculadas.

Esses resultados refletem a característica de variabilidade na distribuição desses acervos nos portais web das entidades, já que quanto maior a quantidade de

formas de acesso, maior o esforço do usuário para navegar no portal para acessar os acervos, e menor sua capacidade de visualizá-los de forma integrada. O mesmo fator dificulta processos de agregação dos conjuntos de objetos dos acervos, que deve então considerar a recuperação de informações de diferentes pontos de acesso, com padrões de extração de dados diferentes.

Figura 29 - Formas de acesso por entidade vinculada

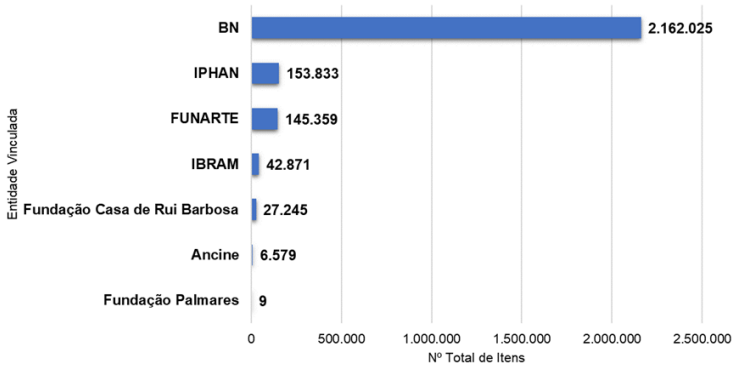


Fonte: Elaboração própria, 2020.

Com relação ao total de objetos informados por entidade (Figura 30), a Biblioteca Nacional apresenta maior evidência, com mais de dois milhões de objetos informados, quase 6 vezes mais objetos do que todas as outras vinculadas juntas. O IPHAN (153.833) e a FUNARTE (145.359) representam as duas entidades seguintes com mais de 100 mil objetos cada uma. Já o Ibram, a Fundação Casa de Rui Barbosa e a Ancine têm entre 6 mil e 43 mil objetos. E com a menor quantidade de objetos informados, a Fundação Palmares com 9 objetos.

É importante ressaltar que essa quantidade de objetos informados foi levantada através do quantitativo exposto pelos portais web, e pode variar à medida que existam formas de acesso não encontradas por essa pesquisa.

Figura 30 - Total de objetos informados por entidade vinculada



Fonte: Elaboração própria, 2020.

5.1.1.2 Análise categorial dos acervos identificados nos portais web das entidades vinculadas ao Minc

Os resultados apresentados a seguir são produtos da análise categorial (BARDIN, 1977), realizada a partir das categorias de análise descritas na metodologia (tópico 4.2.1.2.1). Cada categoria de análise resultou na descoberta dos elementos que compõem o contexto de cada categoria diante do conteúdo analisado através dos acervos identificados.

Esses elementos encontrados são descritos abaixo de acordo com as categorias de análise:

- **Tipo de sistema de recuperação da informação**: especificação do sistema de recuperação da informação usado para acessar e manipular o objeto digital.
    - Página estática (HTML): sistema em que os objetos do acervo são expostos através de página HTML simples, em que foi identificado que o objeto do acervo está integrado à página, tais como fotografias digitais exibidas na página web.
    - Repositório Digital: os objetos do acervo são expostos em um sistema específico para gestão, como DSpace, SophiA Web ou o WordPress+Tainacan, que indica o uso do CMS WordPress com a plataforma de publicação de acervos incorporada, por exemplo.
    - Sistema de Gerenciamento de Conteúdo: os objetos do acervo são expostos em páginas estruturadas a partir de um Content Management System (CMS), como WordPress ou Drupal, por exemplo.

- Documento: os objetos do acervo são expostos em estrutura de arquivos no formato de listas, como lista de objetos publicadas dentro de um arquivo em PDF, ou conjuntos de arquivos em ZIP, ou ainda expostos como pontos em um mapa, utilizando a ferramenta Google My Maps.

- **Software utilizado**: Recurso tecnológico subjacente ao sistema de recuperação de informações.
  - Sistema de Gerenciamento de Conteúdo: Drupal, Joomla e Wordpress.
  - Repositório Digital: DSpace, PHL, Pergamum, SophiA, Wordpress+Tainacan, OJS, OrtoDocs, Fotoweb 7, Google Arts and Culture e SICG.
  - Documento: Adobe PDF
  - Página Estática (HTML): Mecanismos de estruturação da exposição das formas de acesso ao acervo que faz uso de linguagem de marcação para publicação de informação em rede, como o uso de PHP e HTML para criação de conteúdo em páginas da *web*.
- **Licenças de uso e Direitos autorais:** especificações do conjunto de licenças necessárias para publicação de objetos digitais.
  - A partir dessa categoria de análise foram identificados os seguintes tipos de licenças utilizadas pelas entidades vinculadas para conceder acesso aos seus conjuntos de objetos: Copyright, Creative Commons, e CC BY-ND 3.0. Ainda ocorreram casos em que não foi identificada explicitamente a licença de acesso e uso dos objetos.
- **Extração de dados:** Formas de extração de dados no sistema de recuperação de informação.
  - A partir dessa categoria de análise foram identificados os seguintes meios de obtenção dos dados dos objetos: API, Harvester (OAI-PHM), CSV, KML, pedido por E-mail, Download de documentos (no caso de arquivos PDF e ZIP com a lista dos objetos) e Raspagem de Dados (processo automático ou semiautomático de se obter dados diretamente de sua página na Web).
- **Organização e representação da informação**: modelos utilizados para organizar e representar informações nas coleções digitais.
  - Padrões de metadados: foram identificados os seguintes padrões de metadados utilizados para catalogar os objetos dos acervos: Dublin

Core, Dublin Core+ (quando o Dublin Core é estendido com mais metadados), MARC. Existem casos em que o padrão de metadados não foi identificado, quando entende-se que existe um padrão de metadados implícito no conjunto de objetos, mas não foi reconhecido ou explicitado no site.

- Linguagens documentárias: foram identificados os seguintes tipos de linguagens documentárias utilizadas para auxiliar na padronização da descrição dos objetos: CDD, Base de Terminologia da FBN, Vocabulário controlado sobre escravidão, abolição e pós-abolição. Existem casos em que a linguagem documentária não foi identificada por falta de indicação no site analisado.
- Regras de catalogação: foi identificado o seguinte tipo de regra de catalogação utilizada para padronizar a forma de inserção de dados sobre os objetos nos acervos: AACR2. Existem casos em que a regra de catalogação não foi identificada por falta de indicação no site analisado.

- **Visualização dos acervos:** Formato usado para apresentar o conjunto de itens nos sites das instituições filiadas.
  - Foram identificados os seguintes tipos de formas de publicação dos acervos, que representa o modo como o acervo é apresentado no site das entidades vinculadas:
    - Coleções, em que os objetos estão distribuídos em coleções.
    - Exposição, em que os objetos estão organizados no formato de exposições (páginas HTML com parte do acervo, por exemplo).
    - Hierárquica, em que os objetos estão organizados de forma hierárquica (como em listas de registros ou vídeos, ou ainda conjuntos de pastas, por exemplo);
- **Número de itens no acervo:** Número total de itens identificados em cada site.
  - O número total de itens foi identificado a partir do conteúdo dos portais web das entidades, normalmente indicado pelos repositórios digitais dos acervos. Alguns conjuntos de objetos não tiveram tamanhos identificados, pela dificuldade de encontrar essa informação nos próprios sistemas de publicação, como o SophiA Web, que em determinados pontos de acesso não permite fazer busca por todo o

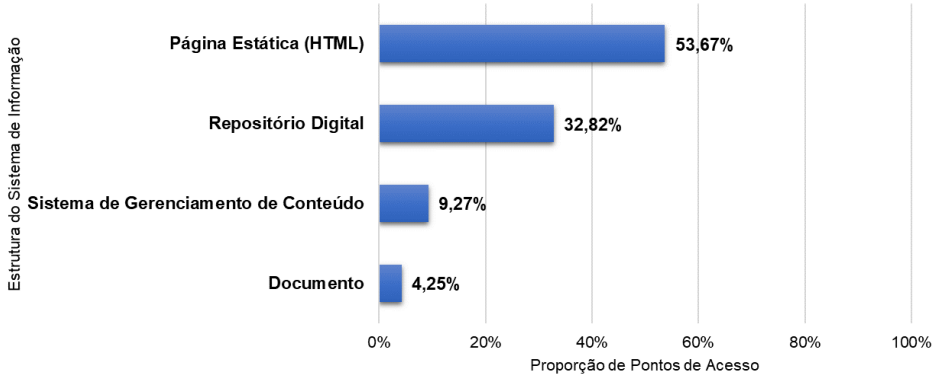
acervo, não informando assim em nenhuma página pública o total de objetos indexados pelo sistema. (o resultado para essa categoria de análise foi apresentado no início deste tópico)

- **Tipo de mídia:** Tipo de conteúdo de mídia (arquivos que continham metadados de objetos do acervo digital).
    - Foram identificados os seguintes elementos: E-book; PDF; PDF/A; Áudio; Imagem; Vídeo; e Texto, quando somente os registros foram identificados e o objeto cultural não foi digitalizado e disponibilizado on-line.

Para cada uma dessas categorias analíticas e elementos que as compõem, foi aplicada a técnica de estatística descritiva, mais especificamente o cálculo de frequência entre as formas de acesso identificadas, e os elementos das categorias.

Quanto ao tipo de sistema de recuperação da informação (Figura 31), é evidente a expressividade do uso de páginas estáticas em HTML e repositórios digitais para a disponibilização do objeto on-line. O uso de páginas HTML reflete um uso limitado de recursos tecnológicos de sistematização dos objetos, que poderiam ser armazenados em repositórios digitais com funcionalidades específicas de recuperação e reúso da informação.

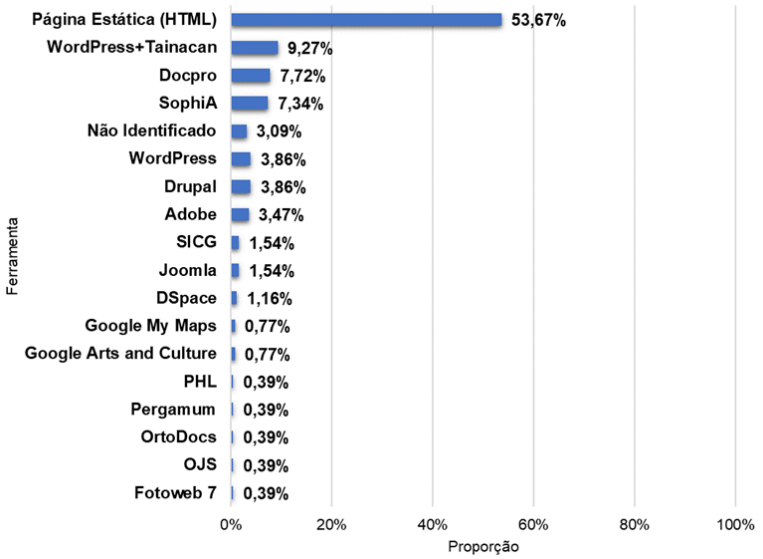
Figura 31 - Acervos por tipo de sistema de recuperação da informação



Fonte: Elaboração própria, 2020.

Quanto aos softwares utilizados para a disposição dos objetos nos acervos (Figura 32), a categoria página estática (HTML) é mais evidente dentre as outras tipologias, as demais se subdividiram em mais de um tipo de ferramenta/software. Esse resultado é esperado, pois quanto ao tipo de sistema de recuperação da informação, essa categoria também foi mais evidente.

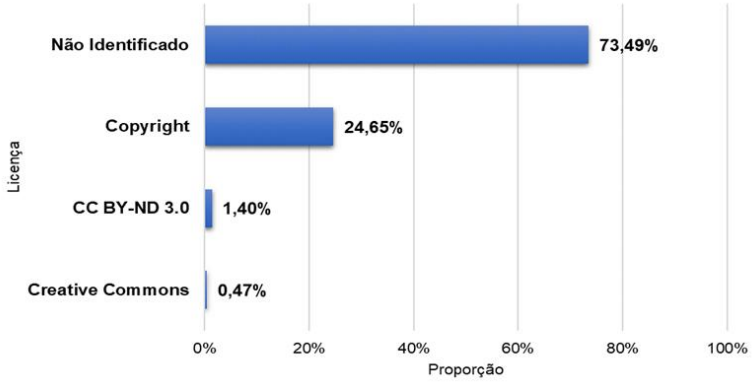
Figura 32 - Quantidade de acervos por ferramenta utilizada



Fonte: Elaboração própria, 2020.

Quanto às licenças de uso e direitos autorais (Figura 33), não foi possível identificar nas formas de acesso aos acervos as licenças vigentes para uso do conteúdo. Esse resultado dá abertura e abre o questionamento de como as entidades culturais explicitam a licenças de uso e os direitos autorais dos seus acervos através de seus portais web, já que nas interfaces de acesso aos objetos essa informação, em grande parte, não foi encontrada de maneira explícita na maioria dos casos (73,49%).

Figura 33 - Quantidade de acervos por licença utilizada

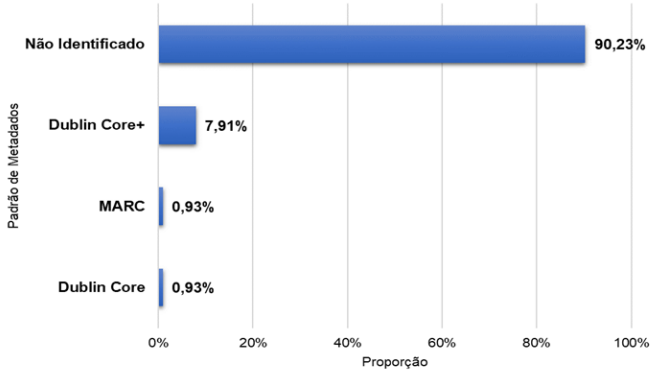


Fonte: Elaboração própria, 2020.

Quanto às formas de organização e representação da informação, mais especificamente quanto aos padrões de metadados utilizados nos acervos (Figura

34), na maioria dos casos (90,23%) o padrão de metadados não é explícito ou não foi possível identificar um padrão pela simples observação dos dados.

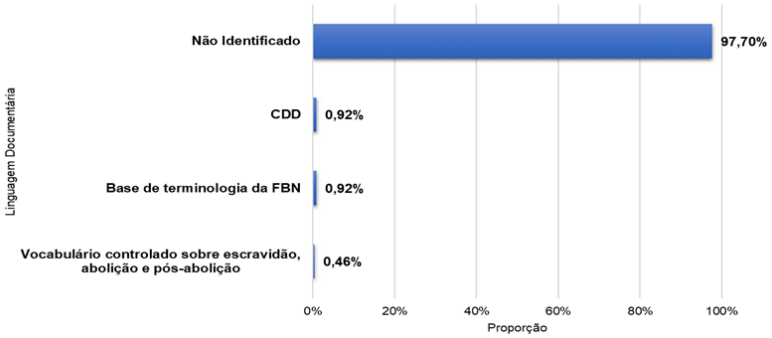
Figura 34 - Quantidade de acervos por padrão de metadados utilizado



Fonte: Elaboração própria, 2020.

Quanto às linguagens documentárias utilizadas nos acervos (Figura 35), os resultados seguem o mesmo contexto e apontam para a maioria (97,70%) dos acervos sem linguagem documentária explícita.

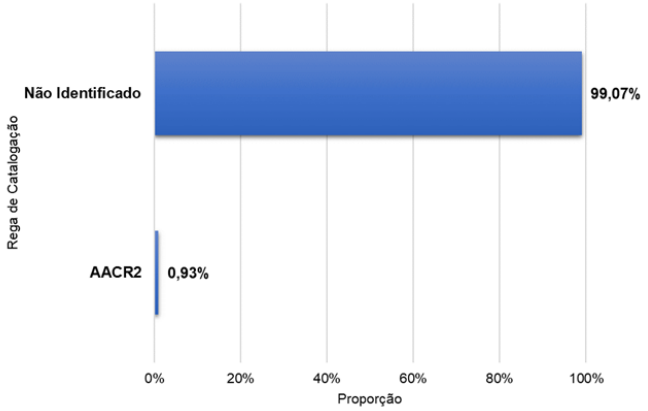
Figura 35 - Quantidade de acervos por linguagem documentária utilizada



Fonte: Elaboração própria, 2020.

Quanto às regras de catalogação (Figura 36), também compartilham da mesma característica, em 99,07% dos casos não foram identificadas quais as regras de catalogação aplicadas na indexação dos objetos dos acervos.

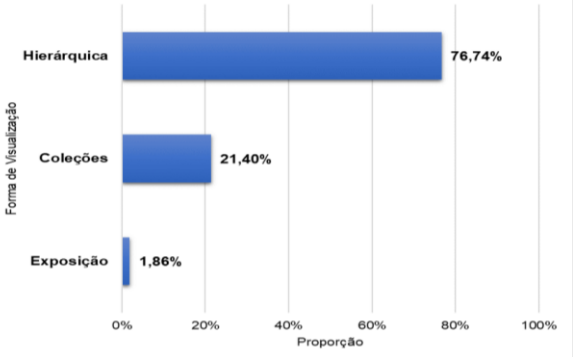
Figura 36 - Quantidade de acervos por regras de catalogação utilizadas



Fonte: Elaboração própria, 2020.

Quanto à visualização dos acervos (Figura 37), em sua maioria (76,74%) estão dispostos em formato hierárquico, forma mais simples de exibição e que não garante necessariamente interconexão entre os objetos.

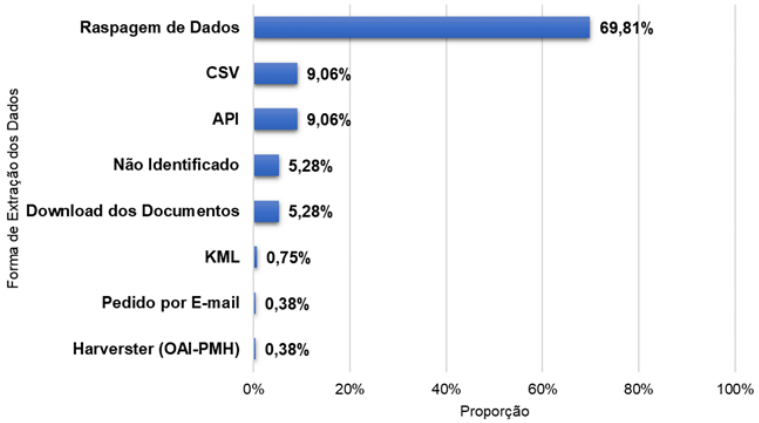
Figura 37 - Quantidade de acervos por forma de visualização do acervo



Fonte: Elaboração própria, 2020.

Quanto às formas de extração de dados (Figura 38), para a maior parte (69,81%) dos acervos é necessário aplicar técnicas de raspagem de dados, o que resultará na dificuldade de se obter os dados dos acervos para agregação, já que são técnicas de complexidade mais alta, comparadas com as exportações por CSV ou API. Tal resultado demonstra uma enorme fragilidade na constituição das estratégias tecnológicas dos acervos culturais brasileiros, tornando o desafio de construir ferramentas de agregação desses acervos mais difícil e complexa.

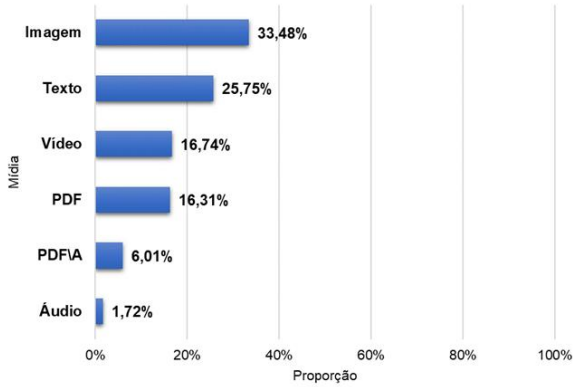
Figura 38 - Quantidade de acervos por formas de extração de dados



Fonte: Elaboração própria, 2020.

Quanto ao tipo de mídia, (Figura 39) houve proporção maior em objetos dos acervos no formato de imagem (33.48%) e no formato texto (25,75%) em que só os registros existem.

Figura 39 - Quantidade de pontos de acesso por mídias disponíveis no acervo



Fonte: Elaboração própria, 2020.

5.1.1.3 Discussão analítica sobre os resultados encontrados

De forma geral os resultados encontrados apontam alguns pontos importantes que ajudam a compreender as características das diferentes formas de organização e representação da informação existentes nos acervos digitais das instituições vinculadas ao Ministério da Cultura do Brasil:

- **Dificuldade de se encontrar documentação sobre os acervos das vinculadas**: Como apresentam os resultados analisados das categorias de

análise: Linguagens Documentárias, Regras de Catalogação, Padrão de Metadados e Licenças. Existe um potencial comprometimento da interpretação da organização do conhecimento e representação da informação desses acervos. Para agregação dos acervos é importante o esclarecimento da documentação dos pontos de acesso, para entender a correspondência conceitual dos metadados utilizados entre as entidades vinculadas, e execução do estágio de mapeamento, que garantirá a uniforme representação dos objetos em um serviço de agregação.

- **Licenças de uso e Direitos autorais:** Vale ainda reforçar a importância do papel das licenças no contexto de acervos culturais. Os objetos desses acervos culturais em sua maioria representam obras de autores que são artistas, figuras públicas, indígenas, entre outros. A autoria desses objetos do acervo digital por vezes leva ao desafio de cunhá-los como disponíveis ao público sem restrições, por isso é muito importante que as licenças e os direitos autorais estejam claros e sigam normas vigentes para esse fim. (Torino, Monteiro, Vidotti, 2023)

- **Grande quantidade de objetos sem condição de coleta automatizada:** poucos acervos permitem a coleta de dados através de aplicações como API ou um Harvester (OAI-PHM) ou ainda formatos abertos de dados estruturados como planilhas em CSV. Isso culmina na necessidade de coleta desses dados de maneira mais complexa, como raspagem de dados, que consiste em identificar como os dados estão estruturados através de suas páginas web e desenvolver programas personalizados para cada modo de disponibilização dos objetos.

Esses resultados encontrados corroboram a necessidade de se implementar estratégias de pactuação entre as instituições e os serviços de agregação para garantir a qualidade dos dados obtidos dos acervos dessas instituições (Siqueira *et al*., 2021; Siqueira, Martins, 2020; Siqueira, Martins, 2022). Vislumbrar uma estratégia de agregação de acervos digitais culturais brasileiros, após os resultados encontrados, se torna uma tarefa mais difícil e exige esforço específico na revisão da forma como o conhecimento é organizado e a informação é representada nesse contexto de acervos estudado.

Outro ponto importante identificado é referente às condições de reúso através da coleta de dados desses acervos. A falta de uma forma automatizada de coleta, como o OAI-PMH ou APIs, dificulta a prospecção de reúso em escala dos objetos desses acervos.

Com o avanço tecnológico expressivo das aplicações desenvolvidas atualmente, é essencial que os sistemas de recuperação da informação desses acervos disponibilizem formas automáticas de reúso de seus dados. Sem uma forma de coleta mais eficiente, como foi identificado na maioria dos casos estudados, esses acervos se apresentam limitados diante das possibilidades que a publicação na internet traz, dificultando o consumo dos dados no contexto da estratégia de agregação, por exemplo.

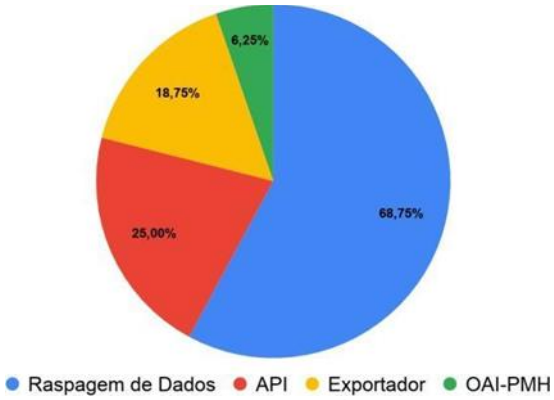
5.1.2 Coleta dos dados dos acervos identificados na etapa de caracterização

Os resultados apresentados neste tópico são referentes ao processo de coleta de dados dos acervos identificados na etapa de caracterização (tópico 5.2.1). As diretrizes de coleta estão descritas na metodologia desta pesquisa (tópico 4.2.1.2.2), e as ferramentas de coleta foram desenvolvidas utilizando linguagem de programação e estão descritas no tópico 5.2.1.

Vale recuperar a informação de que, devido ao contexto de produção desta pesquisa e a quantidade de acervos identificados nos portais web das entidades vinculadas ao MinC, foi realizado recorte de 16 acervos do conjunto total para executar a análise cujos resultados serão apresentados em seguida. Os critérios de seleção de acervos para esse recorte estão descritos no tópico 4.2.1.2.1.

Como resultado geral da coleta de dados, todos os 16 acervos tiveram de alguma forma seus dados coletados; àqueles em que não foi encontrada nenhuma forma de coleta através de funcionalidades como OAI-PMH, exportadores ou API, foi aplicada a técnica de coleta por raspagem de dados. Na Figura 40 abaixo, observa-se a proporção das formas de coleta utilizadas para coletar dados dos acervos.

Figura 40 - Proporção das formas de coleta dos dados dos acervos



Fonte: Elaboração própria, 2020.

Como apresentado (Figura 40), para a maior parte dos acervos (11 acervos) foi utilizada a coleta através de raspagem de dados, sugerindo que não foram identificadas outras formas de recuperação dos dados do acervo.

Para a forma de coleta de dados através de API foram identificados quatro acervos, sendo três deles acervos publicados utilizando o Tainacan, e um deles com o acervo publicado através do Flickr. É importante ressaltar que mesmo seguindo as orientações da documentação da API do software DSpace, não foi possível realizar a coleta de dados desse sistema, de modo que os únicos sistemas listados na seleção dos links para amostra que possibilitaram a coleta através de API foram o Tainacan e o Flickr. Aparentemente, a API do DSpace não estava habilitada para coleta na instalação analisada.

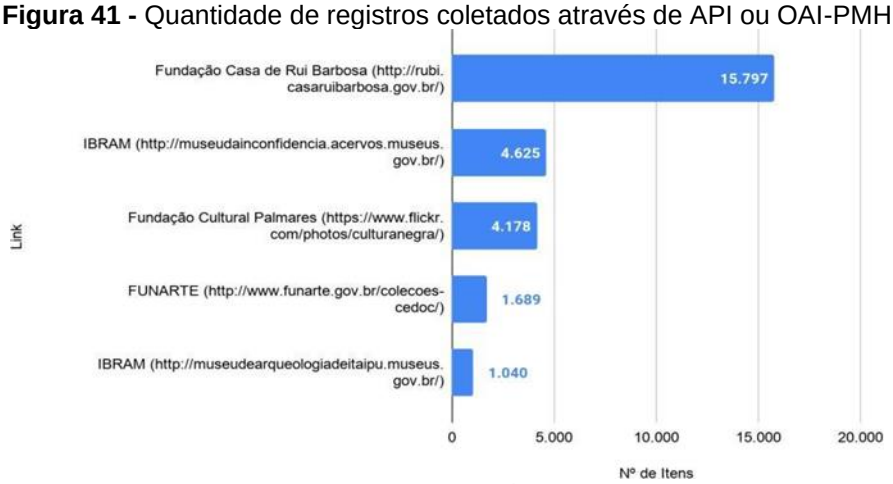
Já quanto à forma de coleta através de exportador, três acervos foram identificados com essa possibilidade, novamente os mesmos que utilizaram o software Tainacan, e que apresentam na interface do repositório a opção de exportação no formato de planilha de dados.

Por fim, somente um acervo da Fundação Casa de Rui Barbosa apresentou a possibilidade de coleta efetiva através do OAI-PMH, com o acervo publicado através do DSpace. Mesmo um dos acervos do IPHAN, publicado em DSpace, não apresenta esse formato de coleta habilitado de acordo com o processo de coleta realizado.

Ainda quanto aos acervos publicados através do Tainacan, que tem a funcionalidade de coleta dos dados através do OAI-PMH, não foi possível realizar a coleta dessa forma, pois é necessário que os metadados do acervo estejam mapeados para um padrão de metadados, e nenhum dos pontos de acesso do OAI-PMH desses acervos apresentaram essa característica, e então são retornados com resultados em branco.

Já considerando a quantidade de registros coletados para cada acervo, como apresenta a Figura 41 abaixo, o acervo da Fundação Casa de Rui Barbosa apresentou a maioria dos registros coletados até então, com 15.797 registros disponíveis através do DSpace e obtidos a partir do OAI-PMH. Logo em seguida o acervo do Museu da Inconfidência, gerido pelo Ibram, com 4.625 registros coletados através da API do Tainacan, na sequência o acervo fotográfico da Fundação Palmares coletado pela API do Flickr com 4.178 registros coletados, e utilizando a API do Tainacan, foi possível realizar a coleta de 1.689 registros de parte do acervo da Funarte, e por fim, 1.040 registros do acervo do Museu de Arqueologia de Itaipu gerido pelo IBRAM.

Figura 41 - Quantidade de registros coletados através de API ou OAI-PMH



Fonte: Elaboração própria, 2020.

5.1.2.1 Funarte

Nos tópicos indicados a seguir serão apresentados os resultados por entidade cultural, evidenciando as características específicas do processo de coleta dos dados, bem como apresentando os números totais de registros coletados para cada acervo selecionado no recorte analisado.

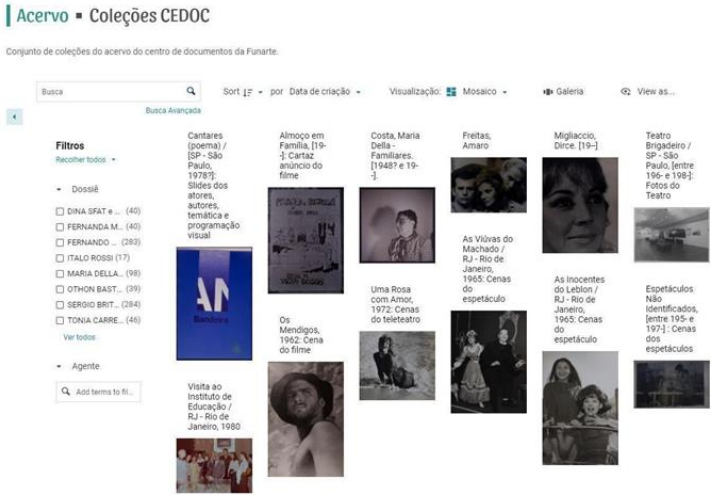
- **Coleções CEDOC**

Para a coleta de dados da Funarte foram selecionados três acervos no recorte, um deles é o acervo denominado “Coleções CEDOC”, que é um conjunto de 847 objetos digitais publicados utilizando o *software* Tainacan (Figura 42), cuja coleta

de dados foi realizada através da API do Tainacan utilizando um programa desenvolvido em Python.

A coleta de dados através da API do Tainacan não apresentou dificuldades para ser realizada, uma vez que sua documentação está disponível on-line e se demonstrou de simples aplicação.

Figura 42 - Página inicial do acervo da Funarte denominado “Coleções CEDOC”



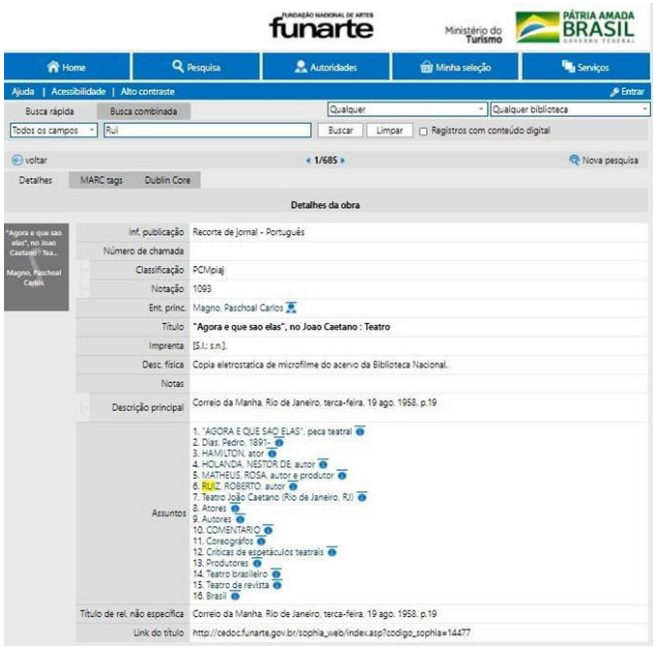
Fonte: Funarte, 2020.

- **Acervo bibliográfico**

Já o segundo acervo selecionado, faz menção ao acervo bibliográfico da Funarte, que é disponibilizado através do software SophiA Biblioteca (Figura 43), em que foi não foi identificada nenhuma função de coleta dos dados através de OAI-PMH, API ou exportadores, levado ao uso da técnica de raspagem de dados.

Sendo assim, foi realizada a coleta dos registros iniciais por intermédio de um programa desenvolvido para acervos publicados através do SophiA Biblioteca.

Figura 43 - Página dos registros de um item do acervo SophiA Biblioteca da Funarte

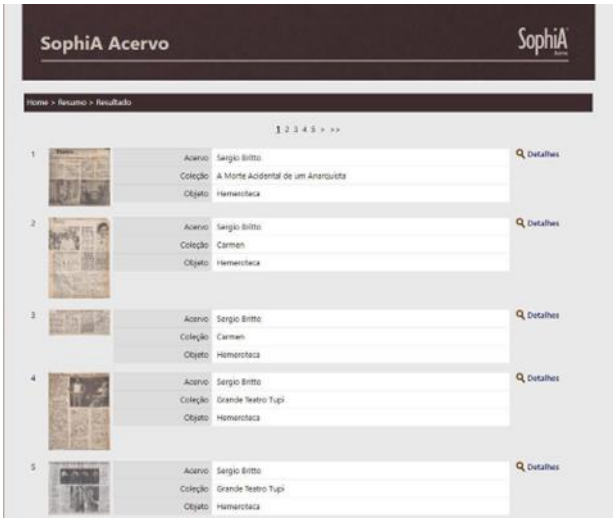


Fonte: Funarte, 2020.

- **Acervo Sérgio Britto**

O terceiro acervo selecionado no recorte para a Funarte é o acervo Sérgio Britto (Figura 44), disponível através do software SophiA Acervo. Neste caso também não foram identificados meios de coleta dos dados através de API, OAI- PMH ou outros exportadores, levando à aplicação da técnica de raspagem de dados para realizar a coleta.

Figura 44 - Página de Objetos do Acervo Sérgio Britto da Funarte



Fonte: Funarte, 2020.

No caso do acervo Sérgio Britto, não foi encontrado nenhum padrão de exibição dos objetos, e foi necessário seguir uma estratégia de aproximação para a coleta dos registros. Essa estratégia foi pesquisar uma palavra-chave abrangente no campo de busca de palavras-chave do repositório, no caso o termo “par”, que retornou 873 registros.

Além do processo apresentar esse contexto de coleta de registros por aproximação, ainda ocorrem erros como a demora de resposta pelo servidor ao navegar entre os objetos para a coleta de dados, o que levou a coletas parciais, e necessitou de atenção aos dados repetidos ao unir as bases segmentadas ao final da coleta.

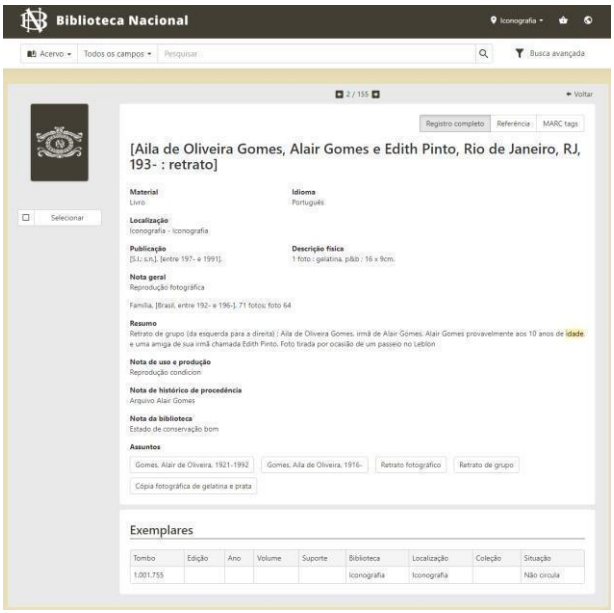
5.1.2.2 Biblioteca Nacional

Quanto à Biblioteca Nacional, foram selecionados dois acervos para coleta, o acervo BNDigital disponível, através do *software* SophiA Biblioteca, e o acervo BN, disponível através do *software* SophiA Acervo. De ambos não foi identificada outra forma de extração a não ser a raspagem de dados, e foram utilizados programas em Python desenvolvidos para a coleta dos registros, utilizando essa técnica.

- **Acervo BN**

Para o acervo BN (Figura 45), foi desenvolvido um programa para a coleta personalizada dos registros, o que se deve à apresentação dos registros em páginas HTML simples, sem a necessidade de interação do usuário. Sendo assim, foi aplicada a raspagem dos dados através do padrão do número identificador do item na URL.

Figura 45 - Página de um item do acervo BN da Biblioteca Nacional

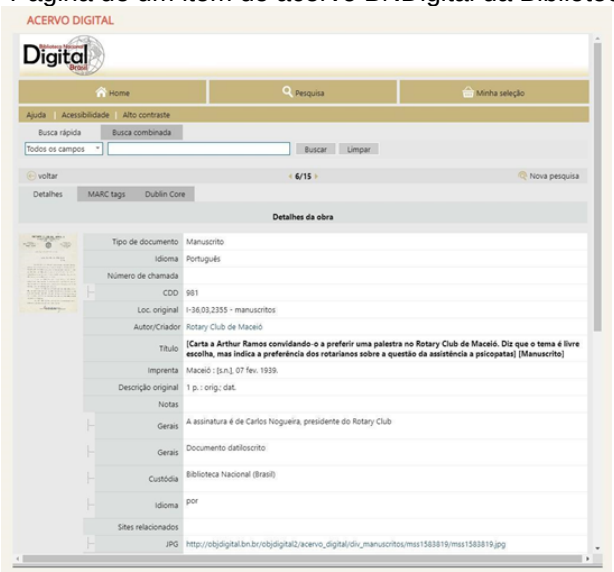


Fonte: Biblioteca Nacional, 2020.

- **Acervo BNDigital**

Já para o acervo BNDigital (Figura 46), que está disponível através do SophiA Biblioteca, foi utilizado o mesmo programa desenvolvido para a coleta dos dados do acervo CEDOC da Funarte, e já que o *software* utilizado é o mesmo, somente foram alterados os parâmetros de coleta de acordo a URL do acervo BNDigital.

Figura 46 - Página de um item do acervo BNDigital da Biblioteca Nacional



Fonte: Biblioteca Nacional, 2020.

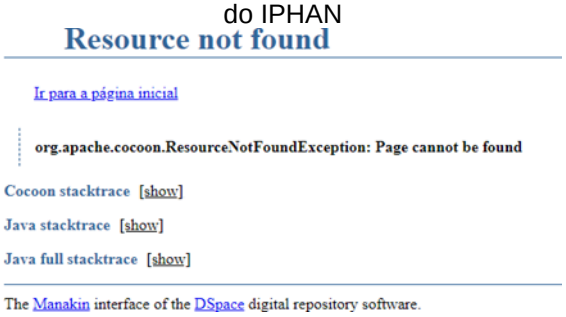
5.1.2.3 IPHAN

Para o IPHAN foram selecionados 4 acervos na amostra, o acervo da rede de arquivos disponível através do *software* DSpace, o acervo bibliográfico disponível por meio do *software* Pergamum, o acervo do Sistema Integrado de Conhecimento e Gestão (SICG) disponível, via *software* próprio, e o acervo de vídeos disponível por intermédio de páginas HTML estáticas. Em todos os casos não foram encontrados outros meios de coleta dos dados a não ser pelo referido método de raspagem.

- **Acervo da Rede de Arquivos**

Quanto ao acervo da rede de arquivos, publicado através do *software* DSpace, mesmo que esse sistema apresente funções de coleta de dados através do protocolo OAI-PMH, foram realizadas tentativas de identificação do *endpoint* (de acordo com a etapa de identificação descrita no tópico 5.2.1), mas não foi encontrada nenhuma opção disponível, e em todas as tentativas foi retornada uma tela de erro mencionando “recurso não identificado” (Figura 47).

Figura 47 - Tela de erro ao identificar acesso ao protocolo OAI-PMH no acervo da Rede de Arquivos do IPHAN



Fonte: IPHAN, 2020.

Dessa forma, foi desenvolvido programa para coleta dos dados utilizando a técnica de raspagem, que interage com a página inicial do acervo (Figura 48) utilizando o botão navegar para exibir todos os registros, e a partir da exibição desses dados realizar a coleta.

Figura 48 - Página inicial do acervo da Rede de Arquivos do IPHAN

Fonte: IPHAN, 2020.

- **Acervo bibliográfico**

Quanto ao acervo bibliográfico disponível através do *software* Pergamum, é caracterizado pelo mesmo contexto de coleta do acervo Sérgio Britto, da Funarte, não havendo padrão explícito na exibição dos objetos, o que leva a interação por palavra-chave no campo de busca. Neste caso, foi utilizada a palavra-chave “IPHAN” para recuperar os registros.

Outro agravante na coleta dos registros desse acervo, além do tempo de resposta lento, foi a limitação de exibição de no máximo 1000 objetos, isso já se apresenta como um limitador, além da própria natureza da estrutura de exibição da página, por falta de padrão de exibição dos objetos, necessitando-se de abordagem por aproximação.

Não foi possível obter imagens da interface do sistema na data atual de produção desta pesquisa, pois o sistema se encontra em manutenção, e a página de acesso ao acervo apresenta conteúdo informando que o sistema está em manutenção (Figura 49).

Figura 49 - Acervo bibliográfico do IPHAN - Página em manutenção

Sistema em manutenção

Fonte: IPHAN, 2020.

- **Sistema Integrado de Conhecimento e Gestão (SICG)**

Quanto ao acervo do Sistema Integrado de Conhecimento e Gestão do IPHAN (Figura 50), ele tem a característica de reunir objetos em 4 contextos diferentes: Ação, Bem Material, Bem Imaterial, e Instituição. Esses contextos de objetos resultam em representação diferente para cada um nas páginas do acervo, já que cada contexto tem seu conjunto de metadados. Dessa forma foi necessário desenvolver um programa de raspagem de dados para cada acervo mapeado, haja vista que o sistema não apresenta nem a opção de coleta via OAI-PMH, API, ou via exportadores.

Figura 50 - Página inicial do SICG

Fonte: IPHAN, 2020.

- **Acervo de vídeos**

O link para o acervo de vídeos (Figura 51), apresenta grande parte do acervo diagnosticado para o IPHAN, que além dos sistemas de repositório que armazenam parte do acervo, o próprio site tem objetos do acervo publicados em páginas HTML dispersas, como é o caso desse acervo de vídeos.

Neste caso, o desenvolvimento do programa para coleta dos dados envolve invariavelmente a aplicação da técnica de raspagem de dados por se tratar de páginas HTML estáticas.

Figura 51 - Acervo de vídeos do IPHAN

Fonte: IPHAN, 2020.

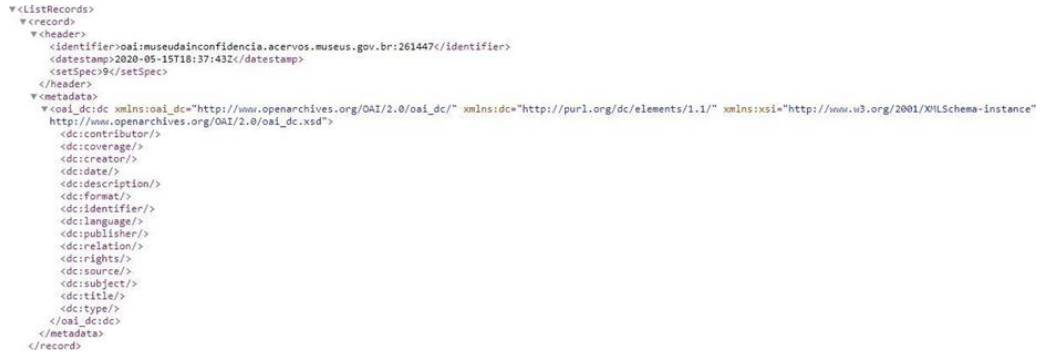
5.1.2.4 Ibram

Para o caso do Ibram foram selecionados dois acervos na amostra, ambos referentes a acervos de museus vinculados à entidade cultural: o Museu da Inconfidência e o Museu de Arqueologia de Itaipu.

O Ibram desenvolveu um projeto de migração dos acervos dos museus sob sua gestão para o software Tainacan , e por isso é comum encontrar esses museus vinculados ao Instituto com acervos on-line disponíveis através desse *software*. Isso já significa uma variabilidade positiva nas possibilidades de coleta dos registros desses acervos, uma vez que o Tainacan possibilita a coleta dos dados públicos através de API, OAI-PHM e exportadores pela interface.

No caso dos dois acervos museológicos selecionados, vale ressaltar que a coleta foi realizada através da API, que se demonstrou a forma automatizada de coleta mais viável, já que apesar de possuir a opção de coleta através do OAI-PMH os metadados do acervo dos dois museus não foram mapeados para um padrão de metadados, e ao tentar realizar a coleta por esse meio são retornados valores em branco (Figura 52). O uso dos exportadores pela interface está funcional, porém como é uma alternativa manual e simples, foi decidido para esta etapa utilizar o programa desenvolvido para coletar dados do Tainacan via API.

Figura 52 - Exemplo de registro do OAI-PMH retornado com valores em branco



Fonte: Museu da Inconfidência, 2020.

- **Museu da Inconfidência**

Dessa forma, como foi possível realizar a coleta dos dados via API, foram coletados 4.625 registros do acervo museológico do Museu da Inconfidência. Não houve maiores barreiras na coleta dos dados, que ocorreu sem erros.

- **Museu de Arqueologia de Itaipu**

Para o Museu de Arqueologia de Itaipu, foi utilizado o mesmo processo de coleta dos dados via API, utilizando o programa já desenvolvido para coleta de acervos disponíveis através do Tainacan, e foram obtidos 1.040 registros, novamente sem problemas identificados.

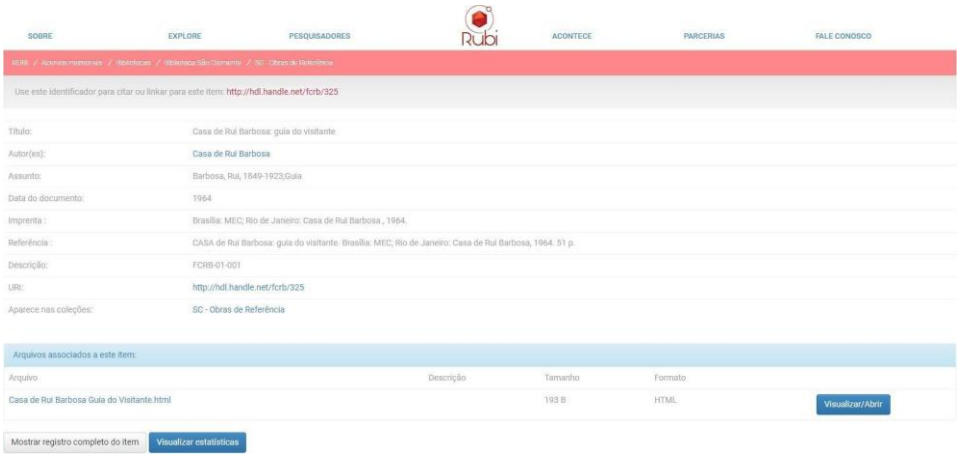
- **Resultados de coleta para a Fundação Casa de Rui Barbosa**

Foram selecionados três acervos para a Fundação Casa de Rui Barbosa. O acervo RUBI, disponível através do *software* DSpace, o acervo Iconográfico disponível através do *software* Fotoweb 7.0, e o acervo bibliográfico disponível através do software SophiA Biblioteca. Para esses dois últimos foram desenvolvidos programas para raspagem dos dados.

- **Acervo RUBI**

O acervo RUBI (Figura 53) está disponível através do *software* DSpace, e com a possibilidade de coleta dos dados via protocolo OAI-PMH; dessa forma foi desenvolvido um programa no modelo apresentado na seção 4.2.1 desta pesquisa, que resultou na coleta de 15.164 registros, já padronizados no formato Dublin Core.

Figura 53 - Página de um objeto do acervo RUBI



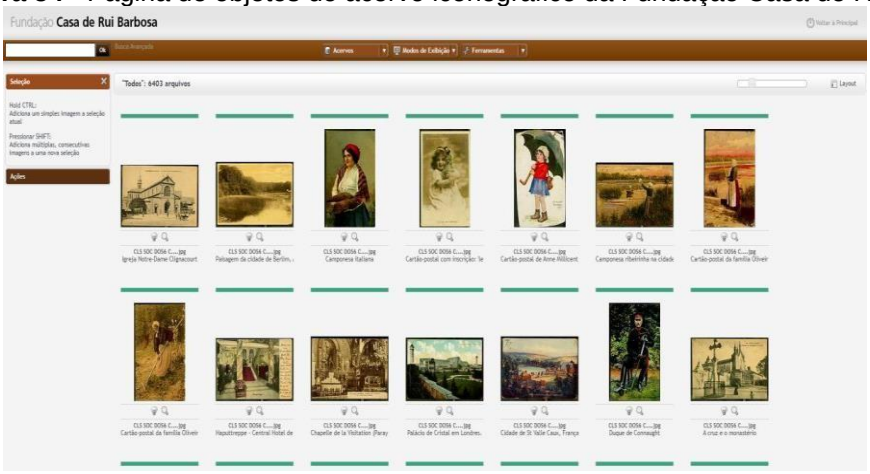
Fonte: Fundação Casa de Rui Barbosa, 2020.

O uso da coleta via protocolo OAI-PMH, se demonstrou eficiente, assim como a coleta via API, que não apresentou barreiras ou erros na execução do programa, e sua utilização foi de fácil execução.

- **Acervo iconográfico**

Para o acervo iconográfico da Fundação Casa de Rui Barbosa (Figura 54), não foram identificadas formas de coleta de dados através de API, OAI-PMH ou ainda algum exportador na interface do repositório. Assim foi desenvolvido um programa de raspagem dos dados para coleta, que aproveitou que o sistema possui a opção de listagem de todos os objetos do acervo sem a necessidade de interagir com filtros ou caixas de busca textual.

Figura 54 - Página de objetos do acervo iconográfico da Fundação Casa de Rui Barbosa

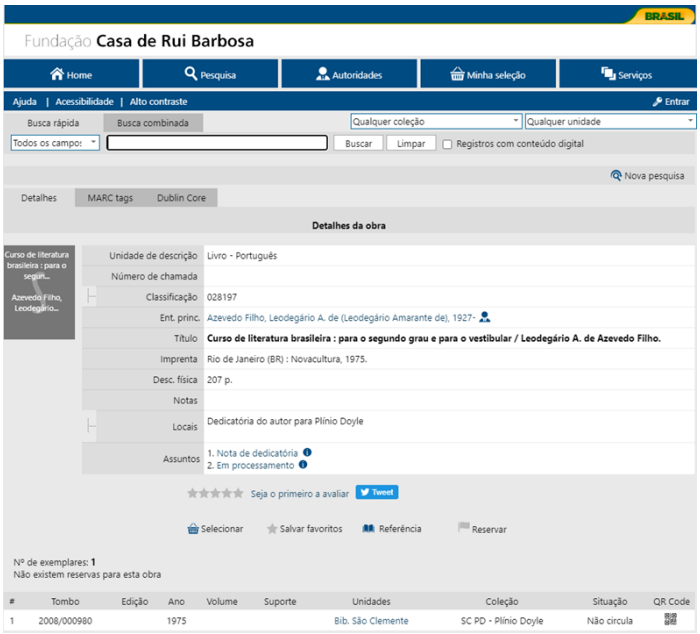


Fonte: Fundação Casa de Rui Barbosa, 2020.

- **Acervo Bibliográfico**

Por sua vez, o acervo bibliográfico (Figura 55), disponível através do *software* SophiA Biblioteca, está situado no mesmo contexto de exibição de objetos que estão os acervos bibliográficos da Biblioteca Nacional e da Funarte, o que permitiu o uso do mesmo programa para a coleta através da aplicação da técnica de raspagem de dados, se atentando somente à raiz da URL dos objetos.

Figura 55 - Página de um objeto no acervo bibliográfico da Fundação Casa de Rui Barbosa



Fonte: Fundação Casa de Rui Barbosa, 2020.

5.1.2.6 Fundação Cultural Palmares

Para a Fundação Cultural Palmares, foram selecionados dois acervos, um acervo fotográfico disponível diretamente no site da entidade através de uma página HTML estática, e um outro acervo fotográfico disponível na plataforma Flickr.

- **Acervo fotográfico do site**

Para o acervo fotográfico disponível diretamente no site da Fundação Cultural Palmares (Figura 56), foi desenvolvido um programa de raspagem de dados mais simples, visto que os objetos estão situados no contexto de uma página HTML estática, e não possuem outra forma de coleta, sendo assim foram obtidos os 13 registros presentes na página.

Figura 56 - Página do acervo fotográfico no site da Fundação Cultural Palmares

Fonte: Fundação Cultural Palmares, 2020.

- **Acervo fotográfico no Flickr**

Para o acervo fotográfico da Fundação Cultural Palmares publicado através do Flickr (Figura 57), foi possível utilizar a API da plataforma para coletar os registros públicos, a possibilidade mais automatizada disponível nesse caso. Para isso, foi desenvolvido um programa na perspectiva da coleta via API citado nesta pesquisa na seção 4.2.1, para acessar os dados das fotografias publicadas pelo usuário da Fundação Cultural Palmares, obtendo 4.178 registros.

Figura 57 - Acervo fotográfico da Fundação Cultural Palmares no Flickr

Fonte: Fundação Cultural Palmares, 2020.

Novamente a coleta dos dados a partir da API se mostrou sem barreiras específicas, porém vale salientar o limite de consultas imposto pelo Flickr, de 3.600 consultas por hora, informação disponível na documentação da API.

5.1.2.7 Discussão analítica sobre resultados de coleta de dados dos acervos

Como conclusão desta etapa analítica, entende-se que foi identificada a viabilidade da coleta dos dados dos acervos digitais das instituições vinculadas ao Ministério da Cultura do Brasil. Mesmo que essa etapa da pesquisa tenha apresentado um recorte dos acervos identificados (tópico 4.2.1.2.1), as condições de representação apontam para a possibilidade de generalização das condições de coleta, o que foi comprovado posteriormente, e a etapa de publicação dos acervos agregados (4.2.3) foi executada com a totalidade de registros passíveis de coleta em todos os acervos identificados.

Além disso, como produto da coleta realizada nesta etapa do projeto, têm-se as bases de dados com os registros coletados dos acervos, o que possibilitará a execução da etapa de mapeamento dos metadados utilizados pelas entidades na descrição dos objetos, que buscará além de descrever e entender a disposição conceitual dos metadados desses acervos, propor e mapear um possível padrão de metadados de agregação dos acervos.

Ainda vale o destaque para alguns pontos evidentes nos resultados desta etapa da pesquisa:

- A possibilidade de coleta de dados dos acervos através do protocolo OAI-PMH, ou através de API, indicam potencial importante de reúso automatizado dos dados. No caso da busca pela agregação dos acervos, é mais eficiente promover coleta automatizada daqueles que possuem esse tipo de funcionalidade, pois já contempla estrutura de disponibilização dos dados mais robusta em formatos estruturados e semiestruturados como JSON e XML, e até com padrão de metadados definido, como é o caso do protocolo OAI-PMH.
- Quando não há meio automatizado de obtenção dos dados, como OAI- PMH ou API, a aplicação da técnica de raspagem de dados é um caminho. Porém, são apresentadas diversas barreiras como o erro de limitação de tempo de resposta do servidor, que por vezes faz com que seja necessário dividir o resultado do programa em partes para serem reunidas depois, ou ainda a

necessidade de aproximação, em alguns casos de coleta em que não é identificado padrão de exibição de objetos que permita a coleta do acervo completo. Desse modo, a raspagem de dados se apresenta como forma menos viável de coleta dos dados, quando o objetivo é a agregação dos acervos, dificultando a recuperação e reúso dos dados.

Ao comparar essa experiência de coleta com as informações levantadas na revisão das iniciativas de agregação de acervos digitais culturais de diferentes nações (tópico 3), percebe-se uma um hiato tecnológico na maioria dos casos encontrados. A necessidade de desenvolvimento de programas personalizados para coleta dos dados, além de indicar a ineficiência de coleta, indica também que não há uma preocupação com a democratização do acesso aos acervos refletida nas escolhas ou configurações dos sistemas de recuperação da informação utilizados.

A publicação dos acervos na internet abre uma série de oportunidades, inclusive o avanço na democratização do acesso, porém utilizar a internet para essa publicação também exige constante atualização e acompanhamento por parte das entidades culturais detentoras destes acervos. A gama de tecnologias disponíveis para melhorar ainda mais as condições de acesso, recuperação e reúso das informações desses acervos vão além da simples disponibilização para consulta on-line.

Em busca de solucionar essas questões técnicas, seria necessária uma revisão da forma como os acervos são estruturados e documentados. Assim como ocorre nas iniciativas de agregação, é imprescindível que exista consenso entre as instituições do patrimônio cultural, em prol da qualidade dos dados e da agregação dos acervos. Princípios atuais como indicados pela IFLA (2009, 2016, 2020) e pelo contexto dos dados abertos ligados (Bizer; Heath; Berners-Lee, 2009; Machado; Souza; Simões, 2019) são essenciais para melhoria da situação tecnológica atual.

5.2 Desenvolvimento do protótipo de agregação dos acervos das entidades vinculadas ao MinC

Neste tópico serão apresentados os resultados do desenvolvimento do protótipo de agregação proposto para esta pesquisa. A produção das ferramentas de coleta dos dados dos acervos, bem como o mapeamento dos metadados dos acervos coletados, e a publicação desses dados agregados vão atender ao objetivo proposto

de desenvolver um protótipo de agregação dos acervos digitais padronizados sob um único padrão de metadados.

5.2.1 Ferramentas de coleta dos dados dos acervos

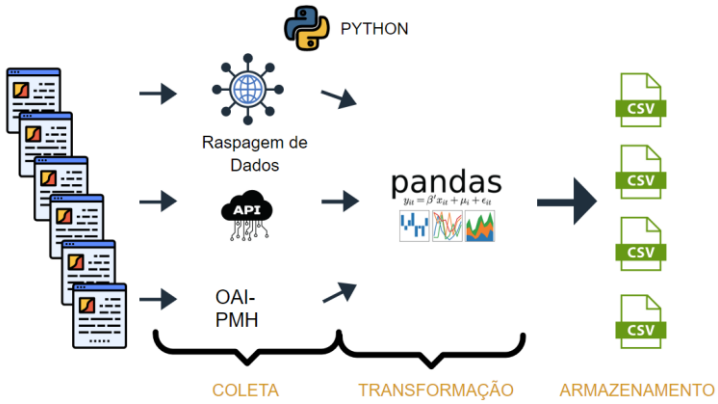
O estágio de coleta de dados aplicado nesta pesquisa é diferente dos estágios identificados na revisão das iniciativas de diferentes nações (tópico 3), principalmente porque a aplicação da estratégia de agregação desta pesquisa envolveu a descoberta dos provedores de dados e de seus acervos a partir da caracterização dos acervos digitais das entidades vinculadas ao MinC (tópico 5.1.1).

Isso implicou a necessidade de promover as formas de coleta por parte única do serviço de agregação, sem a atuação das entidades culturais na adequação e disponibilização de seus acervos em formatos facilitados de acesso e coleta, como ocorre nas iniciativas mapeadas (Siqueira *et al*., 2021; Siqueira; Martins, 2020; Siqueira, Martins, 2022).

Todo o processo de coleta de dados foi desenvolvido utilizando a linguagem de programação Python. Foram produzidos um total de 23 programas na linguagem de programação Python, que tiveram de ser customizados para coletar as 41 fontes de informação identificadas dos acervos digitais envolvidos no projeto. Todos os arquivos dos programas resultantes dessa etapa estão disponíveis no repositório de códigos on-line GitHub (https://github.com/tainacan/data_science/tree/master/FAPESP).

Para esta etapa de coleta, de modo geral, os dados coletados foram armazenados utilizando duas tecnologias de estrutura de matriz (linhas e colunas): o DataFrame da biblioteca de programação Pandas para processamento pelo programa desenvolvido, e o formato CSV para armazenamento dos produtos dos programas em arquivos de planilhas (Figura 58).

Figura 58 - Esquema do estágio de coleta dos dados dos acervos das entidades culturais



Fonte: Fonte: Elaboração própria, 2023.

Abaixo serão apresentadas as metodologias de desenvolvimento dos programas para cada forma de coleta identificada.

5.2.1.1 OAI-PMH

O protocolo OAI-PMH é uma forma comum de comunicação entre sistemas, então já se esperava que alguns *softwares* como Atom e DSpace, por exemplo, tivessem essa funcionalidade. Mas ainda assim, pelo contexto de aplicação desta pesquisa, que envolveu o desenvolvimento de programas de coleta, foi necessário identificar o ponto de acesso (*endpoint*) ao protocolo OAI-PMH, que pode ser entendido como o link que dá acesso à função de consulta dos dados do acervo.

Essa informação pode estar descrita no próprio site da entidade ou na documentação do *software* utilizado, facilitando desse modo o acesso à funcionalidade, ou, ainda, ela pode ser identificada a partir da tentativa de algumas opções de flexão do próprio *endpoint*, como a adição do sufixo “/oai?verb=Identify” ou ainda o sufixo “oai/request?verb=Identify”.

O acesso aos dados dos acervos através do OAI-PMH pode ser configurado pelo software como público, sem restrições, como foi o caso dos acervos identificados publicados através do DSpace, cujo caminho de acesso padrão aos dados é “/oai/request; e do Tainacan , cujo acesso aos dados é alcançado adicionando-se ao link do acervo o caminho “/wp-json/tainacan/v2/oai”. Ou ainda privado, que necessita de uma chave de acesso, como foi o caso dos acervos identificados publicados através do Atom, que inclui a opção de requisitar a autenticação de acesso aos dados

do acervo através do comando de exemplo: “curl -v -H "X-OAI-API-Key: caaac1a110b771bf" "http://example-site.com/;oai?"”.

No nível de processamento dos dados, é realizada a requisição do verbo de listagem dos registros “ListRecords” e a partir de um modelo de metadados identificado através do verbo “ListMetadaFormats”. Normalmente os acervos publicados através do OAI-PMH são disponibilizados utilizando-se o padrão de metadados Dublin Core, ao permitir a recuperação dos metadados já alinhados a esse modelo semântico.

Os dados coletados são exportados por padrão no formato XML, que é um formato semiestruturado cuja exposição é feita através de hierarquia de tags, como apresenta a Figura 59 abaixo, com o exemplo de um registro do acervo RUBI da Fundação casa de Rui Barbosa:

Figura 59 - Exemplo de registro em XML obtido via OAI-PMH

```
<oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:doc="http://www.lyncode.com/xoai"
xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
<dc:title>Balzac ignoré</dc:title>
<dc:creator>Cabanès, Augustin, 1862-1928</dc:creator>
<dc:subject>Balzac, Honoré de, 1799-1850</dc:subject>
<dc:description>Assinalado por Rui</dc:description>
<dc:description>Livro ilustrado com 36 gravuras</dc:description>
<dc:description>Notas de rodapé</dc:description>
<dc:date>2013-05-07T17:47:10Z</dc:date>
<dc:date>2013-05-07T17:47:10Z</dc:date>
<dc:date>1911</dc:date>
<dc:type>Livro raro</dc:type>
<dc:identifier>CABANÈS, Augustin. Balzac ignoré. 2 ed. rev. et augm. Paris (França) : Albin Michel, [1911?].</dc:identifier>
<dc:identifier>http://hdl.handle.net/fcrb/179</dc:identifier>
<dc:language>francês</dc:language>
<dc:relation>2 ed. rev. e aument.</dc:relation>
<dc:rights>Domínio Público</dc:rights>
<dc:format>288 p : il., 36 grav.</dc:format>
<dc:publisher>Paris (França) : Albin Michel, [1911?]</dc:publisher>
</oai_dc:dc>
```

Fonte: Acervo RUBI da Fundação Casa de Rui Barbosa, 2020.

Para executar o processo de listagem dos registros e processá-los em estrutura de dados reutilizável, foi desenvolvido um programa em Python, utilizando a biblioteca Sickle que permite a coleta dos dados de um *endpoint* que utiliza o método “ListRecords” juntamente com a informação do prefixo do padrão de metadados. Para ler os dados em XML foi utilizada a biblioteca ElementTree XML , e para armazenar os dados em DataFrame foi utilizada a biblioteca Pandas, permitindo que esses dados sejam exportados para diferentes formatos posteriormente.

5.2.1.2 API

A coleta de dados através de API tem algumas semelhanças à coleta realizada através do protocolo OAI-PMH, porém a principal diferença está na estruturação e exibição dos dados.

Quanto à identificação da existência de acesso aos dados do acervo através de API, busca-se através da página do acervo alguma informação sobre a API, ou busca-se sobre documentação da API nos sites relacionados aos *softwares* dos repositórios. Nesse caso não há tentativa de identificar o *endpoint* de acesso à API pois ele não segue padrão específico, apresentando variações do caminho de acesso de acordo com a configuração do *software*.

As páginas de documentação das APIs contêm a explicação da estruturação dos dados e a listagem dos métodos de coleta e modificação, que geralmente são construídos com base em contextos. Por exemplo, existe uma estrutura, para as coleções de objetos, para os metadados, para os registros e assim por diante. A forma como esses contextos são estruturados e posteriormente traduzidos para métodos, torna possível o acesso e coleta dos dados do acervo.

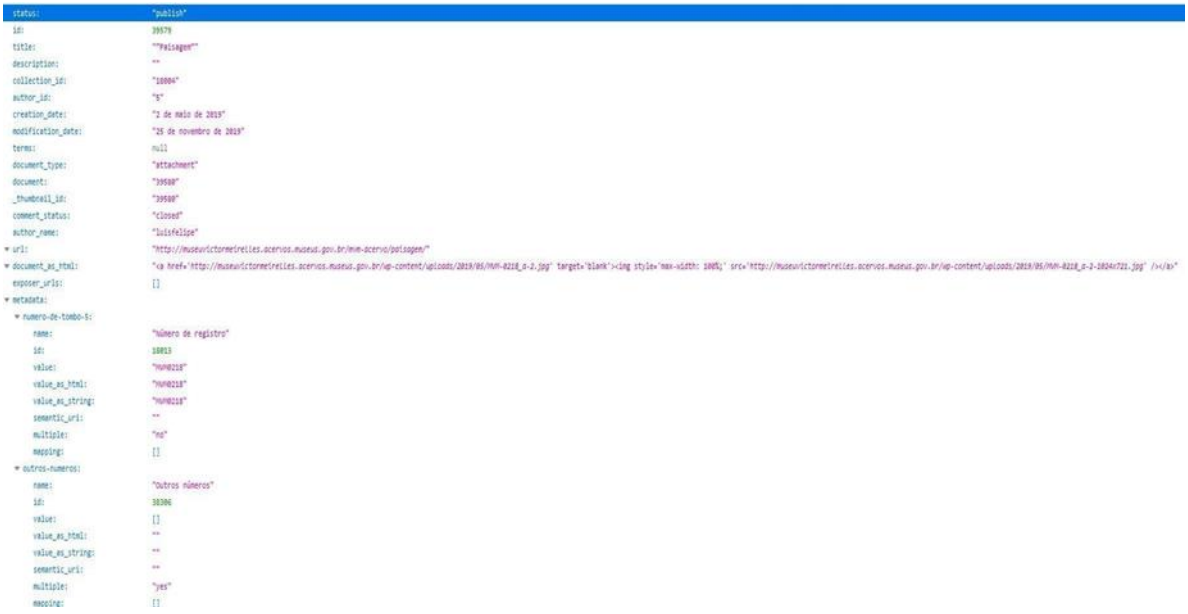
Uma API é uma aplicação que vai além da função de disponibilizar os metadados para coleta e interoperabilidade, e ela permite também, mediante autenticação, realizar modificações através das funcionalidades do sistema. Essa é a realidade de algumas iniciativas descritas no tópico 3 desta pesquisa, principalmente no caso da Europeana.

Como nosso objetivo é a coleta dos dados para promover a agregação dos acervos, o escopo de coleta foi limitado às APIs que permitem a coleta dos dados públicos sem a necessidade de autenticação. Para isso, o primeiro passo foi a identificação do link que dá acesso à API de exibição dos dados do acervo. Esse tipo de link difere para cada sistema, no caso do Tainacan, por exemplo, é identificado através da adição do sufixo “/wp-json/tainacan/v2/” ao link do acervo.

Uma vez identificada a documentação da API, e seu link de acesso, o nível de processamento dos dados se deu através da identificação dos métodos da API que direcionam para a coleta dos dados, normalmente métodos associados ao termo “GET”. Uma vez identificados esses métodos e suas relações, é possível recuperar os dados através dos formatos disponíveis, geralmente uma API disponibiliza os dados em JSON, porém há casos em que é possível escolher mais de um formato.

Na Figura 60 abaixo, é apresentada parte do registro de um item do acervo museológico do Museu Victor Meirelles, no formato JSON, e foi retornado através da consulta à API.

Figura 60 - Representação parcial de um item em JSON obtido através de API



Fonte: Tainacan do acervo museológico do Museu Victor Meirelles, 2020.

Ainda no nível de processamento, para coletar esses dados também foi necessário o desenvolvimento de programa utilizando Python. As bibliotecas implementadas foram: Requests, para fazer a requisição aos links de acesso das APIs e obter como resultado os dados no formato JSON; Pandas, para armazenamento dos dados; Time, para lidar com o limite de requisições por tempo. Para esse programa foi crucial o entendimento da estruturação da exibição dos dados, já que eles são disponibilizados em JSON, os dados são estruturados em formato de chave e valor, o que geralmente pode ser traduzido para os metadados e seus valores; portanto, entender quais chaves são essas e quais valores elas representam é o que garante a confiabilidade dos dados quanto a sua representação.

5.2.1.3 Exportadores

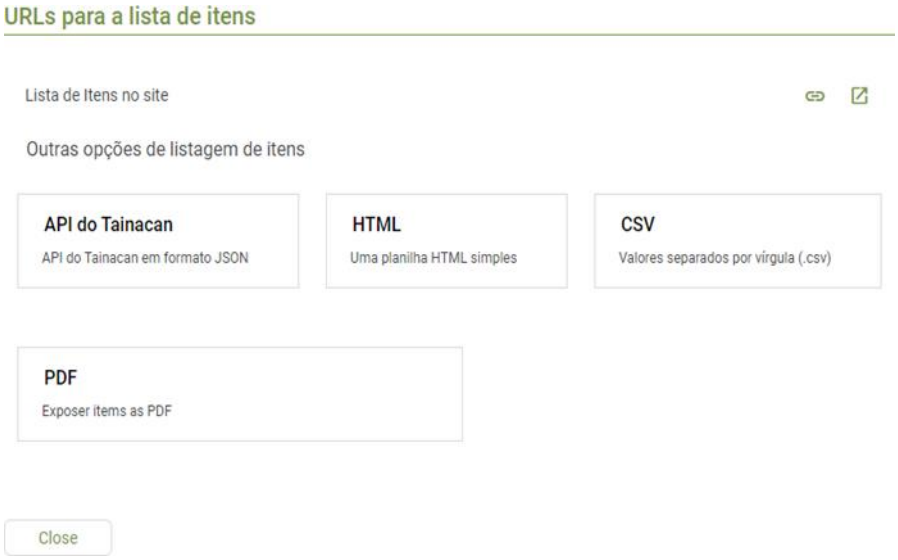
O método de coleta através de exportadores é realizado a partir da própria interface de consulta do acervo; a possibilidade de extração dos dados por exportadores é disponibilizada por alguma funcionalidade do sistema de exposição

dos objetos do acervo, e é conduzida a partir da interação através da seleção dos objetos que se deseja exportar e escolher o formato de exportação.

O processamento de dados se limita às possibilidades da funcionalidade de exportação do acervo, se é possível filtrar ou escolher manualmente os objetos, e ainda se possibilita a extração de dados em mais de um formato. Uma vez configurada a exportação dos dados, o último passo é processar com a confirmação de exportação e fazer o download do arquivo exportado.

A Figura 61 abaixo apresenta um exemplo da funcionalidade de exportação de dados do Tainacan.

Figura 61 - Funcionalidade de exportação de dados do Tainacan



Fonte: Tainacan do acervo museológico do Museu Victor Meirelles, 2020.

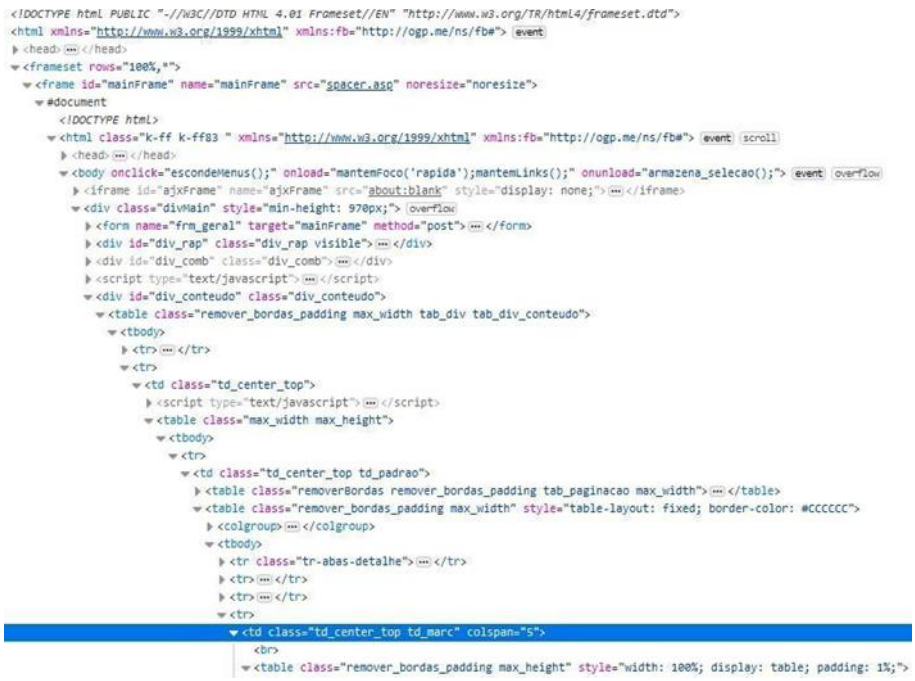
No caso do Tainacan, é possível filtrar os objetos pela interface do acervo, e exportar os dados através do ícone “ver como” no canto direito da interface. Através do clique nesse ícone as opções apresentadas na figura 8 aparecem, e é possível exportar os dados públicos do acervo em cinco formatos diferentes: JSON, através de consulta realizada previamente na API e intermediada pela interface; CSV; HTML; e PDF.

5.2.1.4 Raspagem de dados

A coleta de dados através da raspagem de dados foi a última alternativa utilizada, quando não foram identificadas outras opções de coleta. A raspagem de dados pode ser executada a partir de quaisquer páginas exibidas da web que estão estruturadas sob formato do tipo HTML, que foram identificados utilizando a inspeção da página através do navegador. (Firefox, Chrome, Opera, Safari, etc.)

A Figura 62 abaixo apresenta o exemplo de inspeção da página de um item do acervo bibliográfico da Biblioteca Nacional.

Figura 62 - Inspeção de uma página HTML



Fonte: Acervo bibliográfico da Biblioteca Nacional, 2020.

Vale neste ponto destacar que uma página HTML é uma estrutura hierárquica composta de tags, que são chaves que situam o conteúdo dentro da página. Por exemplo, a tag “ ” armazena o conteúdo textual de um parágrafo; a tag “ ” é utilizada como um contêiner, que armazena outras estruturas de tags que são de um contexto específico. Outro exemplo, é a tag “<table></table>”, que armazena a estrutura de uma tabela através da tag “<td><td>” para indicar as células; a tag “<th><th>” para indicar as colunas, e a tag “<tr><tr>” para indicar as linhas. Essa estrutura de tabela é comumente utilizada para exibir os metadados dos registros na página do objeto.

Outro atributo importante para a coleta de dados através da raspagem de dados é a existência de um padrão de exibição. Por exemplo, quando se tem um acervo no qual é possível listar todos os objetos do acervo, e estes são dispostos divididos em páginas se torna possível desenvolver um programa que percorra todos os objetos e todas as páginas coletando todos os registros, que é o caso do acervo da Rede de Arquivos do IPHAN, em que ao clicar no botão “navegar” na página inicial, o usuário é redirecionado para uma página com 10 objetos, e à medida que o usuário rola a página para baixo, mais 10 objetos são carregados até apresentar o total de 38.022 registros do acervo.

Outro padrão comum é a existência de identificadores em links que indicam uma sequência numérica relacionada ao objeto do acervo, e assim é possível coletar os registros a partir do identificador 1 até o enésimo identificador, que seria uma estimativa do total de objetos do repositório. Esse é o caso dos acervos publicados através do SophiA Biblioteca, que não apresentam a função de listar todos os objetos, porém existe um padrão de identificação dos objetos no link, como por exemplo este link de um item publicado no acervo da Funarte “http://cedoc.funarte.gov.br/sophia_web/index.asp?codigo_sophia=19460” em que esse número após o sufixo “codigo_sophia” é o número do item, e dessa forma é possível acessar a página HTML com o conteúdo alterando o número no final da URL.

E ainda há outros casos em que não são encontrados padrões explícitos na exibição dos registros, e nem é permitida a listagem de todos os objetos, sendo necessária a interação do usuário com a interface através de algum filtro ou busca textual para retornar uma parte do acervo. Nesses casos não fica clara a quantidade total de objetos, limitando a raspagem de dados a alguma estratégia que envolva utilizar palavras-chave genéricas na busca textual do repositório e coletar os registros por aproximação, o que necessita de posterior limpeza da base de dados coletados para remover possíveis duplicações.

Além da estrutura HTML das páginas, em alguns casos, como no contexto de repositórios digitais que não disponibilizam outros formatos de coleta de dados, é necessário automatizar a interação do usuário com a interface do repositório para que os objetos sejam apresentados. É o caso, por exemplo, do acervo Sérgio Britto da Funarte, em que foi necessário inserir um termo de busca para listar alguns registros do acervo, não possibilitando a listagem de todos os registros.

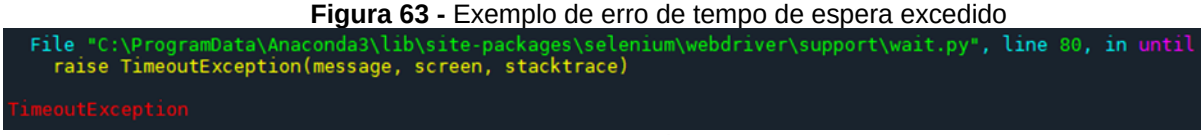
É através da ferramenta de inspeção da página que se dá a identificação primária dessas hierarquias da estrutura de exposição de dados, por meio dessa inspeção é possível interagir diretamente clicando nas áreas da página que se deseja coletar, e identificar na estrutura HTML qual a posição do conteúdo na hierarquia.

Uma vez identificada a estrutura HTML das páginas, o processamento dos dados envolveu o desenvolvimento de um programa, cuja ação de acessar e traduzir o formato exibido em HTML para o formato de um DataFrame é particular para cada link de acervo.

Brevemente, a estrutura geral de raspagem de dados apresentada neste projeto é encaminhada de duas formas: a primeira se dá quando as páginas são estáticas e não é necessária a automatização de interação com a interface. Nesse caso a biblioteca de programação em Python utilizada é a BeautifulSoup, que permite executar o processo de dividir a página HTML e coletar diretamente os registros dos objetos do acervo a serem coletados nos blocos de conteúdo específicos. Já quando o acesso à página HTML depende da interação do usuário com funcionalidades, como filtros e busca textual, é utilizada a biblioteca Selenium.

Uma das desvantagens comuns desse tipo de coleta é a dificuldade de comunicação com o servidor que disponibiliza os dados na web, uma vez que a raspagem de dados é executada através do envio da solicitação de uma página ao servidor e espera-se uma resposta do servidor com a página solicitada; então a lentidão nesse processo pode levar o programa a retornar erro de tempo de espera (Figura 63), já que é necessário indicar um tempo limite de espera pela página requisitada, no caso foi utilizado um tempo máximo de 2 minutos.

Figura 63 - Exemplo de erro de tempo de espera excedido



Fonte: Elaboração própria, 2020.

5.2.3 Mapeamento dos metadados dos acervos coletados para padrão de metadados agregador

O estágio de mapeamento dos metadados desta pesquisa ocorreu em duas etapas: na identificação de padrão de metadados de agregação a partir dos metadados utilizados pelas entidades vinculadas ao MinC, e na implementação do mapeamento dos metadados utilizados por essas entidades cultuais para o padrão de metadados proposto para agregação.

O processo de mapeamento de metadados no caso de estratégias de agregação, como observado na revisão das iniciativas de agregação de diferentes nações (tópico 3), envolve um contexto de articulações entre as instituições do patrimônio cultural, e a instituição que promove o serviço de agregação. Esse contexto se dá justamente para alinhar as expectativas quanto à organização do conhecimento e representação da informação dos acervos entre as instituições.

No caso da Europeana e DPLA, por exemplo, há contexto de pactuação entre os provedores de dados e o serviço de agregação para garantir a qualidade dos dados (Siqueira *et al*., 2021). Já no caso da Brasiliana Museus, a agregação dos acervos digitais se deu dentro do órgão responsável pela gestão dos museus (Ibram). Nas demais iniciativas não fica claro como essa articulação se deu, entretanto é possível observar a necessidade deste alinhamento entre o provedor e o agregador.

Como a proposta de protótipo desta pesquisa se dá em contexto diferente, o método de mapeamento aplicado se baseou no referencial teórico indicado no tópico 2.2. A partir desse referencial foi possível executar as duas etapas que compõem este estágio de mapeamento, que serão apresentadas nos tópicos abaixo.

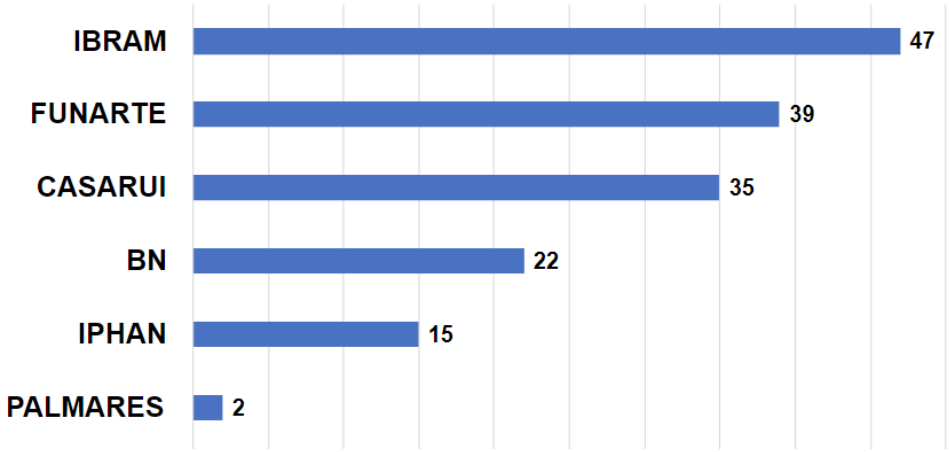
5.2.3.1 Padrão de metadados de agregação

A escolha do padrão de metadados adotado para agregação dos acervos foi executada com base em análise dos metadados dos acervos obtidos na etapa de coleta deste estudo. Esses metadados foram coletados no mesmo recorte de 16 acervos utilizados para a análise dos resultados de coleta apresentados no tópico 5.1.2.

A maior quantidade de metadados diferentes encontrados está presente nos acervos analisados do Ibram (47 metadados diferentes). Já a Funarte, a Fundação

Casa de Rui Barbosa, a Fundação Biblioteca Nacional e o Instituto do Patrimônio Histórico e Artístico Nacional apresentaram entre 15 e 39 metadados diferentes em seus acervos. E quanto à Fundação Cultural Palmares, foram identificados somente dois metadados diferentes (Figura 64).

Figura 64 - Quantidade de metadados identificados nos acervos das entidades vinculadas ao MinC

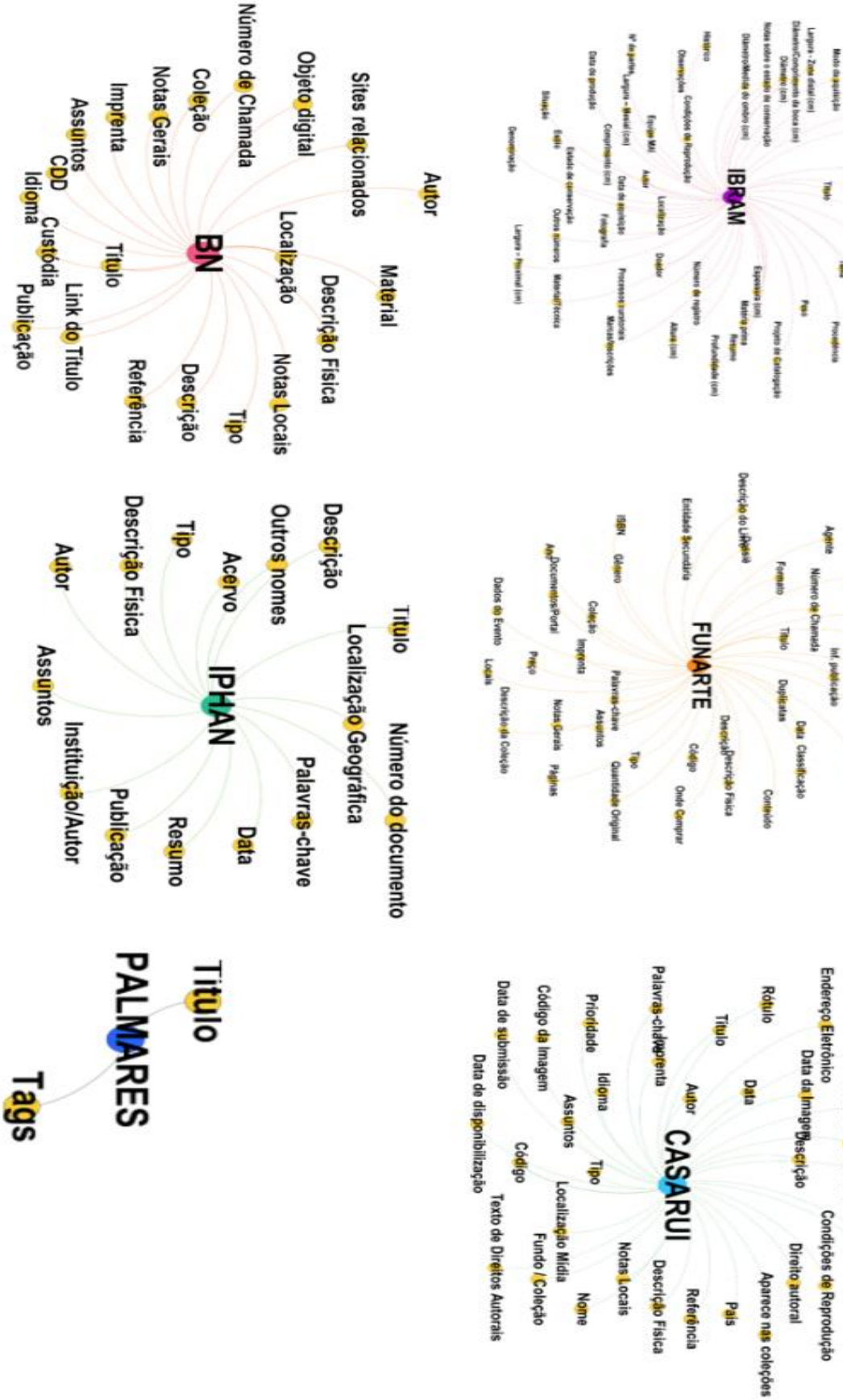


Fonte: Elaboração própria, 2023.

Para complementar esta análise dos metadados coletados em busca de propor um padrão de metadados para agregação, foi utilizada a visualização de grafos (tópico 2.2.2), para apresentar visualmente os conjuntos de metadados entre as entidades culturais.

Os grafos abaixo apresentam o conjunto de metadados dos acervos analisados para cada entidade cultural. É possível perceber, de forma complementar ao quantitativo de metadados coletados, a variabilidade do uso de diferentes metadados entre as entidades vinculadas.

Figura 65 - Grafos dos conjuntos de metadados utilizados pelas entidades culturais analisadas



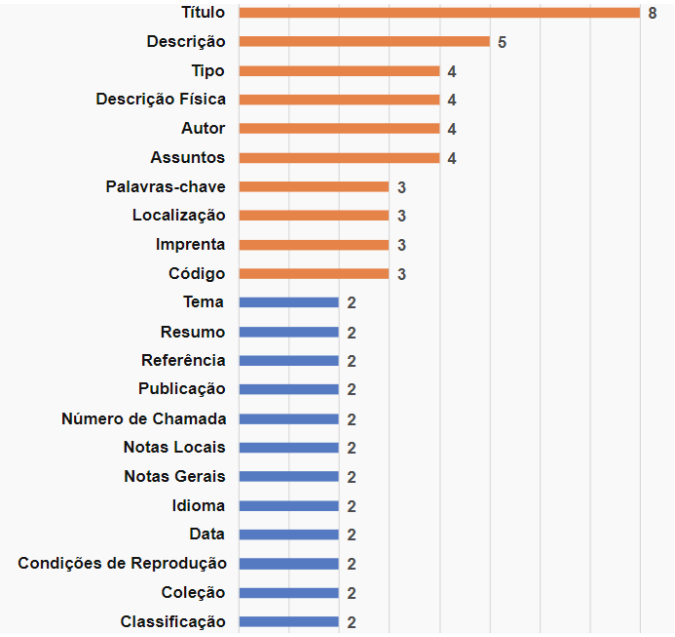
Fonte: Elaboração própria, 2023.

Essa variabilidade de metadados identificada até então, corrobora os resultados identificados através da categoria analítica de “Organização e representação da informação” descritos na etapa de caracterização dos acervos (4.1.1). Não há padrão de metadados explícito através da exposição dos acervos no site das entidades culturais. Isso já indica o desafio que é definir um padrão de metadados para agregação desses acervos.

Dessa forma, a definição do padrão ocorreu ao realizar o cruzamento dos metadados utilizados em comum entre os acervos das entidades culturais analisadas. Essa co-ocorrência de metadados é interpretada no contexto desta pesquisa como um indicativo do conjunto de metadados que mais represente os acervos de maneira geral.

Como apresenta a Figura 66 abaixo, um conjunto de 10 metadados aponta aqueles que mais apareceram entre os acervos analisados. São eles: “título”, “descrição”, “tipo”, “descrição física”, “autor”, “assuntos”, “”palavras-chave”, “localização”, “imprenta”, e “código”. Esses metadados apareceram entre três e oito dos 16 acervos analisados.

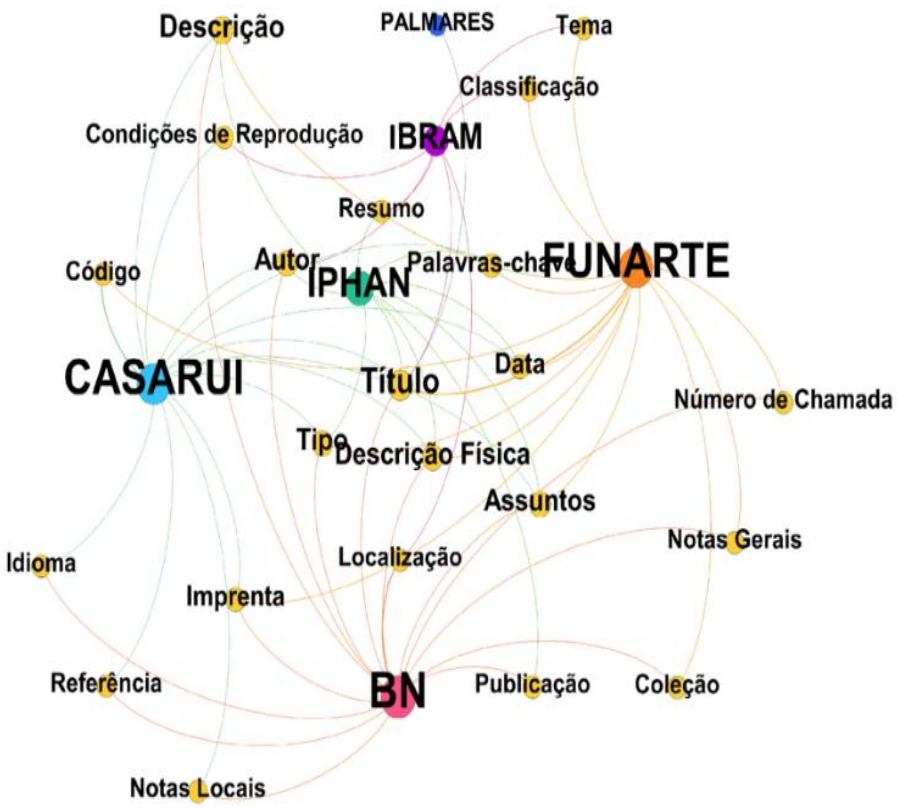
Figura 66 - Metadados recorrentes entre os acervos analisados



Fonte: Elaboração própria, 2023.

Da mesma maneira, através da visualização por grafos é possível observar como esses metadados em comum se distribuem entre as entidades culturais (Figura 67). As entidades com maior tamanho do vértice no grafo são aquelas cujos acervos têm mais metadados em comuns com as demais.

Figura 67 - Grafo dos metadados em comum entre os acervos das entidades culturais analisadas



Fonte: Elaboração própria, 2023.

Um dos principais resultados dessa análise de co-ocorrência de metadados entre os acervos analisados é a semelhança deste conjunto de metadados com o esquema de metadados principais proposto pelo Dublin Core. São 15 os elementos propostos por esse esquema (Tabela 6), que tem como premissa a representação de objetos digitais na web de forma geral.

Tabela 6 - Esquema dos principais metadados do Dublin Core

Metadado Dublin Core	Definição
Título	Nome dado ao recurso.
Autor	Entidade responsável principalmente pela produção do recurso.
Assunto/palavras-chave	O tópico do recurso.
Descrição	A descrição do recurso.
Editor	Entidade responsável por disponibilizar o recurso.
Contribuidor/colaborador	Entidade responsável por fazer contribuições para o recurso.
Data	Ponto ou período associado a um evento no ciclo de vida do recurso.
Tipo	A natureza ou gênero do recurso.

Formato	O formato do arquivo, meio físico ou dimensões do recurso.
Identificador	Referência inequívoca ao recurso dentro de determinado contexto.
Fonte	Recurso do qual o recurso descrito é derivado.
Idioma	Idioma do recurso.
Relação	Recurso relacionado.
Abrangência/Cobertura	O tópico espacial ou temporal do recurso, a aplicabilidade espacial do recurso ou a jurisdição sob a qual o recurso é relevante.
Licenças/Direitos autorais	Informações sobre direitos mantidos dentro e sobre o recurso.

Fonte: Dublin Core, 2012.

Dada essa relação entre os elementos principais do Dublin Core e os metadados comuns entre os acervos analisados, foi definido que o padrão de metadados utilizado para agregação dos acervos nesta pesquisa será o conjunto de 15 elementos principais do Dublin Core.

5.2.3.2 Mapeamento dos metadados para o padrão de agregação

Depois de definido o modelo de metadados para agregação como sendo o conjunto de elementos principais do Dublin Core, assim como descrito na metodologia (tópico 4.2.2),um conjunto de pesquisadores realizou o processo de equivalência entre os metadados coletados e os metadados do Dublin Core. Essa fase do mapeamento já foi executada sobre o quantitativo geral de dados coletados de todos os acervos passíveis de coleta, e não mais sobre o recorte de 16 acervos.

Esse mapeamento da equivalência dos metadados dos acervos com os metadados do Dublin Core é denominado *crosswalk* pela literatura técnica da Ciência da Informação, em que partindo de um modelo de dados se mapeia para outro modelo. Após a identificação das equivalências entre os metadados, o processo de união dos dados sob um único padrão de metadados foi implementado por meio de um programa em Python (https://github.com/tainacan/data_science/blob/master/FAPESP/croswallk.py).

Para esse mapeamento, alguns procedimentos técnicos são adotados (CHAN; ZENG, 2006a; CHAN; ZENG, 2006b; HASLHOFER; KLAS, 2010). Alguns metadados diferentes na origem são agregados e mapeados para um mesmo metadado no destino. Outros metadados, por não terem nenhuma correspondência com o padrão de destino, são descartados e não farão parte do modelo final.

Essa é a etapa que representa perda de informação, porém necessária para a manutenção da interoperabilidade semântica entre um modelo comum de todos os acervos. No total, foram identificados um total de 587 metadados na pesquisa, sendo que 227 (38,7%) foram mapeados e 360 (61,33%) foram descartados.

Na Tabela 7, apresentamos o total de metadados que foram mapeados para cada um dos campos do padrão Dublin Core simplificado. Por exemplo, 105 metadados foram mapeados para o metadado "Título" do padrão Dublin Core, 57 para o metadado "Descrição" e assim por diante. Essa tabela mostra quais são os campos em que mais informação foi possível se obter, e aqueles nos quais menos informações foram encontradas.

Tabela 7 - Número de metadados mapeados para os campos do padrão Dublin Core

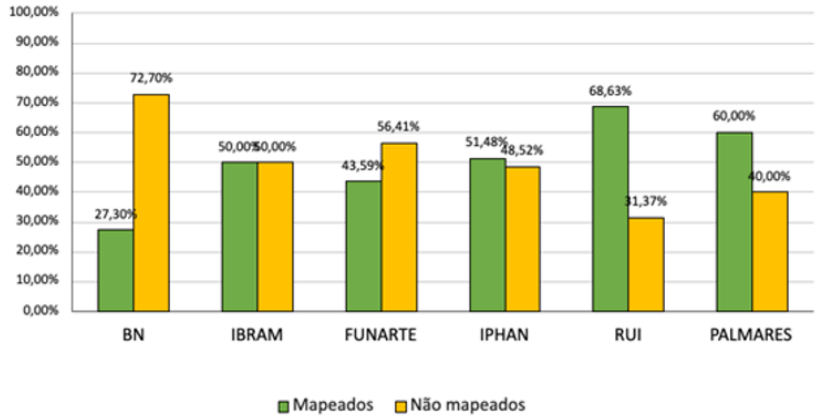
Metadado Dublin Core	Nº de Metadados	Proporção (%)
Título	105	46,56%
Descrição	57	25,11%
Assunto/palavras-chave	54	23,79%
Abrangência/cobertura	47	20,70%
Identificador	40	17,62%
Data	34	14,98%
Autor	19	8,37%
Colaborador	16	7,05%
Fonte	11	4,85%
Tipo	9	3,96%
Direitos	7	3,08%
Relação	6	2,64%
Editor	5	2,20%
Idioma	5	2,20%
Formato	4	1,76%

Fonte: Elaboração própria, 2023.

Alguns casos chamam atenção, como o metadado "Direitos", em que foram mapeados apenas sete metadados, mostrando como a declaração de direitos sobre os objetos digitais não é algo expressamente declarado nos acervos digitais encontrados.

Também foi possível analisar os resultados do mapeamento em relação a cada instituição vinculada, permitindo identificar aquelas onde foi possível obter melhor cobertura de mapeamento de metadados e aquelas onde foi mais difícil encontrar correspondência com o modelo Dublin Core. O resultado pode ser visto na Figura 68.

Figura 68 - Mapeamento relativo dos metadados por entidade cultural vinculada



Fonte: Elaboração própria, 2023.

Observa-se na Figura 68 que a BN é a instituição que tem maior impacto em termos de perda de informação no processo de mapeamento, e a Fundação Casa de Rui Barbosa a que tem o melhor desempenho. Esse resultado se deve ao fato da BN utilizar o modelo de dados MARC, que apresenta um conjunto de metadados bem mais abrangente e que apresenta perdas importantes quando mapeado para o Dublin Core.

Já no caso da Fundação Casa de Rui Barbosa, o repositório RUBI já apresentava metadados no modelo Dublin Core, e tal fato facilitou bastante o reúso dos metadados e sua maior cobertura nos resultados obtidos. Nas demais instituições, observam-se resultados intermediários entre esses relatados acima, no entanto a perda de metadados nunca é menor que 30% em todos os casos.

Pode-se afirmar, portanto, que a ausência da adoção de modelo de dados, comum entre as instituições de forma deliberada e planejada, leva a expressiva perda de informação quando se tenta mapear os modelos de dados para um padrão comum, tal como o Dublin Core no caso da presente pesquisa.

Tal resultado, certamente, vai impactar nas possibilidades de busca e recuperação da informação para o usuário final, levando a perda de expressividade e, potencialmente, maior dificuldade de encontrar e recuperar dados.

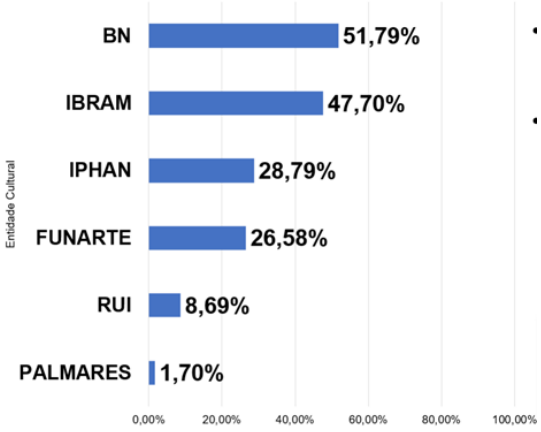
5.2.3 Publicação dos acervos agregados

Para a publicação dos acervos agregados, já unificados sob o padrão de metadados do Dublin Core, o universo de dados publicados é resultante da execução

dos programas de coleta para todos os links de acesso aos acervos identificados na etapa de caracterização desta pesquisa (tópico 5.1.1).

O resultado da coleta gerou 41 arquivos que foram armazenados em formato CSV, totalizando 396.557 registros coletados. A distribuição relativa da quantidade de dados coletados pode ser vista na Figura 69.

Figura 69 - Distribuição dos dados coletados por instituição cultural vinculada.



Fonte: Elaboração própria, 2023.

Cabe ressaltar que se esperava um total maior de registros coletados. No entanto, o período de realização da coleta dos dados foi concomitante a um problema de ataque hacker no site da Biblioteca Nacional, sendo que o site e todos os serviços digitais da BN ficaram fora do ar por várias semanas e apenas retornaram gradualmente, não estando todos disponíveis para coleta no cronograma disponível para a realização da presente etapa. Além da Biblioteca Nacional, os dados da ANCINE também estavam fora do ar no momento da coleta e ficaram de fora da base final. Optou-se, portanto, para não prejudicar o restante da pesquisa que seguiu apenas com os dados já coletados.

5.2.3.1 Armazenamento e indexação dos dados

Para armazenamento e indexação dos dados resultantes do processo de mapeamento foi adotada a ferramenta Elastic Search. A ferramenta é preparada para trabalhar com dados massivos e realizar busca em escala, sendo capaz de lidar com dados da ordem de bilhões de registros.

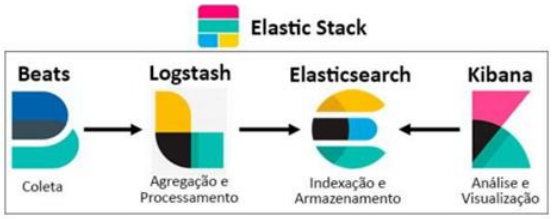
A ferramenta foi considerada adequada para o projeto devido ao volume de dados coletados, da ordem de mais de 300.000 registros, e preparada para integrar

com outros soluções que facilitam a automatização futura de processos de coleta de dados, tais como FileBeats e LogStash, além da possibilidade de automação e geração de painéis analíticos para visualização de dados, como o Kibana. (ELASTIC, 2023)

Ou seja, o Elastic Search faz parte de uma ecologia de soluções livres que constituem camadas de serviços de dados que podem ser exploradas em etapas futuras do projeto, ampliando o potencial de resultados da pesquisa. Um exemplo que corrobora esse fator é seu uso efetivo no serviço de agregação Brasiliana Museus. (Siqueira; Martins; Lemos, 2022; Siqueira; Martins, 2021)

Além disso, considerando a baixa expressividade semântica dos dados, optou-se por explorar melhor o potencial analítico descritivo dos dados, o que pôde ser mais bem trabalhado com os serviços ligados ao Elastic Search. As soluções fornecidas pela Elastic podem ser integradas da maneira proposta na Figura 70.

Figura 70 - Proposta de integração das soluções da ElasticStack



Fonte: Elaboração própria, 2023.

Para o presente projeto, cabe dizer que foi utilizada apenas a camada ElasticSearch, pelo fato de as etapas de coleta e agregação de dados serem feitas utilizando programas em Python. No entanto, vale a reflexão, de que se as instituições culturais adotassem repositórios digitais para seus acervos culturais, seria possível automatizar todo o processo de coleta e usar as soluções da ElasticStack, como a FileBeats e Logstash, para realizar a coleta, agregação e processamento em um único fluxo de tratamento de dados.

Entende-se que a compreensão desse modelo é um importante resultado da pesquisa, pois sinaliza para um possível modelo automatizado de agregação de dados a partir de adoção de modernos repositórios digitais pelas instituições culturais. Além disso, obviamente, a adoção de boas práticas de catalogação e modelagem de dados se faz fundamental para a boa interoperabilidade entre os acervos.

5.2.3.2 Repositório dos acervos agregados

Conforme já mencionado no tópico 4.2.1.4 desta pesquisa, o software de repositório digital escolhido para a publicação dos dados foi o Tainacan, por três motivos: 1 - é um *software* difundido entre os museus brasileiros, e utilizado no caso da iniciativa nacional da Brasiliana Museus; 2 – é um *software* que fornece o meio de acesso ao repositório através de um site, já que foi desenvolvido na mesma estrutura do sistema de gerenciamento de conteúdo WordPress; 3 – é o *software* idealizado, desenvolvido e atualmente mantido pela equipe do Laboratório de Inteligência de Redes (UnB), da qual faço parte.

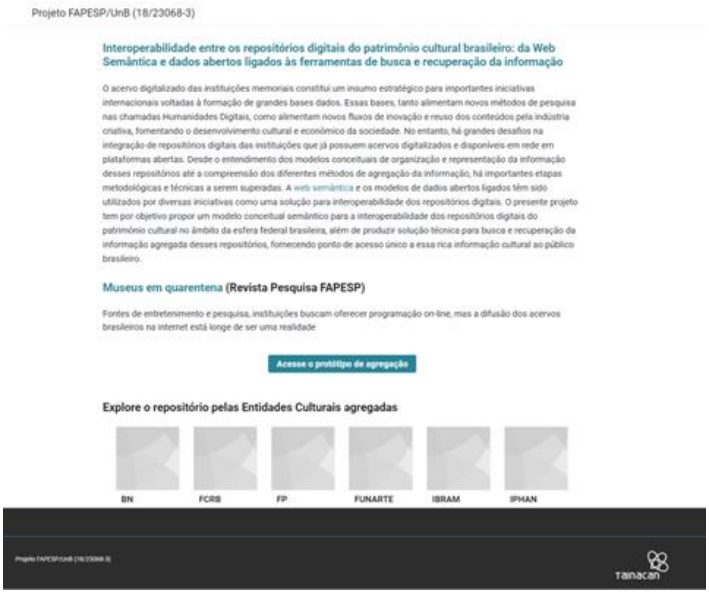
Outra questão importante é que, pelo fato de o Tainacan ser um software construído sob a estrutura do WordPress, a conexão com a ferramenta Elastic Search foi facilitada através do plugin Elastic Press, permitindo aproveitar todos os benefícios desta ferramenta como mencionado no tópico anterior (5.2.3.1).

O repositório digital do projeto está disponível em: https://culturabr.tainacan.org/ (acesso alternativo: https://rosaluis.com/tese_fapesp_unb/). O repositório foi implementado utilizando o software WordPress e o plugin Tainacan para as funcionalidades de organização e representação da informação.

O repositório possui conexão direta com os dados armazenados no Elastic Search e utiliza o Tainacan apenas como uma interface de visualização e navegação nos dados. Dessa forma, tem-se uma combinação de soluções, um banco de dados NoSQL para processamento de dados massivos e uma interface interativa para visualização e exploração dos dados. A ferramenta e suas características são apresentadas nas figuras a seguir.

A tela inicial de entrada do repositório encontra-se na Figura 71. São apresentados os dados gerais do projeto, seu número institucional e uma matéria da revista FAPESP que aborda o projeto logo no início da quarentena relacionada a pandemia do coronavírus. Logo abaixo, são apresentadas as coleções digitais que foram agregadas no projeto. As coleções são denominadas pelo nome da instituição de cultura vinculada de forma a facilitar exploração do acervo por meio de sua origem.

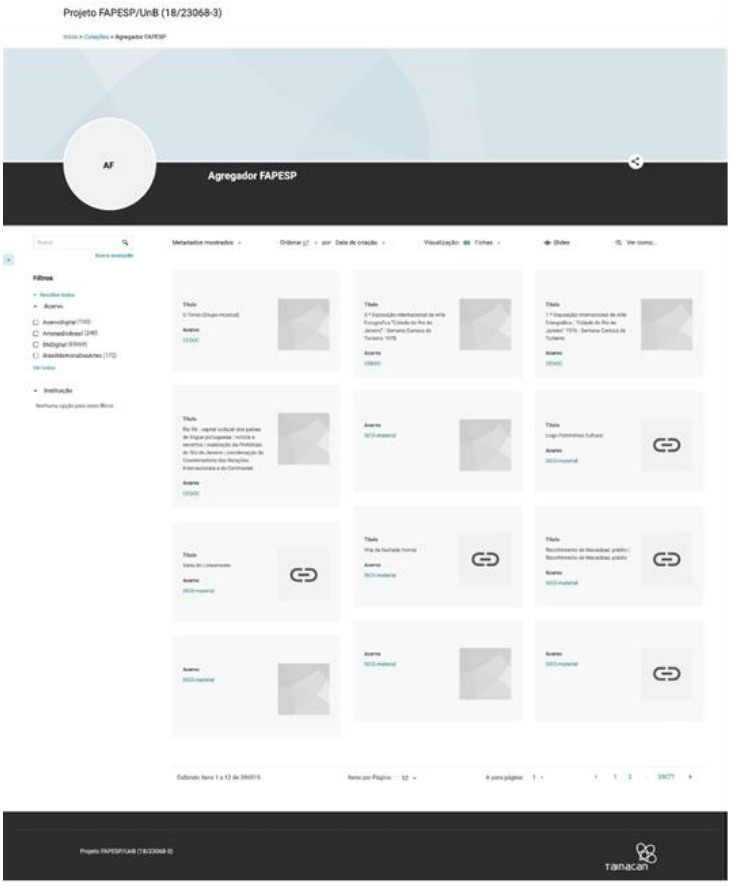
Figura 71 - Tela inicial do repositório digital



Fonte: Repositório Digital Tainacan, 2021.

Na Figura 72, apresenta-se a tela de entrada da interface de exploração dos dados. Algumas áreas são destacadas para se observar detalhes da implementação. Na parte esquerda da imagem, pode-se visualizar as facetas de exploração dos dados, sendo o nome das coleções encontradas e o nome da instituição de cultura vinculada. No centro da imagem, aparem os metadados dos objetos culturais coletados. Embaixo da imagem, pode-se observar a paginação dos dados, mostrando os mais de 300.000 registros catalogados.

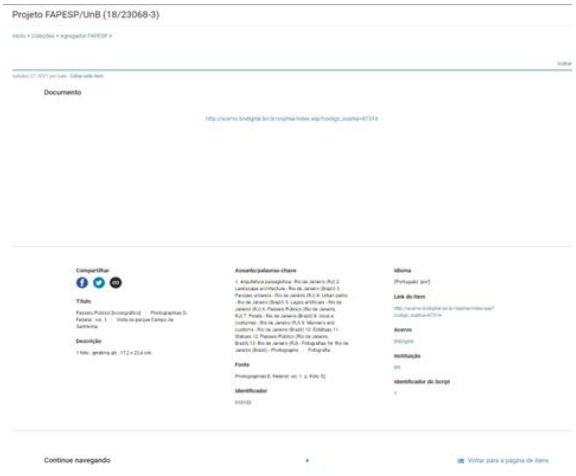
Figura 72 - Tela inicial da interface de exploração dos dados



Fonte: Repositório Digital Tainacan, 2021.

Na Figura 73, apresenta-se a página de um item. É possível observar todos os metadados do item já padronizados no modelo Dublin Core e como documento consta o link de origem do objeto cultural em sua fonte de informação inicial.

Figura 73 - Página de um objeto cultural no repositório digital



Fonte: Repositório Digital Tainacan, 2021.

6 CONCLUSÕES

Ao recuperar o problema que esta pesquisa se propôs a responder, que trata-se de “Como estão disponibilizados os acervos digitais publicados na internet pelas instituições vinculadas ao Ministério da Cultura do Brasil, quanto às suas características tecnológicas e de organização do conhecimento e representação da informação, diante da possibilidade de implementação de um serviço de agregação destes acervos”, entende-se que o questionamento foi satisfatoriamente respondido por meio dos resultados aqui apresentados. As características tecnológicas e de organização do conhecimento e representação da informação dos acervos digitais das entidades culturais estão descritas no tópico 5.1 dos resultados, e a comparação dessa caracterização diante de um contexto de possibilidade de implementação de um serviço de agregação desses acervos foi exemplificada através do desenvolvimento do protótipo de agregação, apresentado no tópico 5.2.

Em relação à hipótese inicial da pesquisa, que partiu do entendimento de que “promover a implementação de um serviço de agregação desses acervos heterogêneos do modo em que estão publicados atualmente, do ponto de vista da manutenção e reprodução, de mecanismos tecnológicos e de modelos de organização da informação para garantir a integração dos dados dos objetos digitais, exigirá um conjunto complexo de ações e poderá ter como resultado um modelo e formato pouco sustentável e escalável”, conclui-se a sua corroboração a partir dos resultados alcançados. Destaca-se que tanto no quesito da caracterização dos acervos, que apontou a condição da heterogeneidade em vários aspectos e, também por este sentido, prospectou a dificuldade de implementação de um serviço de agregação, quanto no desenvolvimento do protótipo de agregação, que foi desenvolvido, em sua maioria, a partir de programas de raspagem de dados com alta complexidade e capacidade mínima de sustentabilidade do serviço de agregação, é que chegando-se ao entendimento da validação da hipótese apresentada.

Dessa forma, nos tópicos abaixo serão apresentadas reflexões gerais sobre o estudo realizado nesta pesquisa, bem como serão apontados o que se entende como alcances e limitações identificadas ao longo do desenvolvimento deste trabalho.

6.1 Reflexão geral

Diante das etapas percorridas até aqui, entende-se que, de forma geral, a documentação encontrada nos repositórios das entidades estudadas se mostrou de baixa padronização, o que acabou por gerar um ambiente com poucas possibilidades de facetas para exploração dos acervos, sendo utilizadas facetas criadas no âmbito desta pesquisa: o nome da coleção e o nome da instituição cultural. Nesse contexto, entende-se que o usuário final terá mais dificuldade de explorar esse acervo sem uma busca específica, e tal fato demonstra a ausência de linguagens documentárias entre as coleções, impossibilitando padronizar palavras-chave, taxonomias ou qualquer tipo de vocabulário controlado, seja de denominação de objetos, entidades ou descrição temática do objeto cultural.

Quanto à caracterização dos acervos digitais culturais brasileiros, vale retomar a reflexão sobre a importância de se deixar explícita a documentação sobre licenças direitos autorais, principalmente em um contexto em que os objetos culturais podem assumir autorias tão historicamente e etnicamente heterogêneas.

Outro aspecto importante é quanto às funções dos *softwares* para o armazenamento e publicação dos acervos. É preciso esclarecer quais as possibilidades da ferramenta utilizada quanto à permissão de coleta automatizada dos dados do acervo. Em alguns casos, como o uso de *softwares* proprietários, é necessário entender se há necessidade de adquirir ou configurar módulos extras que permitam essa função.

Apesar do contexto apresentado acima, os resultados aqui alcançados apresentam importantes funcionalidades de busca e cruzamento de dados que seriam impossíveis de serem realizados em um mesmo sistema de informação sem o processo técnico proposto na presente pesquisa.

Esta pesquisa demonstra, ao menos em potencial, que a agregação dos acervos é possível e pode ser profundamente melhorada com a adoção de melhores práticas de gestão da informação. A demonstração da possibilidade mostra que, mesmo com o precário cenário atual encontrado e descrito na pesquisa, é possível coletar dados, identificar possibilidades mínimas de interoperabilidade e produzir um repositório digital para agregar a documentação resultando das etapas técnicas aqui propostas.

Considerando a adoção de melhores práticas de organização do conhecimento e representação da informação, a saber, o uso de linguagens documentárias em comum, a definição e uso comum de um modelo expressivo de metadados, a adoção de práticas comuns de catalogação e a adoção de repositórios digitais para os acervos culturais, tal agregação não apenas poderia expressar dados de melhor qualidade, com maiores possibilidades de exploração e cruzamento, como também automatizar o processo de coleta e manutenção dos registros já existentes.

Da perspectiva técnica, o uso de tecnologias atuais de indexação e recuperação de informação, como o Elastic Stack, que é utilizado na iniciativa nacional Brasiliana Museus, é um indicativo favorável à implementação nos demais casos com potencial de agregação. Já da perspectiva semântica, o uso do Dublin Core como um conjunto inicial de elementos para mapear os metadados dos acervos e agregá-los, também é um apontamento favorável na direção da agregação dos acervos digitais culturais brasileiros.

Além dessas reflexões conceituais e práticas sobre a agregação de acervos digitais culturais, vale incitar uma reflexão posterior aos alcances dessa pesquisa, que diz respeito a condição de articulação entre as instituições de cultura brasileira e sua possibilidade de acordos, baseados no avanço de políticas nacionais de cultura e digitalização de acervos, para promover uma gestão sustentável de iniciativas de agregação, como apresentam os exemplos de agregadores culturais de outras nações (tópico 3).

Assim, pode-se concluir, que esta pesquisa demonstra seu objetivo principal, a identificação e descrição das condições informacionais e tecnológicas necessárias para a agregação de acervos digitais culturais do Brasil, conforme delimitados pelo escopo deste estudo.

6.2 Alcances e limitações

É importante delimitar a abrangência da capacidade que o material desta pesquisa tem como referência para o contexto científico brasileiro da agregação de acervos digitais culturais. Abaixo são listados o que se entende por elementos que foram alcançados na produção deste estudo, como também a indicação de elementos entendidos como limitações e que não puderam ser contemplados por esta pesquisa.

6.2.1 Alcances

- **Caracterização dos acervos digitais das entidades culturais brasileiras:** entende-se que a caracterização dos acervos digitais das entidades culturais apresentada nesta pesquisa é um referencial técnico importante, tanto para a tomada de decisões no contexto da cultura digital brasileira, quanto para futuras pesquisas da área que utilizem das informações para explorar em estudos científicos os resultados evidenciados de outras maneiras.
- **Metodologia aplicada:** entende-se que a metodologia aplicada nesta pesquisa, o estudo de casos múltiplos complementado pela análise de conteúdo categorial e pelas técnicas de estatística descritivas, está devidamente detalhada, de modo a possibilitar que outras pesquisas da área possam se fundamentar no estudo realizado e, até mesmo o replique, se necessário.
- **Modelo de estágios de agregação utilizado:** entende-se que o modelo de estágios de agregação dos acervos digitais, produzidos através de referencial teórico e pré-análise dos casos de uso, servirá como um *framework* a ser estudado, reutilizado e até melhorado em pesquisas que tenham objetivos semelhantes a esta.
- **Protótipo de agregação dos acervos das entidades culturais:** entende-se que o protótipo desenvolvido pode servir como insumo às iniciativas brasileiras que desejem investir no desenvolvimento de um serviço de agregação dos acervos digitais das entidades culturais brasileiras. As ferramentas utilizadas e o processo de encadeamento dos estágios de agregação até a publicação dos acervos agregados em um repositório digital, podem servir de base para um serviço mais completo e mais eficiente, se alinhado com as articulações entre os entes culturais brasileiros.
- **Mapeamento dos metadados:** entende-se que o resultado que o mapeamento de metadados aponta, mesmo com suas limitações para este estudo, é importante para indicar a necessidade de atenção especial à forma como os aspectos de organização do conhecimento e representação da informação são abordados nos acervos em questão. O resultado da semelhança entre os elementos comuns dos acervos com o esquema principal de elementos do Dublin Core é importante para nortear possível modelo semântico de

agregação mais completo, porém, ainda se faz necessário um trabalho holístico que envolva várias frentes e ações na perspectiva de implementar esse modelo semântico.

- **Visualização das relações dos metadados a partir de grafos:** entende-se que o uso da visualização de grafos para representação visual das relações entre acervos digitais a partir de seus metadados é uma técnica efetiva para apresentação desse tipo de aplicação em trabalhos científicos. Dessa forma, espera-se que essa técnica seja difundida, para que seu uso aumente as condições de interpretação de outras pesquisas que tenham como objetivo a análise de relações informacionais.

6.2.2 Limitações

- **Coleta dos dados dos acervos digitais das entidades culturais:** como foi possível observar nos resultados apresentados na pesquisa, houve dificuldades no processo de coleta dos dados dos acervos da ordem temporal e técnica. Esta pesquisa foi desenvolvida em meio à pandemia do coronavírus, o que dificultou seu desenvolvimento pela necessidade de adaptação às novas circunstâncias de trabalho, sendo este um dos fatores que incidiram no uso de recortes de acervos para análise, sendo possível utilizar os dados completos somente depois da execução total da coleta. De ordem técnica, podemos citar a falha nos acessos aos sites dos acervos, pelos motivos já citados anteriormente. Dessa forma, a coleta dos dados não representou a totalidade de registros mapeados na etapa de caracterização dos acervos, mas entende-se que isso não comprometeu a execução da pesquisa ou seus resultados.
- **Mapeamento dos metadados:** O mapeamento dos metadados executado foi de cunho sintático. Houve a participação de pesquisadores no processo de mapeamento que atuaram de forma qualitativa na interpretação das reconciliações com os elementos do Dublin Core, mas não houve um aprofundamento semântico nesse processo que levasse em conta os valores e suas relações com sistemas de organização do conhecimento, por exemplo. Isso se deu em grande parte pelas circunstâncias apresentadas pelo estudo, uma vez que a documentação sobre os acervos era limitada, e esse tipo de abordagem normalmente deve envolver de forma ativa a participação das

entidades culturais, o que não fez parte do escopo desta pesquisa. Dessa maneira, entende-se que dentro do contexto de produção desta pesquisa essa limitação não comprometeu a execução do estudo ou seus resultados.

- **Modelo de estágios de agregação utilizado:** o modelo de estágios de agregação não incluiu o estágio de enriquecimento dos dados coletados dos acervos, como apresentaram algumas iniciativas identificadas na etapa de revisão da pesquisa. Essa etapa de enriquecimento é o processo em que os dados são relacionados com outros dados de outras fontes, em busca de aumentar o contexto informacional dos objetos dos acervos. O motivo pela não inclusão desta etapa é novamente pelas limitações de aprofundamento semântico da pesquisa, uma vez que não foi possível, pelos motivos citados no parágrafo anterior, processar de forma semântica os valores dos metadados mapeados. Porém, a importância desta etapa foi ressaltada nos tópicos de pertinência do estudo. Assim, entende-se que dentro do contexto de produção desta pesquisa, essa limitação não comprometeu a execução do estudo ou seus resultados.

REFERÊNCIAS

ABADAL, E.; LLUIS, C**. Bases de datos documentales:** características, funciones y método. Síntesis, 2005.

ABBAS, J. **Structures for organizing knowledge:** Exploring taxonomies, ontologies, and other schema. Neal-Schuman. 2010.

ADAMSKI, K.; ALKHAEIR, T.; HELIŃSKI, M.; NOWAK, A.; WERLA, M.; WOŹNIAK, P. **Europeana DSI 2–Access to Digital Resources of European Heritage**. 2017. Disponível em: https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_DSI-2/Milestones/ms7.2-infrastructure-layer.pdf. Acesso em: 20 ago. 2023.

ALMEIDA, M. B. Revisiting ontologies: A necessary clarification. **Journal of the American Society for Information Science and Technology**, v. 64, n. 8, p. 1682-1693, 2013.

ALMEIDA, M.; SOUZA, R.; FONSECA, F. Semantics in the semantic web: A critical evaluation. **Knowledge Organization**, v. 38, n. 3, p. 187–203, 2011.

ALVARENGA, L. Representação do conhecimento na perspectiva da ciência da informação em tempo e espaço digitais. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, n. 15, 2003.

ATKINSON, R. Library functions, scholarly communication, and the foundation of the digital library: Laying claim to the control zone. **The library quarterly**, v. 66, n. 3, p. 239-265, 1996.

ARARIPE, F. M. A. Do patrimônio cultural e seus significados. **Transinformação**, v. 16, p. 111-122, 2004.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval:** The concepts and technology behind search. Addison Wesley, 2011.

BARDIN, L.; **Análise de Conteúdo**. Lisboa: Edições 70, 1977.

BARITE, M. G. The notion of “category”: Its implications in subject analysis and in the construction and evaluation of indexing languages. **Knowledge Organization**, v. 27 n. 1–2, p. 4–10, 2000.

BERNERS-LEE, T. **Linked Open Data 5 Star Scheme for assessing the stages of open data deployment and use.** 2009. Disponível em: https://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 25 jul. 2023.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data—the story so far. **International Journal on Semantic Web and Information Systems**, v. 5, n. 3, p. 1–22, 2009.

BONDY, J. A.; MURTY, U. S. R. **Graph theory**. Springer Publishing Company, Incorporated, 2008.

BORGMAN, C. L. **Scholarship in the digital age:** Information, infrastructure, and the Internet. MIT press, 2010.

BORKO, H. Information science: what is it?. **American documentation**, v. 19, n. 1, p. 3-5, 1968.

BRASILIANA MUSEUS. **Página Inicial**. 2023a. Disponível em: https://brasiliana.museus.gov.br/. Acesso em: 15 set. 2023.

BRASILIANA MUSEUS. **Sobre a Brasiliana Museus**. 2023b. Disponível em: https://brasiliana.museus.gov.br/. Acesso em: 15 set. 2023.

CATALANO, C. E.; VASSALLO, V.; HERMON, S.; SPAGNUOLO, M. Representing quantitative documentation of 3D cultural heritage artefacts with CIDOC CRMdig. **International Journal on Digital Libraries**, v. 21, n. 4, p. 359–373, 2020.

CGI. **Pesquisa sobre o uso das Tecnologias de Informação e Comunicação nos equipamentos culturais brasileiros - TIC Cultura 2020**, São Paulo, 2021.

CHAN, L. M.; ZENG, M. L. Metadata interoperability and standardization–a study of methodology part I. **D-Lib magazine**, v. 12, n. 6, p. 1082-9873, 2006a.

CHAN, L. M.; ZENG, M. L. Metadata interoperability and standardization–a study of methodology part II. **D-Lib magazine**, v. 12, n. 6, p. 1082-9873, 2006b.

CHERVEN, K. **Mastering Gephi network visualization**. Packt Publishing Ltd, 2015.

CLEMENT, T. E.; CARTER, D. Connecting theory and practice in digital humanities information work. **Journal of the Association for Information Science and Technology**, v. 68, n. 6, p. 1385-1396, 2017.

CLINK, K. DPLA: Digital Public Library of America. **Reference Reviews**, v. 28, n. 3, p. 54-55, 2014.

COLLA, D. *et al*. Bringing semantics into historical archives with computer-aided rich metadata generation. **Journal on Computing and Cultural Heritage**, v. 15, n. 3, p. 1-24, 2022.

CONCORDIA, C. Integration of heterogeneous metadata in Europeana. *In*: **LIDA Workshop:** Issues and Challenges with the Implementation of Metadata Schemes, Zadar, Croatia. 2009. Disponível em: https://www.dublincore.org/groups/tools/docs/LIDA09WorkshopC_1.pdf. Acesso em: 15 ago. 2023.

DAHLBERG, I. A referent-oriented, analytical concept theory for INTERCONCEPT. **International Classification**, v. 5, n. 3, p. 142–151, 1978.

DAHLBERG, I. Knowledge organization: its scope and possibilities. **Knowledge Organization**, v. 20, n. 4, p. 211-222, 1993.

DEKKERS, M.; GRADMANN, S.; MEGHINI, C. **Europeana Outline Functional Specification: for development of an operational european digital library.** 2008. Disponível em: https://www.academia.edu/22251368/Europeana_Outline_Functional_Specification_For_development_of_an_operational_European_Digital_Library_Europeana_Outline_Functional_Specification. Acesso em: 20 ago. 2023.

DIJKSHOORN, C. *et al*. The Rijksmuseum collection as linked data. **Semantic Web**, v. 9, n. 2, p. 221-230, 2018.

DOERR, M. *et al.* A Repository for 3D Model Production and Interpretation in Culture and Beyond. In: **VAST**, 2010a.

DOERR, M. *et al*. The europeana data model (edm). *In*: **World Library and Information Congress:** 76th IFLA general conference and assembly. 2010b.

DPLA. **Introduction to the DPLA Metadata Application Profile, version 5.0.** 2017. Disponível em: https://drive.google.com/file/d/1kMxXgFrGwu3i7LBLFOj6VZRQFuQzqkHk/view. Acesso em: 15 jun. 2023.

DPLA. **DPLA Github repositories.** 2023a. Disponível em: https://github.com/dpla. Acesso em: 22 ago. 2023.

DPLA. **Ingestion 3.** 2023b. Disponível em: https://github.com/dpla/ingestion3. Acesso em: 22 ago. 2023.

DPLA. **Ingestion 3 harvest.** 2023c. Disponível em: https://github.com/dpla/ingestion3#harvest . Acesso em: 22 ago. 2023.

DPLA. **Ingestion 3 mapping and validation.** 2023d. Disponível em: https://github.com/dpla/ingestion3#mapping-and-validation. Acesso em: 22 ago. 2023.

DPLA. **Metadata apllication profile.** 2023e. Disponível em: https://pro.dp.la/hubs/metadata-application-profile . Acesso em: 22 ago. 2023.

DPLA. **Ingestion 3 Enrichment and normalization.** 2023f. Disponível em: https://github.com/dpla/ingestion3#enrichment-and-normalization . Acesso em: 22 ago. 2023.

DPLA. **Wikimedia.** 2023g. Disponível em: https://github.com/dpla/ingestion3#wikimedia . Acesso em: 22 ago. 2023.

DPLA. **DPLA front-end.** 2023h. Disponível em: https://github.com/dpla/dpla-frontend. Acesso em: 22 ago. 2023.

DPLA. **API.** 2023i. Disponível em: https://github.com/dpla/api. Acesso em: 22 ago. 2023.

DPLA. **Dashboard analytics.** 2023j. Disponível em: https://github.com/dpla/dashboard-analytics . Acesso em: 22 ago. 2023.

EUROPEANA TECH. **AI in relation to GLAMs Task Force:** Report and recommendations. 2021. Disponível em: https://pro.europeana.eu/files/Europeana_Professional/Europeana_Network/Europeana_Network_Task_Forces/Final_reports/AI%20in%20relation%20to%20GLAMs%20Task%20Force%20Report.pdf. Acesso em: 05 jan. 2023.

EUROPEANA. **Share your data – process**. 2023. Disponível em: https://pro.europeana.eu/share-your-data/process. Acesso em: ago. 2023.

EUROPEANA PRO. **Europeana SPARQL API**. 2023. Disponível em: https://pro.europeana.eu/page/sparql. Acesso em: 22 ago. 2023.

FARIAS, A. M. L. **Estatística descritiva**. Universidade Federal Fluminense. Instituto de Matemática, 2020.

FERREZ, H. D. **Tesauro de objetos do patrimônio cultural nos museus brasileiros.** Rio de Janeiro: Fazer Arte. Gerência de Museus da Secretaria Municipal de Cultura, 2016.

FERREZ, H. D.; BIANCHINI, M. H. S. **Thesaurus para acervos museológicos.** Ministério da Cultura, Secretaria do Patrimônio Histórico e Artístico Nacional, Fundação Nacional Pró-Memória, Coordenadoria Geral de Acervos Museológicos, 1987.

FOSKETT, A. C. **A abordagem temática da informação**. Briquet de Lemos, 1973.

FREIRE, K. M. W.; SALES, L. F.; SAYÃO, L. F. Curadoria digital no contexto artístico e cultural: possibilidades de reúso de dados de arte. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 25, p. 01-21, 2020.

GILLILAND, A. J. Setting the stage. In: BACA, M. (Ed.), **Introduction to metadata**. Getty Publication. 2016.

GREENBERG, J. Big metadata, smart metadata, and metadata capital: Toward greater synergy between data science and metadata. **Journal of Data and Information Science**, v. 2, n. 3, p.19–36, 2017.

GUIZZARDI, G. Ontology, ontologies, and the “I” of FAIR. **Data Intelligence**, v. 2 n. 1–2, p. 181–191, 2020.

HARPRING, P. **Metadata Standards Crosswalk.** 2022. Disponível em: https://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html. Acesso em: 15 dez. 2023.

HASLHOFER, B.; KLAS, W. A survey of techniques for achieving metadata interoperability. **ACM Computing Surveys (CSUR)**, v. 42, n. 2, p. 1-37, 2010.

HISPAÑA. **Presentación**. 2023. Disponível em: https://hispana.mcu.es/es/contenido/presentacion.do. Acesso em: 26 ago. 2023.

HJORLAND, B. Does the traditional thesaurus have a place in modern information retrieval? **Knowledge Organization**, v. 43 n. 3, p. 145–159, 2016.

HJORLAND, B. Fundamentals of knowledge organization. **Knowledge organization**, v. 30, n. 2, p. 87-111, 2003.

HJORLAND, B. Semantics and knowledge organization. **Annual Review of Information Science and Technology**, v. 41, p. 367-405, 2007.

HJORLAND, B. Theories are knowledge organizing systems (KOS). **Knowledge Organization**, v. 42, n. 2, p. 113–128, 2015.

HODGE, G. **Systems of knowledge organization for digital libraries:** Beyond traditional authority files. Digital Library Federation, Council on Library and Information Resources, 2020.

HOLLEY, R. Trove: Innovation in access to information in Australia. **Ariadne**, n. 64, 2010.

HYVONEN, E. **Publishing and using cultural heritage linked data on the semantic web**. Springer Nature, 2022.

HYVÖNEN, E.; HEINO, E.; LESKINEN, P.; IKKALA, E.; KOHO, M.; TAMPER, M.; TUOMINEN, J.; MÄKELÄ, E. WarSampo data service and semantic portal for publishing linked open data about the second world war history. *In*: **European semantic web conference.** Springer, p. 758–773, 2016.

IFLA (International Federation of Library Associations and Institutions). **Functional requirements for bibliographic records. Study Group on the Functional Requirements for Bibliographic Records**. 2009. Disponível em: https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf. Acesso em: 20 jun. 2023.

IFLA (International Federation of Library Associations and Institutions). **Statement of international cataloguing principles**. 2016. Disponível em: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/icp/icp_2016-en.pdf. Acesso em: 12 jun. 2023.

IFLA (International Federation of Library Associations and Institutions). **Statement on libraries and artificial intelligence**. 2020. Disponível em: https://repository.ifla.org/handle/123456789/1646. Acesso em: 16 jun. 2023.

JACOB, E. K. Classification and categorization: A difference that makes a difference. **Library Trends**, v. 52, n. 3, p. 515–540, 2004

JÚNIOR, A. C; LEMOS, D. L. Tratamento da informação em acervos culturais: avaliação do uso de vocabulários controlados em coleções museológicas sob gestão do Instituto Brasileiro de Museus. **Revista Ibero-Americana de Ciência da Informação**, v. 16, n. 1, p. 131-145, 2023.

KHOO, C. S. G.; NA, J.-C. Semantic relations in information science. **Annual Review of Information Science and Technology**, v. 40, n. 1, p. 157–228, 2006.

KOLTAY, T. 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.

LANCASTER, F. W. **Vocabulary control for information retrieval**. Information Resources Press, 1986.

LEFFERTS, M.; CIOCOIU, A.; MUHR, M.; GASIA, A.; DUNNING, A.; **Europeana Cloud D4.2 - Content Ingestion Plan**. 2015. Disponível em: https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Cloud/Deliverables/D4.2%20Content%20Ingestion%20Plan.pdf Acesso em: 20 ago. 2023.

LEMOS, D. L. S., MARTINS, D. L., CARMO, D. Quality standards for data and metadata addressed to data science applications. **Advanced Notes in Information Science**, v. 2, p. 161-170, 2022.

LEMOS, D. L.; SOUZA, R. R. 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.

LIMA, J.L.O.; ALVARES, L. Organização e representação da informação e do conhecimento. **Organização da informação e do conhecimento:** conceitos, subsídios interdisciplinares e aplicações. São Paulo: B4 Editores, v. 248, p. 21-48, 2012.

LIU, A. 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.

LUHN, H. P. A new method of recording and searching information. **American Documentation**, v. 4, n. 1, p. 14–16, 1953.

MACHADO, L. M. O.; SOUZA, R. R.; SIMÕES, M.G. Semantic web or web of data? A diachronic study (1999 to 2017) of the publications of Tim Berners-Lee and the worldwide web consortium. **Journal of the Association for Information Science and Technology**, v. 70, n. 7, p. 701-714, 2019.

MARTELETO, R. M. Cultura da modernidade: discussões e práticas informacionais. **Revista da Escola de Biblioteconomia da UFMG,** v. 23, n. 2, 1994.

MARTINS, D. L. *et al*. Information organization and representation in digital cultural heritage in Brazil: Systematic mapping of information infrastructure in digital collections for data science applications. **Journal of the Association for Information Science and Technology**, v. 74, n. 6, p. 707-726, 2022.

MARTINS, D. L.; CARMO, D.; SILVA, M. F. Modelos de governança em serviços de acervos digitais em rede: elementos para a produção de uma política pública nacional para objetos culturais digitais. **Perspectivas em Ciência da Informação,** v. 27, p. 81-109, 2022.

MARTINS, D. L.; LEMOS, D. L.; ANDRADE, M. C. 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, D. L.; LEMOS, D. L.; CARMO, D.; SIQUEIRA, J.; OLIVEIRA, L. F. R. Requisitos de qualidade para dados de agregação em museus. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 14, 2021.

MARTINS, D. L.; SILVA, M. F.; CARMO, D. Acervos em rede: perspectivas para as instituições culturais em tempos de cultura digital. **Em Questão**, v. 24, n. 1, p. 194-216, 2018.

MESSAOUDI, T.; VÉRON, P.; HALIN, G.; DE LUCA, L. An ontological model for the reality-based 3D annotation of heritage building conservation state. **Journal of Cultural Heritage**, v. 29, p. 100–112, 2018.

MOOERS, C. N. The next twenty years in information retrieval; some goals and predictions. **American Documentation**, v. 11, n. 3, p. 229–236, 1960.

NISO (National Information Standards Organization). (2005). **Guidelines for the construction, format, and management of monolingual controlled vocabularies**. ANSI/NISO Z39.19, NISOPress, 2005.

OLIVEIRA, A. de A.; FEITOSA, A. C. A. A difusão digital nos museus IBRAM: a implantação do projeto Tainacan. **Revista Eletrônica Ventilando Acervos**, n. 1, p. 70-90, 2021.

PARK, J.; BRENZA, A. Evaluation of semi-automatic metadata generation tools: A survey of the current state of the art. **Information technology and libraries**, v. 34, n. 3, p. 22-42, 2015.

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

POORT, J. *et al*. **The value of Europeana:** the welfare effects of better access to digital cultural heritage. SEO, 2013.

POTENZIANI, M. *et al*. Publishing and consuming 3D content on the web: A survey. **Foundations and Trends® in Computer Graphics and Vision**, v. 10, n. 4, p. 244-333, 2018.

PURDAY, J. Think culture: Europeana. eu from concept to construction. **BIBLIOTHEK Forschung Und Praxis**, v. 33, n. 2, 2009.

RANGANATHAN, S. R. **Prolegomena to library classification**. Asia Publishing House, 1967.

ROBLEDANO-ARILLO, J.; NAVARRO-BONILLA, D.; & CERDÁ-DÍAZ, J. Application of linked open data to the coding and dissemination of Spanish civil war photographic archives. **Journal of Documentation**, v. 76, n. 1, p. 67–95, 2020.

ROLLITT, K. DigitalNZ ā-tihi o Aotearoa: Connecting the Digital Content of New Zealand: Advice, Open Standards and Interoperability. *In*: **International Conference on Dublin Core and Metadata Applications**. 2009.

SALINAS, C. M. Desafíos de la preservación digital del patrimonio cultural en México: el caso de Mexicana. **Cuadernos. info**, n. 55, p. 211-232, 2023.

SAMPAIO, N. A. S.; ASSUMPÇÃO, A. R. P.; FONSECA, B. B. **Estatística descritiva**. Belo Horizonte: Poisson, 2018.

SARACEVIC, T. Information science. **Journal of the American Society for Information Science**, 1999.

SCOPIGNO, R. *et al*. Delivering and using 3D models on the web: are we ready?. **Virtual Archaeology Review**, v. 8, n. 17, p. 1-9, 2017.

SECRETARÍA DE CULTURA. MEXICANA Repositório del patrimônio cultural del México. 2018. Disponível em: https://mexicana.cultura.gob.mx/work/models/repositorio/Resource/126/2/images/documentacion.pdf. Acesso em: 27 de agosto de 2023.

SIQUEIRA, J.; CARMO, D.; MARTINS, D. L.; LEMOS, D. L., MEDEIROS, V.N., OLIVEIRA, L.F.R. Elements for Constructing a Data Quality Policy to Aggregate Digital Cultural Collections: Cases of the Digital Public Library of America and Europeana Foundation. *In*: **Data and Information in Online Environments: Second EAI International Conference, DIONE 2021, Virtual Event**, Proceedings. Cham: Springer International Publishing, p. 106-122, 2021.

SIQUEIRA, J.; MARTINS, D. L. Acervos agregados do Instituto Brasileiro de Museus: desenvolvimento do painel de visualização analítica. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 14, 2021.

SIQUEIRA, J.; MARTINS, D. L. Recuperação de informação: descoberta e análise de workflows para agregação de dados do patrimônio cultural. **Ciência da Informação**, v. 49, n. 3, 2020.

SIQUEIRA, J.; MARTINS, D. L. Workflow models for aggregating cultural heritage data on the web: A systematic literature review. **Journal of the Association for Information Science and Technology**, v. 73, n. 2, p. 204-224, 2022.

SIQUEIRA, J.; MARTINS, D. L.; LEMOS, D. L. S. Brasiliana museus: serviço de busca e recuperação da informação agregada dos acervos digitais do instituto brasileiro de museus. **XXII Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação**, 2022. Disponível em: http://hdl.handle.net/20.500.11959/brapci/201748. Acesso em: 14 jul. 2023.

SIQUEIRA, J.; MARTINS, D. L.; MEDEIROS, V. N. BRASILIANA MUSEUS: teste funcional do agregador de dados museais do Instituto Brasileiro de Museus. **IV Workshop de Informação, Dados e Tecnologia**. 2022.

SUPPLEJACK. **Arquitetura**. 2023. Disponível em: https://digitalnz.github.io/supplejack/architecture.html. Acesso em: 25 ago. 2023.

SVENONIUS, E. **The intellectual foundation of information organization**. The MIT Press, 2000.

TAYLOR, A. G. **The organization of the information**. Libraries Unlimited, 2004.

TESSLER, A. **Economic valuation of the British Library**. London: Oxford Economics, 2013.

TORINO, E; MONTEIRO, E. C. S. A; VIDOTTI, S. A. B. G. Plano de gestão de dados de pesquisa de povos indígenas: considerações acerca dos princípios FAIR e CARE. **Revista Brasileira de Preservação Digital**, 2023.

TRIQUES, M. L.; GONÇALEZ, P. R. V. A.; ALBUQUERQUE, A. C. A integração de dados culturais de repositórios digitais um panorama dos Hubs da DPLA. **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação**, v. 20, 2023.

TROVE. **Technical ecosystem**. 2023. Disponível em: https://trove.nla.gov.au/about/what-trove/technical-ecosystem. Acesso em: 25 ago. 2023.

VIRKUS, S.; GAROUFALLOU, E. 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.

VIRKUS, S.; GAROUFALLOU, E. Data science from a library and information science perspective. **Data Technologies and Applications**, v. 53, n. 4, p. 422–441, 2019.

WILKINSON, M. D. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. **Scientific data**, v. 3, n. 1, p. 1-9, 2016.

WINER, D.; ROCHA, I. E. Europeana: um projeto de digitalização e democratização do patrimônio cultural europeu. **Patrimônio e Memória**, v. 9, n. 1, p. 113-127, 2013.

XAVIER, A.; HERNANDEZ, F. OAI-PMH and Linked Open Data in the context of Hispana and Europeana: some historical reflections. **Italian Journal of Library, Archives and Information Science**, p. 1-16, 2020.

YIN, R. K. **Estudo de Caso, planejamento e métodos**. 2.ed. São Paulo: Bookman, 2001.

ZENG, M. L. Interoperability. **Knowledge Organization**, v. 46, n. 2, p.122–146, 2019.

ZENG, M. L.; QIN, J. **Metadata**. ALA Neal-Schuman. 2016.