Tendências da Pesquisa Brasileira em Ciência da Informação

PLATAFORMAS DE AGREGAÇÃO DE DADOS CULTURAIS: A TRAJETÓRIA DO PROJETO MIDIATECA CAPIXABA[1]

CULTURAL DATA AGGREGATION PLATFORMS: THE TRAJECTORY OF THE MIDIATECA CAPIXABA PROJECT

Daniela Lucas da Silva Lemos[2]
Dalton Lopes Martins[3]
Dirceu Flavio Macedo[4]

Resumo: Instituições provedoras do patrimônio cultural têm buscado cada vez mais recursos para disponibilizar suas fontes de informação em rede para acesso e consumo de dados. Entretanto, apesar de grande parte de instituições brasileiras do patrimônio cultural se encontrar imbuída em processos de digitalização de seus acervos e coleções, não utilizam infraestruturas informacionais sofisticadas que permitam o tratamento de conteúdo informacional com qualidade para variados tipos de mídia e em variados contextos informacionais. Assim, o objetivo é apresentar a trajetória do projeto Midiateca Capixaba, incluindo principais conceitos e aspectos teórico-metodológicos que sustentaram frentes de pesquisa centrais na disponibilização de informações históricas, culturais e científicas à sociedade. Metodologicamente, a criação e a manutenção da plataforma Midiateca demandaram uma arquitetura de gestão da informação que se fundamentou num modelo de mineração de dados, conhecido como CRoss-Industry Standard Process for Data Mining, voltado, nesta pesquisa, a geração de acervos digitais online. O resultado foi a construção de uma plataforma digital que integra vários repositórios institucionais culturais do Estado do Espírito Santo por meio de um agregador que viabiliza a busca, a recuperação e a navegação de forma integrada, o que exigiu habilidades em técnicas de ciência de dados e computacionais para aprimorar os processos de organização, tratamento e análise de dados nos espaços culturais. Conclui-se que os aspectos teórico-metodológicos apresentados na pesquisa, bem como as tecnologias utilizadas, mostraram-se viáveis de serem aplicados em outros projetos e potenciais políticas públicas voltadas à organização, difusão e acesso ao patrimônio cultural digital.

---

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

[2] Doutora. Docente na Universidade Federal do Espírito Santo. E-mail: daniela.l.silva@ufes.br. ORCID: https://orcid.org/0000-0003-1565-7366.

[3] Doutor. Docente na Universidade de Brasília. E-mail: daltonmartins@unb.br. ORCID: https://orcid.org/0000-0002-6244-6791.

[4] Mestre e Pesquisador na Universidade Federal do Espírito Santo. E-mail: dfmacedo@gmail.com. ORCID: https://orcid.org/0000-0003-4837-3618.

Palavras-Chave: Patrimônio cultural digital. Plataforma de agregação. CRoss-Industry Standard Process for Data Mining. Midiateca Capixaba. Acervos digitais.

Abstract: *Cultural heritage-providing institutions have increasingly sought resources to make their information sources available online for data access and consumption. However, although a significant number of Brazilian cultural heritage institutions are engaged in digitizing their collections and holdings, they often do not employ sophisticated information infrastructures capable of processing informational content with the quality required for various media types and in diverse informational contexts. This paper aims to present the trajectory of the Midiateca Capixaba project, including its main concepts and the theoretical-methodological aspects that supported core research initiatives focused on making historical, cultural, and scientific information accessible to society. Methodologically, the creation and maintenance of the Midiateca platform required an information management architecture based on a data mining model known as the Cross-Industry Standard Process for Data Mining, which, in this research, was directed toward the generation of online digital collections. The result was the development of a digital platform that integrates multiple cultural institutional repositories from the State of Espírito Santo through an aggregator that enables integrated search, retrieval, and navigation. This process required expertise in data science and computational techniques to enhance the organization, processing, and analysis of data within cultural spaces. It is concluded that the theoretical-methodological aspects presented in this study, as well as the technologies employed, have proven to be feasible for application in other projects and in potential public policies aimed at the organization, dissemination, and access to digital cultural heritage.*

Keywords: *Digital cultural heritage. Aggregation platform. CRoss-Industry Standard Process for Data Mining. Midiateca Capixaba. Digital Collections.*

1 INTRODUÇÃO

Instituições provedoras do patrimônio cultural têm buscado cada vez mais recursos para disponibilizar suas fontes de informação em rede para acesso e consumo de dados, colocando em evidência a prática comunicacional e a difusão de acervos e coleções culturais digitalizados (Siqueira; Martins, 2020).

A disponibilização desses acervos e coleções em formatos digitais de maneira aberta e distribuída é resultado do uso de padrões abertos, da interoperabilidade e da organização e tratamento de fontes de informação em infraestruturas informacionais, tal como as plataformas digitais, consideradas,

logo, serviços de informações essenciais para garantir que os objetos digitais sejam acessíveis, integráveis com outros sistemas e utilizáveis funcionalmente, promovendo um acesso democrático à cultura e fortalecendo a preservação digital (Hyvönen, 2012).

Borgman (2010) entende infraestrutura informacional em ambientes abertos e distribuídos como um conjunto de elementos técnicos, sociais e políticos, incluindo pessoas, tecnologias, ferramentas e serviços, usados em favor da disseminação colaborativa de conteúdos em rede.

Nesse contexto, pesquisas de âmbito internacional e nacional têm sido realizadas sobre plataformas digitais (Siqueira; Martins, 2020; Siqueira; Martins; Lemos, 2022; Martins *et al.*, 2022; Pereira; Lemos, 2024), as quais têm proporcionado inovação na forma de produção e tratamento das fontes de informação, dentro do contexto cultural, disponibilizando serviços e conteúdos custodiados por instituições culturais, a exemplo da *Digital Public Library of America* (DPLA)[5], Europeana[6], Mexicana[7], Brasiliana Museus[8] e Midiateca Capixaba[9].

Entretanto, apesar de grande parte de instituições brasileiras do patrimônio cultural se encontrar imbuída em processos de digitalização de seus acervos e coleções para disponibilização na internet, não utilizam infraestruturas informacionais alinhadas com as principais tendências científicas e tecnológicas que permitam o tratamento de conteúdo informacional com qualidade para variados tipos de mídia e em variados contextos informacionais, o que pode redundar em desafios significativos para a organização, a representação, a

[5] Disponível em: https://dp.la/.
[6] Disponível em: https://www.europeana.eu/pt
[7] Disponível em: https://mexicana.cultura.gob.mx/
[8] Disponível em: https://brasiliana.museus.gov.br/
[9] Disponível em: https://midiateca.es.gov.br/site/

recuperação, a agregação, o acesso, o uso e a reutilização eficiente da informação por instituições, governos e indivíduos (Martins *et al.*, 2022).

O tratamento, neste caso, requer uma representação da informação e do conhecimento diferenciada, visando melhorar os sistemas de catalogação, classificação, indexação, busca, navegação, agregação e recuperação, de forma a obter resultados mais significativos a várias tipologias de usuários em ambiente digital. Para além, o profissional da informação, envolvido nessas ações institucionais, geralmente não possui formação adequada para lidar com novos serviços de inteligência, no que diz respeito a conhecimentos em tecnologias relevantes para manipulação de dados na web, como a implantação de padrões para a estruturação de base de dados com qualidade em plataformas digitais, seguindo modelos contemporâneos para tratamento documental que visem normalização, qualidade e intercâmbio de descrições (Lemos; Martins; Souza, 2023).

Tal cenário dificulta a prestação de serviço de qualidade aos usuários, o que implica em garantir que tenham acesso a informações precisas, atualizadas, acessíveis e confiáveis, contribuindo, portanto, para uma experiência positiva em relação à plataforma. Além disso, dificulta a reutilização de fontes de informação na web, considerando que os objetos digitais são elementos preponderantes na geração de valor em suas circulações e reúso social. Isso, inclusive, prejudica o avanço do crescimento de campos emergentes, como as Humanidades Digitais (Poole, 2017), que buscam a extração de valor em objetos digitais por meio de ações de reúso para o desenvolvimento econômico, social e humano.

Nessa perspectiva, com o propósito de tornar os acervos amplamente acessíveis, o governo do Estado do Espírito Santo, sob a gestão da Secretaria de Estado da Cultura (SECULT/ES), promoveu um investimento no desenvolvimento

da plataforma digital de agregação de dados culturais, denominada Midiateca Capixaba. A plataforma é considerada um sistema de informação gratuito e democrático que preserva e difunde acervos de instituições e agentes culturais capixabas, incluindo documentos, fotografias, vídeos, músicas, matérias de jornal, artes gráficas, livros, periódicos, obras de arte e filmes.

A Midiateca é constituída atualmente por um total de 4.847 itens que compõem as coleções digitais das instituições culturais envolvidas, administradas diretamente pelo Estado, quais sejam a Biblioteca Pública do Estado, o Arquivo Público do Estado, o Museu de Arte, a Galeria Homero Massena, o Museu do Colono, a Orquestra Sinfônica, o Palácio Anchieta, o Conselho Estadual de Cultura, o Funcultura e a Rádio TV do Estado, proporcionando facilidade de acesso público a conteúdos históricos e culturais sob sua guarda (MIDIATECA, 2024).

A plataforma Midiateca Capixaba, portanto, se alinha com o trabalho das Humanidades Digitais no Brasil e no exterior, especialmente no que diz respeito ao cenário das fontes documentais na área da cultura em que, no Brasil, se constata uma carência de informação sistematizada e a ausência de bases de dados curadas e de ampla representatividade científica e cultural para fins de organização integrada de documentos, de busca e de recuperação de informação (Martins *et al.*, 2022).

Assim sendo, torna-se relevante e oportuno despontar a Midiateca Capixaba no sentido de apresentar um projeto de pesquisa brasileiro com uma proposta de soluções inovadoras de processamento, navegação e visualização de dados em rede, que buscou tornar a gestão das instituições de cultura capixaba orientadas por dados, no sentido de endereçá-los com qualidade à internet e, consequentemente, auxiliar seu enorme e heterogêneo público a explorar seus

acervos históricos e culturais, por meio de estratégias voltadas ao engajamento deste e estimulando o desenvolvimento de outros produtos e serviços, como desdobramento.

A presente pesquisa busca, então, responder a seguinte questão: *como criar um serviço de informação digital para ampliar a capacidade de busca, recuperação e circulação em rede dos acervos digitais das instituições de cultura do Estado do Espírito Santo?* Logo, o objetivo desta pesquisa é apresentar a trajetória do projeto Midiateca Capixaba, incluindo principais conceitos e aspectos teórico-metodológicos que sustentaram frentes de pesquisa centrais na disponibilização de informações históricas, culturais e científicas à sociedade conectada na rede, especialmente para o Estado do Espírito Santo, Brasil.

O artigo está estruturado como se segue. Esta introdução contextualiza e problematiza a pesquisa, trazendo o seu objetivo e justificando a sua realização; a seção 2 traz conceitos e teorias subjacentes às plataformas digitais que lidam com a função de agregar dados, especialmente os da área da cultura; a seção 3 descreve os procedimentos metodológicos que sustentaram a concretização dos projetos associados à plataforma; a seção 4 apresenta os resultados relacionados aos projetos envolvendo a Midiateca Capixaba; a seção 5 tece uma discussão acerca dos resultados; e a seção 6 expõe as considerações finais e sugestões para trabalhos futuros.

2 PLATAFORMAS DE AGREGAÇÃO DE DADOS CULTURAIS

Instituições culturais estão inovando suas formas de interagir com o público, especialmente na disponibilização de objetos digitais e seus metadados em sites, bibliotecas digitais, repositórios institucionais ou plataformas digitais, colocando em evidência a prática comunicacional e a difusão de seus acervos

digitalizados (Marcondes, 2016; Martins *et al.*, 2022; Siqueira; Martins, 2020; Siqueira; Martins; Lemos, 2022)

De acordo com Marcondes (2016), as tecnologias da informação sempre foram usadas para agregar diferentes acervos, potencializando suas sinergias e complementaridades, e provendo melhores serviços aos seus usuários. Contudo, a agregação de dados culturais não é uma tarefa trivial, pois os metadados e objetos digitais são diversos e singulares, dificultando, sobremaneira, a definição de padrões tecnológicos e de documentação para ambiente web, no qual possui toda uma dinâmica de socialização, potencialidades de integração e acessibilidade aos acervos (Siqueira; Martins, 2020; Martins *et al.*, 2022).

Assim, instituições e governos buscam manter seus dados abertos na internet para democratização e transparência em seus produtos e serviços de informação, contudo, precisam, acima de tudo, de formatos e padrões comuns interoperáveis em suas bases, cujos dados geralmente são manifestados em variados tipos de mídias. Tais padrões de qualidade são recomendados, especialmente, pelos princípios orientadores do FAIR, um acrônimo para *Findable*, *Accessible*, *Interoperable* e *Reusable*, de catalogação, e do *Linked Open Data* (Lemos; Souza, 2024).

A busca por esses padrões, portanto, prevê a construção de um ecossistema informacional aberto em rede, no qual humanos e máquinas podem encontrar, acessar, interoperar e reutilizar de forma transparente dados e metadados das variadas fontes de informação disponíveis na internet.

Entretanto, o que ocorre geralmente é que diversas soluções de *software* são utilizadas para publicar dados e coleções online, sem uma estratégia comum entre as instituições, resultando em dados produzidos em diferentes aspectos de descoberta, formatos, idiomas, direitos autorais e operacionais sobre esses

dados, integração, o que representa um desafio para instituições com interesse em divulgar seus dados e metadados em rede (Hyvönen, 2012; Martins *et al.*, 2022). Logo, um dos maiores desafios nesses projetos envolvendo agregação de dados encontra-se na coleta, na extração e na organização de dados legados da documentação cultural para a construção de sistemas de informação eficientes do ponto de vista da recuperação por parte de seus usuários.

Nesse contexto, as plataformas de agregação de dados culturais surgem como ferramentas essenciais para a preservação e disseminação de dados de instituições de cultura, memória e patrimônio, reunindo conteúdos de bibliotecas, museus, arquivos e outras em ambientes unificados, cuja base tecnológica permite a estruturação de coleções, a ampliação de acesso, a entrega personalizada a várias tipologias de usuários, e a agregação de serviços e conteúdos oriundos de provedores desses recursos (Siqueira; Martins, 2020). Hyvönen (2012) complementa assinalando que as plataformas digitais dessa natureza oferecem não apenas uma forma de preservar, mas também de democratizar o acesso ao patrimônio cultural, facilitando a interação e o engajamento de usuários em um espaço digital ampliado.

Nesse engajamento, a curadoria digital (Freire; Sales; Sayão, 2020; Marcondes, 2022) exerce um importante papel na organização, interpretação e disponibilização de dados de acervos de maneira estratégica, visando recuperação, navegação e reúso de conteúdos relevantes por parte de usuários. No entanto, a curadoria digital carece de ser entendida como uma prática interdisciplinar pelos seus inúmeros desafios envolvendo o tratamento em vários aspectos sobre objetos digitais diversos e complexos, destinados a um público global e diversificado. Para estruturar essa curadoria, adotam-se referenciais que garantem a integridade e relevância dos dados, incluindo critérios de qualidade,

padrões de descrição, indexação e classificação, e a observância de diretrizes éticas e legais (Marcondes, 2022; Lemos; Martins; Souza, 2023).

Exemplos internacionais de plataformas de agregação de dados culturais ilustram bem essa dinâmica curatorial (Siqueira; Martins, 2020). A Europa, por exemplo, lançou, em 2008, o protótipo Europeana, que deu acesso, logo no lançamento, a 4.5 milhões de objetos digitais de bibliotecas, museus, arquivos audiovisuais e galerias. Em 2020, forneceu acesso a 58 milhões de objetos digitais, com sofisticadas ferramentas de pesquisa e filtro, além de coleções temáticas, exposições, galerias e blogs (Europeana, 2020, on-line). No cenário nacional, destacam-se diversas iniciativas de plataformas de agregação de dados culturais (Siqueira; Martins; Lemos, 2022; Martins *et al.*, 2022; Pereira; Lemos, 2024). Citam-se as plataformas de agregação de dados culturais Midiateca Capixaba [Secult/ES] e a Brasiliana Museus [Instituto Brasileiro de Museus], ambas permitindo serviços de integração e interoperabilidade de dados culturais, ampliando o acesso e o reúso das fontes de informação disponíveis.

Em suma, as plataformas de agregação de dados culturais podem ser consideradas infraestruturas informacionais em ambientes abertos e distribuídos que desempenham um papel essencial na democratização do acesso ao patrimônio cultural, estimulando a pesquisa, a colaboração e a inclusão digital, além de promoverem a transparência e a organização dos ambientes digitais por meio de padrões tecnológicos e documentais para fins de disseminação colaborativa e consumo de conteúdos com qualidade em rede (Borgman, 2010; Lemos; Martins; Souza, 2023).

3 PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa é classificada como sendo de natureza aplicada, de abordagem qualitativa, e de caráter exploratório e descritivo, envolvendo a plataforma digital de agregação de dados culturais Midiateca Capixaba.

O processo de criação da plataforma Midiateca Capixaba (primeira fase) e sua continuidade como pesquisa (segunda fase) demandou um modelo abstrato de arquitetura e gestão da informação que se desdobrou em três frentes principais de ação:

- Modelo tecnológico: envolveu a arquitetura da rede de sistemas de informação e seus relacionamentos de forma a garantir a interoperabilidade e a integração.
- Modelo informacional: envolveu o modelo de metadados, as práticas de catalogação e os vocabulários controlados que foram utilizados no projeto.
- Modelo de governança: envolveu o desenho de gestão do projeto, descrevendo seus atores, os papéis desempenhados, as formas de tomada de decisão e as formas de organização do projeto.

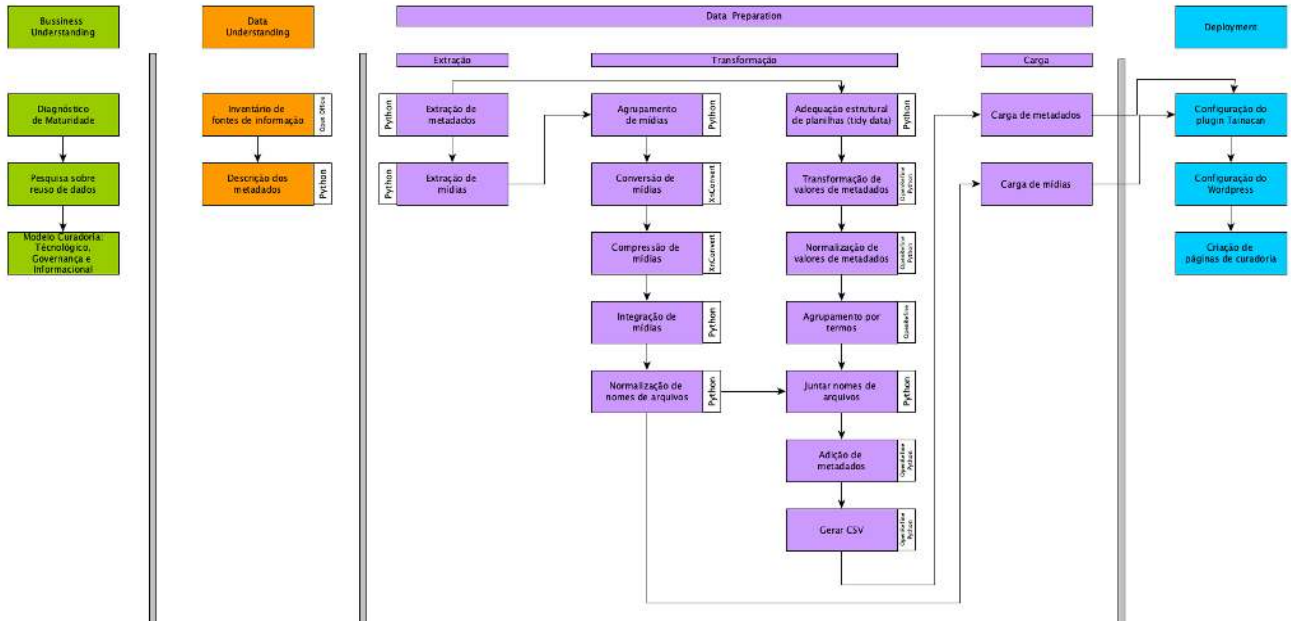
Após uma revisão na literatura nas áreas de Ciência de Dados (CD) e Ciência da Informação (CI), e de modo que o modelo abstrato de arquitetura e gestão da informação pudesse ser operacionalizado para a geração de acervos digitais online, foi proposto como método de pesquisa, para a construção do fluxo de informações do projeto, o uso do modelo de mineração de dados conhecido como *CRoss-Industry Standard Process for Data Mining* (CRISP-DM) (Shearer, 2000).

O modelo CRISP-DM teve sua origem na década de 1990 em projetos de *Dataware House* e *Business Inteligence* no universo corporativo. Teve destaque por ser considerado um método de trabalho eficiente para o conhecimento do ambiente de negócio para fins de levantamento de requisitos úteis à modelagem

de dados, visando extrair conhecimento útil e estratégico dos dados aos negócios de interesse. Além disso, o método proposto pelo modelo se afina com o campo da informação e tecnologia ao permitir uma modelagem apropriada e a produção de metadados pelo profissional da informação em processos geralmente orientados por soluções inteligentes de organização e tratamento de informações oriundas de campos científicos como a CI e a CD (Virkus; Garoufallou; 2020).

O modelo, conforme etapas sugeridas na Figura 1, é dividido em 4 blocos, incluindo o entendimento do negócio (*Business Understanding*), o entendimento dos dados (*Data Understanding*), a preparação, modelagem e avaliação dos dados (*Data preparation, modeling, evaluation*) e, por fim, a visualização dos dados (*Deployment*).

Figura 1- O modelo CRISP-DM para geração de acervos digitais online



Fonte: elaborado pelos autores (2024)

Para efeitos do trabalho que é realizado, visando a construção de um repositório digital, o método é utilizado de modo a garantir o entendimento dos

requisitos do projeto e da execução de um diagnóstico completo do estado atual dos processos de digitalização e documentação da instituição cultural, etapa denominada "*Bussiness understanding*". Uma vez compreendido o contexto organizacional da instituição cultural, ao entrar na próxima etapa, denominada "*Data understanding*", todas as bases de dados e fontes de informação que serão utilizadas serão analisadas. Para tal, busca-se identificar todos os tipos de dados presentes (se são numéricos, data, textuais, categóricos, entre outras), a quantidade de registros disponíveis e as mídias utilizadas, de modo que se tenha um inventário informacional completo da situação atual dos objetos físicos e digitais que serão utilizados para a constituição do repositório digital.

A seguir, a próxima etapa pode ser considerada a que exige maior tempo de execução e maior uso de ferramentas tecnológicas para operacionalização, a etapa denominada "*Data preparation*". Nesta, dois fluxos são criados de forma simultânea para tratamento de dados, sendo um deles dedicado ao desenvolvimento de atividades de tratamento dos metadados, visando melhorias substanciais em sua qualidade, e outro dedicado ao tratamento das mídias digitais, visando adequações necessárias para a sua boa relação com os seus respectivos metadados bem como adaptações para carga em repositório digital. Por fim, uma vez os dados e mídias tratados, entra-se na última etapa do método, denominada "*Deployment*", em que o sistema é configurado, páginas são criadas e os metadados e mídias digitais carregados para a implantação do repositório digital.

4 RESULTADOS

(Primeira Fase) Processamento da documentação digital legada das instituições

A primeira fase do projeto consistiu na criação de uma plataforma digital online para preservação e difusão de acervos nos mais diferentes suportes. A plataforma é composta pelo Arquivo Público do Estado do Espírito Santo (APEES), que disponibiliza cartazes; pela Biblioteca Pública do Espírito Santo "Levy Cúrcio da Rocha" (BPES), que fornece livros e periódicos; pela Galeria Homero Massena, que oferece esculturas, gravuras em metal e pinturas; pelo Museu de Arte do Espírito Santo Dionísio Del Santo (MAES), que dispõe desenhos, esculturas, gravuras e pinturas; pelo Museu do Colono, que expõe objetos museológicos; pelo Palácio Anchieta, que expõe achados arqueológicos, catálogos, esculturas, gravuras, livros, mobiliário, pinturas, porcelanas, relatórios, tapeçarias e utilitários; e, por fim, pela Rádio e Televisão do Espírito Santo (RTV/ES), que possibilita o acesso a arquivos audiovisuais digitais (MIDIATECA, 2024). Tais fontes de informação representam os acervos bibliográficos, museológicos e arquivísticos da plataforma, evidenciando a sua característica heterogênea nos formatos e tipos documentais.

Para que as fontes de informação fossem disponibilizadas pela plataforma de agregação de dados culturais em formato digital, permitindo a encontrabilidade, a identificação, a seleção, a aquisição, o reúso, além de possibilitar a navegação e exploração do conteúdo digital (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2016), foi necessário realizar a análise do ambiente atual, o que se desdobrou numa ação de inventário dos itens envolvidos, o tratamento documental e a inserção dos itens em repositório digital, juntamente com as informações provenientes das

instituições culturais, considerando toda a sua complexidade e abrangência. Tais etapas podem ser observadas nas quatro camadas indicadas no diagrama da Figura 2. Cabe ressaltar que todos os procedimentos sugeridos na Figura 1 são antecedentes técnicos necessários para se chegar à camada 1 da Figura 2, ou seja, existir um repositório digital com dados disponibilizados para que os mesmos possam ser agregados em conjunto com outras instituições num agregador unificado.

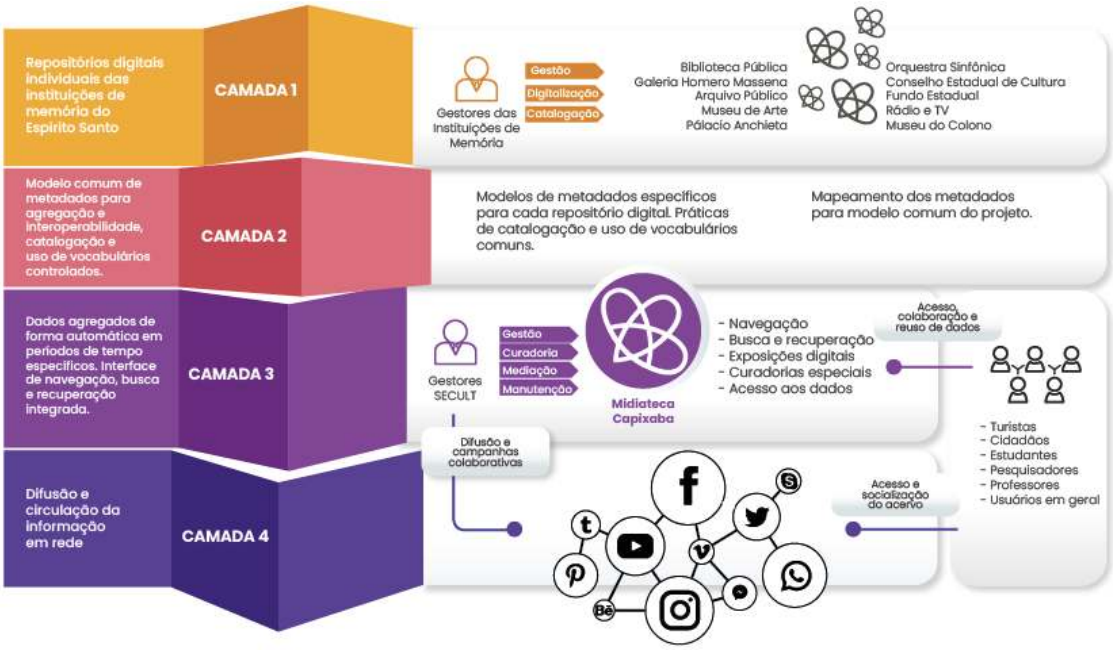
A camada 1 apresenta a maneira pela qual os objetos digitais e seus metadados foram obtidos e enviados ao Tainacan[10] (*software* de repositório utilizado no projeto de pesquisa) de cada instituição, a partir da documentação legada dessas instituições das quais geriram, digitalizaram e catalogaram seus documentos ao longo dos anos, com suas regras de negócio específicas, em vários formatos diferentes, incluindo catálogos em PDF, documentos descritivos em formato Word, planilhas de dados contendo os metadados descritivos e temáticos dos objetos culturais, bases de dados em formato Access, entre tantos outros formatos que ainda hoje são produzidos pelas instituições de memória e utilizados como recursos computacionais para a documentação de objetos culturais.

Apesar de essas ferramentas serem de fácil acesso aos profissionais da cultura e produzirem boa produtividade de escritório para as instituições, são pouco acessíveis para ambiente web, no qual possui toda uma dinâmica de socialização, potencialidades de integração e acessibilidade aos acervos (Martins; Lemos; Andrade, 2021). Logo, um dos maiores desafios desse projeto foram as ações de extração, transformação e carga (conforme Figura 1) dos dados legados da documentação cultural (envolvendo metadados e mídias) para a

[10] https://tainacan.org/.

construção dos repositórios digitais. Para essas ações, fundamentadas em aspectos teórico-metodológicos da CI e da CD, foram desenvolvidos vários *scripts* em linguagem Python[11] que estão disponíveis publicamente no repositório Github[12]. Também foi utilizada a ferramenta Open Refine[13], sobretudo para o trabalho de normalização terminológica e agrupamento dos termos dos vocabulários controlados (Lancaster, 1986) utilizados no projeto, por meio dos quais se torna possível auxiliar os processos de análise e descrição de documentos, permitindo a criação padronizada de metadados ao nomear, de forma consistente, os pontos de acesso aos documentos e a informação neles contida; além de auxiliar na geração de facetas e filtros consistentes na interface da plataforma através da padronização e expansão do vocabulário das consultas (Siqueira; Martins; Lemos, 2022).

Figura 2 - Etapas constitutivas da primeira fase para o lançamento da Midiateca



Fonte: elaborado pelos autores (2024)

[11] https://www.python.org/
[12] https://github.com/tainacan/data_science/tree/master/MIDIATECACAPIXABA
[13] https://openrefine.org/

Na camada 2, uma vez construído o repositório de cada instituição com metadados específicos para cada perfil de aplicação (Baker; Nilsson; Johnston, 2008), foram realizados procedimentos de modelagem conceitual da informação, utilizando-se do padrão de documentação Dublin Core[14] como modelo comum de metadados para a ação de mapeamento ou cotejamento semântico com os dados oriundos das instituições provedoras de fontes de informação, intencionando, para tal, agregação e interoperabilidade, catalogação e uso de vocabulários controlados.

Há de se destacar que o padrão de metadados Dublin Core foi escolhido para o projeto por ser uma referência global para descrição de recursos digitais, incluindo domínios bibliográficos e de patrimônio cultural em geral. É um padrão que cobre de maneira satisfatória metadados para criação e produção, classificação, informação e uso sobre o recurso digital. É válido destacar que o padrão possui qualificadores que ampliam a cobertura de seus elementos para um dado perfil de aplicação, o que agrega positivamente na extensão de seus modelos. Em adição, possui extensões para linguagens de representação do *World Wide Web Consortium* (W3C), o qual concede padrões abertos a fim de interligar e anotar dados sob licença aberta (*linked open data*) por meio de um *Uniform Resource Identifier* (URI). A proposta possibilita que provedores de conteúdo (neste caso, as instituições de cultura) enriqueçam seus esquemas de metadados com especificações estruturadas e bem definidas de conhecimento a partir de vocabulários controlados, ontologias e padrões de metadados, o que pode viabilizar o consumo e o reúso de informações de alta qualidade disponíveis em rede (Marcondes, 2016), sendo, de antemão, uma frente de pesquisa estratégica na evolução da plataforma Midiateca Capixaba.

[14] https://www.dublincore.org/specifications/dublin-core/

Na camada 3, os dados organizados e persistidos nos repositórios digitais institucionais, incluindo arquivos de mídias e seus metadados, são então coletados de forma automática, em períodos específicos, por meio de suas *Application Programming Interface* (APIs), utilizando-se componentes da tecnologia Elastic Search[15] (Siqueira; Martins; Lemos, 2022). Estando os metadados e arquivos de mídias integrados em uma única base de dados, surge a solução agregadora de navegação, exploração, busca e recuperação integrada da informação, a Midiateca Capixaba, também construída utilizando a tecnologia Tainacan. A camada 3 possibilita, portanto, ofertar soluções inovadoras para que os gestores da SECULT possam exercer papéis de curadores e mediadores de bases de dados de acervos bem construídas a partir de padrões de documentação. É válido ressaltar que bases de dados são produtos de informação considerados cruciais para a sociedade quando adotados para realizar a mediação entre documentos e comunidade de usuários interessada na preservação, busca, recuperação, acesso e reúso, pois referenciam e divulgam o conhecimento a partir do uso qualificado da informação (Lemos; Martins; Souza, 2023). Nesse contexto, na plataforma Midiateca, a curadoria digital desempenha papel crucial na montagem de exposições a partir do acesso e reúso de dados pelo público, permitindo, assim, que a coleção digital seja visualizada e explorada de forma interativa e contextualizada para finalidades diversas (Marcondes, 2022).

Por fim, a camada 4 tornou-se um resultado de uma infraestrutura informacional bem projetada para que objetos digitais sejam acessíveis a variadas tipologias de usuários, integráveis com outros sistemas e com facilidades em seu manuseio, em ambientes abertos e distribuídos, usados, portanto, em favor da

[15] https://www.elastic.co/pt/

disseminação colaborativa de conteúdos em rede. Nesse cenário, os acervos e coleções das instituições provedoras conseguem ser difundidos e socializados em mídias sociais que se integram ao paradigma da Web social, visto que a tecnologia Tainacan se percebe com muita facilidade em funcionalidades voltadas a compartilhamento de conteúdo em rede, capacidade de engajamento de usuários na revisão de metadados com sugestão de melhorias na descrição de conteúdo, e mecanismos de votação para que os usuários possam escolher conteúdos de sua preferência (Martins; Lemos; Andrade, 2021). Um bom exemplo de tal aplicação pode ser encontrado no perfil do Instagram[16] da Midiateca Capixaba, por meio do qual são executadas diversas experiências com formas de publicação de acervos e recursos de engajamento social com o público digital. O perfil possui mais de 1480 seguidores e já realizou mais de 154 publicações.

(Segunda Fase) Inventário, Digitalização e Catalogação de acervos

A plataforma Midiateca Capixaba em sua origem já se faz um centro de documentação, um banco de dados com acervo público, que possibilita pesquisas e atividades educacionais, análise de indicadores, mecanismos de busca integrada, viabilizando o monitoramento, a avaliação e o aprimoramento das políticas culturais do Espírito Santo. Nesse sentido, a segunda fase do projeto (ainda em curso), denominado Midiateca Capixaba Conectando[17], encontra-se orientada em universalizar o acesso a história, a arte e a cultura, estimulando suas presenças no ambiente educacional e ampliando o intercâmbio da cultura capixaba no mundo contemporâneo.

[16] https://www.instagram.com/midiatecacapixaba/
[17] https://midiateca.es.gov.br/site/midiateca-capixaba-conectando/

Para a presente pesquisa, o foco concentra-se no relato dos processos provenientes do inventário, digitalização e catalogação (Eixo 2 da pesquisa) das fontes de informação dos acervos, incluindo a Revista "Vida Capichaba", o Conselho Estadual de Cultura, a Orquestra Sinfônica do Espírito Santo, e o tratamento dos bens e produtos gerados a partir dos editais do Funcultura, que, atualmente, não chegam ao conhecimento da sociedade.

O modelo de arquitetura metodológica elaborado para a condução desse projeto de pesquisa pode ser visualizado na Figura 3. Nele, encontram-se quatro etapas integradas e interdependentes voltadas principalmente à modelagem de metadados dos documentos (físicos ou nato digitais) envolvidos nos acervos e, consequentemente, à criação das coleções destinadas à catalogação, que irão compor os repositórios institucionais previamente criados na primeira fase do projeto.

Em verde, tornou-se necessário o conhecimento da situação real do ambiente informacional das instituições envolvidas em conjunto aos seus acervos, a fim de se obter um diagnóstico de maturidade (INSTITUTO BRASILEIRO DE MUSEUS, 2020) de tal ambiente em relação à política institucional (procedimentos e regras institucionais), pessoas envolvidas, tecnologias e sistemas adotados, e documentação do acervo (formas de gestão e organização da informação).

Figura 3 - Etapas constitutivas do projeto Midiateca Capixaba Conectando

Fonte: elaborado pelos autores (2024)

Em azul, a partir do conhecimento dos acervos, tornou-se possível a realização dos inventários dos itens, cada qual com sua complexidade, para fins de descrição das fontes de informação, levantamento inicial dos metadados, verificação do estado de conservação visando digitalização, e verificação de cópias, versões e possíveis duplicações de itens documentais. O produto resultante de cada inventário de acervo foi um pré-requisito ao processo de digitalização dos itens identificados, para o qual foram usadas diretrizes básicas advindas do Conselho Nacional de Arquivos[18] (CONARQ) para a produção de mídias com qualidade, o que se desdobrou em objetos digitais com variados formatos (PDF, TIFF, JPEG, dentre outros) organizados em pastas específicas na rede do Instituto de Tecnologia da Informação e Comunicação do Estado do Espírito Santo (Prodest) com identificações pertinentes à integridade referencial aos seus respectivos itens de coleções. É importante ressaltar que os produtos resultantes desta etapa se encontram organizados em planilhas de dados para fins de documentação do projeto.

[18] https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/Diretrizes_digitalizacao__2021.pdf

Em laranja, as atividades envolvidas são consideradas sensíveis e cruciais ao projeto Midiateca, talvez em função de envolver abstrações do domínio por parte do profissional da informação responsável pela modelagem e tratamento dos dados. Em adição, a importância de envolver padrões e práticas de catalogação consistentes (regras de catalogação, padrões de metadados e linguagens documentárias) (Lemos; Martins; Carmo, 2022; Lemos; Martins; Souza, 2023) pela instituição de patrimônio cultural envolvida nessa rede interoperável de agregação, possibilita a identificação e a modelagem consistente de informações cruciais e necessárias para descrever um item, de modo a localizá-lo nas bases de dados para fins de busca, recuperação e interoperabilidade.

Deste modo, a primeira atividade incluiu a modelagem do esquema de metadados para cada coleção dimensionada no processo de inventário. Para tal, foram modelados metadados de uso geral e metadados para cada tipologia documental levantada na realidade do acervo. Adicionalmente, foram registrados e destacados os metadados que pudessem ser alinhados (mapeados) com o padrão Dublin Core, modelo central da plataforma de agregação, as regras de cardinalidade, as obrigatoriedades, os filtros de busca, os tipos de dados, se faz uso de vocabulário controlado (se sim, indicar qual), se faz uso de regras de catalogação (se sim, indicar qual), e outras características técnicas. Do mesmo modo como foi feito para os inventários, para cada modelagem produzida, criou-se uma planilha de dados correspondente, a qual passou por uma validação de requisitos com os gestores responsáveis da SECULT.

A segunda atividade envolveu a produção de planilhas de dados, fundamentadas na modelagem do esquema de metadados realizada na primeira atividade, para catalogar inicialmente os itens previamente inventariados. Torna-

se importante destacar que, para cada item catalogado, foi registrado um valor de dado referente ao link de localização da sua respectiva mídia digital, produzida no processo de digitalização.

Já a terceira atividade incluiu o tratamento dos dados (baseado na Figura 1) usando-se de ferramentas de CD (Knime[19], Open Refine e *scripts* em python) para correção ortográfica, limpeza, normalização, criação de arquivos de autoridade, taxonomias, listas de seleção, tratamento das mídias digitais, a partir dos itens de dados registrados nas planilhas de catalogação (produzidas na atividade anterior), o que viabilizou a importação automatizada desses dados nas coleções digitais criadas no Tainacan à luz da modelagem de metadados realizada.

Por fim, em rosa, os dados já se encontram padronizados e normalizados nas bases de dados das coleções dos repositórios institucionais, permitindo, assim, que se dê prosseguimento com o processo de catalogação diretamente no Tainacan. Por conseguinte, os dados de cada provedor já podem ser colhidos e mapeados para a plataforma de agregação Midiateca Capixaba.

5 DISCUSSÃO

A construção de um projeto que objetiva criar repositórios digitais para a organização e a difusão do patrimônio cultural das instituições culturais do estado do Espírito Santo bem como construir um agregador que permita integrar e facilitar a busca, a recuperação e a navegação (INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS, 2016) de forma integrada em todo o acervo cultural do Estado consistiu numa ação com inúmeros desafios e ao mesmo tempo inovadora e pioneira como política pública de cultura e informacional entre os 27 estados brasileiros. Para tal, foi necessário o

[19] https://www.knime.com/

desenvolvimento de ações em várias frentes de atuação, incluindo análise e inventário de dados legados e bases de dados já existentes, processos de digitalização e tratamento de mídias digitais, formação de pessoas, aplicação de procedimentos de CD e CI, visando a melhoria e a adequação da qualidade dos dados, design gráfico, implementação de *softwares* em servidores, e, finalmente, criação de uma infraestrutura informacional (Borgman, 2010) mínima de gestão e governança do projeto.

Nesse ensejo, um dos maiores desafios encontra-se no tratamento dos dados legados da documentação cultural para a construção da solução agregadora de navegação, exploração, busca e recuperação da informação, o que exigiu habilidades em técnicas de ciência de dados e computacionais (Siqueira; Martins; Lemos, 2022) para aprimorar os processos de organização, tratamento e análise de dados nos espaços culturais.

É importante destacar como a área da CI torna-se área de conhecimento fundamental para a convergência dos diferentes saberes, metodologias e recursos técnicos e humanos (Lemos; Martins; Souza, 2023) necessários para ações complexas e multidisciplinares descritas na pesquisa. Apenas uma ciência que entende "informação" como um fenômeno humano e social, que permite a incorporação de diferentes epistemologias em seu tratamento, ora contemplando os aspectos culturais e sociais dos procedimentos históricos da construção, seleção, curadoria e documentação dos acervos das instituições culturais do Estado, e ora contemplando aspectos computacionais e estatísticos para aplicação de técnicas de pré-processamento de dados, desenvolvimento de *scripts* e ações de limpeza, normalização e tratamento de dados (Martins *et al.*, 2022), pode oferecer um espaço de encontro e convergência que permita a

construção de uma política pública no campo dos acervos digitais na área da cultura.

Dentre os inúmeros desafios enfrentados, cabe ressaltar que a formação e a fixação de pessoas nos postos de trabalho dentro das instituições culturais consistem em problema significativo a ser enfrentado no futuro do projeto. Para que se efetive a transferência de conhecimento da universidade pública para os técnicos da área da cultura do Estado, o incentivo à qualificação profissional, para atender aos novos desafios que a gestão de projetos digitais agrega ao trabalho, torna-se fundamental para garantir a boa gestão e a governança do projeto.

6 CONSIDERAÇÕES FINAIS

O projeto Midiateca Capixaba encontra-se no que pode ser chamado de sua terceira fase, qual seja, uma expansão substancial de escala e ampliação de seu campo de atuação. O projeto passará a incluir as instituições culturais dos 78 (setenta e oito) municípios capixabas e a inclusão de acervos culturais que hoje estão em posse de Organizações da Sociedade Civil no Estado. Espera-se com isso ampliar o mapeamento, a difusão e o acesso ao público da cultura capixaba, institucionalizada e mantida pelos diversos atores sociais que passarão a integrar a plataforma Midiateca.

A metodologia desenvolvida para a presente pesquisa, bem como as tecnologias utilizadas, mostraram-se viáveis de serem executadas por equipes de pesquisadores, incluindo estudantes de graduação e pós-graduação, oriundos das áreas de Arquivologia, Biblioteconomia, Museologia e Ciência da Informação.

Adicionalmente, os aspectos teórico-metodológicos apresentados podem ser aplicados em outros projetos e em potenciais políticas públicas voltadas à organização, difusão e acesso ao patrimônio cultural digital. Logo, tanto a questão

quanto o objetivo desta pesquisa foram cumpridos na perspectiva de divulgar um modelo de criação de infraestruturas informacionais endereçadas à disseminação colaborativa de conteúdos em rede para o campo da cultura.

Finalmente, espera-se que a Midiateca Capixaba se torne um projeto inspirador e que estimule outras políticas públicas estaduais e municipais ao redor do Brasil a criarem ações que facilitem a construção de repositórios digitais e tecnologias de agregação, ampliando de forma significativa a possibilidade acesso ao público e também o acesso a algoritmos, mecanismos de busca e inúmeros usos e recursos de dados em potencial que transformem o patrimônio cultural brasileiro em fonte de inovação para a sociedade.

REFERÊNCIAS

BAKER, Thomas; NILSSON, Mikael; JOHNSTON, Pete. **The Singapore Framework for Dublin Core Application Profiles**. DCMI, 2008. Disponível em: http://dublincore.org/documents/singapore-framework/. Acesso em: 10 abr. 2025.

BORGMAN, Christine L. **Scholarship in the digital age**: Information, infrastructure, and the internet. Cambridge, MA: MIT Press, 2010.

EUROPEANA. **Brief History**. [*S.l.*], 2020. Disponível em: https://pro.europeana.eu/about-us/mission#brief-history. Acesso em: 10 abr. 2025.

FREIRE, Klara Martha W.; SALES, Luana Farias; SAYÃO, Luis Fernando. Curadoria digital no contexto artístico e cultural: possibilidades de reuso de dados de arte. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 25, p. 01-21, 2020.

HYVÖNEN, Eero. **Publishing and using cultural heritage linked data on the semantic web**. EUA: Morgan & Claypool Publishers, 2012.

INSTITUTO BRASILEIRO DE MUSEUS. **Acervos digitais nos museus**: manual para realização de projetos. Brasília, DF: Ibram, Universidade Federal de Goiás, 2020.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS (IFLA). **Declaração dos Princípios Internacionais de Catalogação**. Haia, 2016. Disponível em: https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/icp/icp_2016-pt.pdf. Acesso em: 10 abr. 2025.

LANCASTER, Frederick Wilfrid. **Vocabulary control for information retrieval**. [*S.l.*]: Information Resources Press, 1986.

LEMOS, Daniela Lucas da Silva; MARTINS, Dalton Lopes; DO CARMO, Danielle. Padrões de qualidade para dados e metadados endereçados a aplicações em ciência de dados. *In*: DIAS, Thiago Magela Rodrigues (Org.). **Advanced Notes in Information Science**. 1ed. Tallinn, Estonia: ColNes Publishing, 2022, v. 2, p. 161-170.

LEMOS, Daniela Lucas da Silva; MARTINS, Dalton Lopes; SOUZA, Renato Rocha. Organização e representação da informação e do conhecimento em contextos informacionais: uma proposta de um modelo teórico-conceitual para a qualidade de objetos culturais digitais. **Fronteiras da representação do conhecimento**, Belo Horizonte, v. 3, n. 2, 2023.

LEMOS, Daniela Lucas da Silva; SOUZA, Renato Rocha. Ontologies for Semantic Annotation: Proposal for an Ontological Multimedia Reference Model. **Knowledge Organization**, [*s.l.*], v. 51, n. 8, p. 561-581, 2024.

MARCONDES, Carlos Henrique. Interoperabilidade entre acervos digitais de arquivos, bibliotecas e museus: potencialidades das tecnologias de dados abertos interligados. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 21, n. 2, p. 61-83, abr./jun. 2016.

MARCONDES, Carlos Henrique. Um modelo de curadoria para integrar acervos digitais em memória e cultura publicados na web como dados abertos interligados. **Fronteiras da representação do conhecimento**, Belo Horizonte, v. 2, n. 2, p. 26-56, 2022.

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

MARTINS, Dalton Lopes; LEMOS, Daniela Lucas da Silva; DE OLIVEIRA, Luis Felipe Rosa; SIQUEIRA, Joyce; DO CARMO, Danielle; MEDEIROS, Vinicius

Nunes. Information organization and representation in digital cultural heritage in Brazil: Systematic mapping of information infrastructure in digital collections for data science applications. **Journal of the Association for Information Science and Technology**, [*s.l.*], p. asi.24650, 2022.

MIDIATECA. **Acervo Midiateca Capixaba**: plataforma online de difusão e preservação digital de acervos culturais sobre a cultura e a história do Espírito Santo. [*S. l.*], 2024. Disponível em: https://midiateca.es.gov.br/site/. Acesso em: 10 abr. 2025.

PEREIRA, Joanicy Leandra; LEMOS, Daniela Lucas da Silva. Reuso da informação em plataformas de agregação: uma análise utilizando a Revisão Sistemática de Literatura no âmbito da cultura digital. **Ciência da Informação em Revista**, [*s.l.*], v. 11, p. e15829, 2024.

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

SHEARER, Colin. The CRISP-DM model: the new blueprint for data mining. **Journal of data warehousing**, [*s.l.*], v. 5, n. 4, p. 13-22, 2000.

SIQUEIRA, Joyce; MARTINS, Dalton Lopes. Recuperação de informação: descoberta e análise de workflows para agregação de dados do patrimônio cultural. **Ciência da Informação**, Brasília, v. 49, n. 3, p. 97- 114, set./dez. 2020.

SIQUEIRA, Joyce; MARTINS, Dalton Lopes; LEMOS, Daniela Lucas da Silva. Brasiliana Museus: serviço de busca e recuperação da informação agregada dos acervos digitais do Instituto Brasileiro de Museus. *In*: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 2022, 22., Porto Alegre. **Anais** [...]. Porto Alegre: UFRGS, 2022.

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

AGRADECIMENTOS

Agradecemos a FAPES pelo apoio no financiamento desta pesquisa juntamente à Secretaria de Estado da Cultura do Espírito Santo.

tpbci@ancib.org @anciboficial