Título do projeto: Interoperabilidade entre os repositórios digitais do patrimônio cultural brasileiro: da web semântica e dados abertos ligados às ferramentas de busca e recuperação da informação

Nome do Pesquisador Responsável: Dalton Lopes Martins

Instituição(ões) Sede do projeto: Faculdade de Ciência da Informação/ Universidade de Brasília

Equipe de pesquisa, incluindo nomes, qualificações e instituições de vínculo:

Dalton Lopes Martins – pesquisador responsável - Doutor em Ciência da Informação - Faculdade de Ciência da Informação da Universidade de Brasília

Daniela Lucas de Silva Lemos – pesquisadora associada – Doutora em Ciência da Informação - Centro de Ciências Jurídicas e Econômicas da Universidade Federal do Espírito Santo

José Eduardo Santarém Segundo – pesquisador associado – Doutor em Ciência da Informação - Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto/FFCLRP/USP

Luciana Conrado Martins – pesquisadora associada – Doutora em Educação - Percebe Pesquisa, Consultoria e Treinamento Educacional S/C Ltda. - ME/PERCEBE.

Número do Processo FAPESP: 2018/23068-3

Período de vigência do projeto: 01/12/2019 a 30/11/2021

Período coberto pelo Relatório Científico em questão: 01/12/2019 a 30/11/2020.

Assinatura do Pesquisador Responsável:

Resumo do projeto proposto

O presente projeto tem por objetivo desenvolver, de forma piloto, um serviço de agregação de dados de acervos digitais das instituições de memória do patrimônio cultural brasileiro no âmbito federal. Foca-se, sobretudo, nas instituições ligadas a atual Secretaria da Cultura lotada no Ministério da Cultural. As instituições de objeto de pesquisa são a Agência Nacional de Cinema (ANCINE), Fundação Biblioteca Nacional, Fundação Casa de Rui Barbosa, Fundação Cultural Palmares, Fundação Nacional das Artes, Instituto Brasileiro de Museus e Instituto do Patrimônio Histórico e Artístico Nacional. Considera-se que estão entre as mais importantes instituições culturais brasileiras, seja tanto no âmbito do exercício e influência na política cultural do país quanto na custódia e preservação de acervos culturais dos mais distintos tipos e categorias. O projeto parte da constatação da inexistência de um serviço de informação que agrega e disponibiliza de maneira centralizada as informações e, quando disponível, imagens e arquivos dos objetos culturais dessas instituições completos para busca, recuperação e exploração pelo usuário brasileiro. É também premissa do projeto que a inexistência de tal serviço dificulta o acesso, o reuso e a própria valorização da cultura e identidade cultural brasileira no âmbito da Internet, o que representa uma perda significativa de potencial de socialização e circulação da informação, além do potencial de inovação e criatividade que tal informação pode agregar a sociedade. Sabe-se que as instituições objeto desta pesquisa possuem diferentes projetos de digitalização e publicação de acervos digitais na Internet, valendo-se de diferentes estratégias, tecnologias, padrões e práticas de gestão da informação cultural. No entanto, até o presente momento, não se tem um diagnóstico preciso e a partir de critérios sistemáticos aportados pela área da Ciência da Informação do estágio de organização e adoção de boas práticas desses acervos publicados e nem se os mesmos se encontram em condições de serem coletados e eventualmente fazerem parte de um projeto de agregação, busca e recuperação.

O presente projeto parte desse quadro inicial e propõe desenvolver de forma piloto um serviço de agregação dos acervos e, por meio desse trabalho técnico e conceitual, realizar um diagnóstico que produza insumos analíticos para o desenvolvimento e implantação de tal serviço pelo poder público. Para tal, o projeto parte de uma **etapa inicial** onde se propõe, a partir de metodologia criada especificamente para a realização do projeto, a **analisar a situação informacional dos acervos digitais** já publicados e disponíveis para acesso público na Internet das instituições participantes da pesquisa. Na sequência, o projeto propõe uma **coleta amostral dos dados** para identificar os desafios técnicos e informacionais para a coleta dos dados existentes e fornecer elementos de como esses acervos estão estruturados

para o planejamento de coleta da população dos itens disponíveis e proposição de modelo conceitual semântico para interoperabilidade dos dados. Uma vez realizada a coleta amostral, o projeto se dedica a **análise e proposição de modelo conceitual semântico** visando a melhor interoperabilidade dos acervos. Para tal, é fundamental conhecer os dados e metadados
existentes, sua estrutura, o uso de linguagens documentárias, a aplicação de regras de catalogação, entre outros instrumentos de boas práticas de gestão da informação. Com base nesse resultado, espera-se identificar um modelo mínimo conceitual que atenda ao padrão de qualidade dos dados encontrados e permita a proposição de um nível mínimo de interoperabilidade semântica entre os acervos. Por fim, o projeto propõe uma **coleta extensiva dos dados dos acervos** e a implementação de uma **interface de busca e recuperação da informação**.

Espera-se, como contribuição do projeto, identificar os passos técnicos e conceituais que a realização de um serviço de informação de agregação, busca, recuperação e exploração dos acervos culturais digitalizados exija, ao mesmo tempo identificando recomendações, melhorias, sugestão de adoção de boas práticas, atualizações tecnológicas, migrações, entre outros elementos que podem melhorar substancialmente a qualidade dos serviços e da informação dos acervos culturais brasileiros disponíveis em rede.

Em termos de **contribuições acadêmicas**, o projeto tem por objetivo fornecer um **diagnóstico** de como se encontram, em termos das práticas e estratégias de gestão da informação, os acervos digitais já disponibilizados pelas instituições culturais objeto da pesquisa. A segunda contribuição do projeto consiste na **avaliação de diferentes modelos semânticos** de representação da informação cultural que têm sido utilizados e desenvolvidos mundo afora
e, com base nessa análise, propor um modelo que melhor se adapte a realidade informacional brasileira. A terceira contribuição do projeto consiste na proposição ou adoção de um **modelo conceitual** mínimo que permita a interoperabilidade semântica entre os acervos identificados. A quarta contribuição do projeto consiste no **modelo computacional** adotado para as fases da coleta dos dados, a criação de um índice de busca integrado, o motor de busca e recuperação além da interface gráfica de acesso aos dados e serviços oferecidos para o usuário. Entende-se que um considerável desafio tecnológico e informacional no desenho de tal solução e que os resultados alcançados podem colaborar de forma significativa para o desenvolvimento de projetos e futuras políticas voltadas para a promoção de acervos digitais culturais na Internet.

Realizações no período

As realizações no período serão apresentadas conforme as fases descritas no cronograma original do projeto, listadas em sequência a seguir.

Fase 1 – Duração: 2 meses - Análise da informação dos repositórios digitais: consiste na identificação dos modelos de dados dos repositórios, dos tipos de metadados utilizados, da tipologia de acervo, da estrutura organizacional do acervo, das linguagens documentárias utilizadas, regras de catalogação, tipologia de documentos, mídias e formato de coleta dos

dados.

A presente pesquisa foi classificada como sendo de natureza teórica e aplicada, quali quantitativa, e de cunho exploratório e descritivo, envolvendo acervos culturais localizados na Web. Como procedimento técnico para coleta e análise dos dados, utilizou-se de pesquisa bibliográfica e documental. A primeira no intuito de fundamentar conceitos e fornecer sustentabilidade teórica ao estudo, e para revisão na literatura, de modo a visualizar o panorama atual sobre o problema de pesquisa. Para tal, procedeu-se a um levantamento nos campos das Humanidades Digitais, Ciência da Informação e Ciência da Computação que sustentaram as discussões teóricas e empíricas do estudo. E a segunda no intuito de acessar (a partir de endereços eletrônicos pré-identificados) fontes documentais institucionais, incluindo bases de dados, manuais, tutoriais, normas técnicas, linguagens documentárias, regras, normas e instruções de catalogação, modelos conceituais, tabelas de classificação, dentre outras.

Os acervos culturais determinados como objetos empíricos de pesquisa são oriundos de instituições vinculadas à Secretaria Especial da Cultura (2013c), elencada especialmente por sua relevância e abrangência nacional, a saber: o Instituto do Patrimônio Histórico e Artístico Nacional – Iphan; o Instituto Brasileiro de Museus – Ibram; a Agência Nacional do Cinema – Ancine; a Fundação Casa de Rui Barbosa – FCRB; a Fundação Cultural Palmares – FCP; a Fundação Nacional das Artes – Funarte; e a Fundação Biblioteca Nacional - FBN.

A etapa de coleta e análise dos dados contou com o método de pesquisa conhecido como análise de conteúdo (Bardin, 2016), que adota como estratégia um conjunto de métodos e técnicas com abordagens quali-quantitativas visando a criação de categorias de análise para compreensão mais abrangente do fenômeno investigado de modo a facilitar a extração, a análise e a interpretação mais concisa dos dados presentes nos materiais elencados no estudo. Desse modo, a pesquisa se organizou de forma cronológica em torno de três fases: i) pré-análise; ii) exploração do material; e iii) tratamento dos resultados, a inferência e a interpretação.

O método foi escolhido pelo entendimento de que a pesquisa consiste na identificação da presença específica de categorias analíticas em documentos, no caso sítios web, que permitam identificar a presença ou não de elementos estruturantes da organização e representação dos acervos digitais disponibilizados publicamente pelas instituições analisadas. Como se sabe, o método da análise de conteúdo permite compreender uma massa documental por meio de regras específicas analíticas que facilitem a identificação de regularidades e uma ordem objetiva na presença ou ausência das categorias analíticas.

> "(...) é método das categorias, espécie de gavetas ou rubricas significativas que permitem a classificação dos elementos de significação constitutivos da mensagem. É, portanto, um método taxonômico bem concebido para satisfazer os colecionadores preocupados em introduzir uma ordem, segundo critérios, na desordem aparente" (BARDIN, 2016, pg. 43).

- A pré-análise: definição do corpus central e das categorias de análise

A fase inicial partiu do objetivo da pesquisa, que foi a base para a construção do instrumento de coleta de dados. Consistiu em três ações baseadas na construção do corpus central da pesquisa (elementos que a análise vai fazer falar), na identificação das hipóteses e objetivos da pesquisa e na elaboração dos indicadores de análise que permitam a interpretação dos resultados. Os indicadores se tornaram unidades comparáveis de categorização para análise temática a partir do contato inicial com fontes bibliográficas e documentais para a obtenção de impressões e orientações acerca de seus conteúdos.

A primeira ação, executada de forma manual, se deu a partir da identificação dos endereços eletrônicos das instituições vinculadas à Secretaria Especial da Cultura. A partir desses endereços, tornou-se possível explorar o ambiente digital e identificar fontes documentais em potencial inerentes aos conjuntos de itens presentes nos acervos de cada instituição vinculada (Tabela 1). Torna-se válido ressaltar que a pesquisa considerou acervo como todo conjunto de itens (ou documentos) organizado em possíveis tipos de SRIs, tais como repositórios e bibliotecas digitais, sistemas de gestão de conteúdo e itens sistematizados em páginas web dispersas. Para este último, considerou que potencialmente poderiam fazer parte do acervo da instituição vinculada e que, em uma possível agregação e busca integrada, seria interessante que houvesse a recuperação desses itens. Sobretudo, deseja-se conhecer aqui os objetos culturais que já foram digitalizados e estão disponibilizados ao público de alguma maneira acessível pela web, sendo que foram eliminados da pesquisa documentos administrativos ou ligados a gestão da instituição.

Na tabela 1, a seguir, apresenta-se as os sítios web de cada instituição cultural e as fontes documentais que foram identificadas nessa etapa de pré-análise e que serão objeto da análise de conteúdo proposta na presente pesquisa. Na coluna "fontes documentais", pode-se observar o total de endereços web que representam conjuntos de objetos culturais publicados na web para cada instituição. Vale ressaltar aqui que um endereço web pode ser um endereço de um repositório digital que agrega milhares de objetos culturais ou apenas uma página web com uma dezena de objetos culturais disponíveis para visitação como imagem. O foco da pesquisa é exatamente analisar essa variabilidade nas estratégias de publicação de objetos culturais e como isso pode impactar na ideia de construção de uma ferramenta de busca e recuperação da informação de forma integrada.

Tabela 1 - *Corpus* central da pesquisa - fontes documentais por instituição vinculada

Instituição Vinculada	Endereço eletrônico (*link* principal)	Nº Fontes documentais
Agência Nacional do Cinema	https://www.ancine.gov.br/	3
Fundação Biblioteca Nacional	https://www.bn.gov.br/	38
Fundação Casa de Rui Barbosa	http://www.casaruibarbosa.gov.br/	22
Fundação Cultural Palmares	http://www.palmares.gov.br/	1
Fundação Nacional das Artes	https://www.funarte.gov.br/	10
Instituto Brasileiro de Museus	https://www.museus.gov.br/museus-ibram/	28
Instituto do Patrimônio Histórico e Artístico Nacional	http://portal.iphan.gov.br/	113

A segunda ação da etapa de pré-análise consiste na identificação das hipóteses e objetivos da pesquisa. Tem-se por hipótese, baseado na experiência dos pesquisadores em lidar cotidianamente com as instituições culturais brasileiras, que os acervos culturais digitalizados são publicados na web de forma ainda bastante precária, seja tanto em termos das tecnologias para acesso e navegação nos documentos quanto nas práticas de gestão da informação estabelecidas, o que envolve os usos de padrões de metadados, vocabulários controlados e regras de catalogação, para citar alguns exemplos. Logo, entende-se que a hipótese estabelecida para esses acervos, apesar de digitalizados, é de que estão disponíveis de forma a dificultar sua agregação, busca e recuperação pelos seus usuários em potencial. Desse modo, tem-se como objetivo da presente pesquisa apresentar um mapeamento sistemático das formas de organização e representação da informação que estão sendo aplicadas aos dados pertencentes aos acervos das instituições vinculadas à Secretaria Especial da Cultura.

A terceira ação se baseou em princípios advogados por Bardin (2016) quanto ao uso de categorias para procedimentos de análise qualitativa e quantitativa. Nesse sentido, o procedimento de categorização se orientou em questões fundamentais com garantia literária em áreas ligadas aos campos da informação (ex.: política de qualidade de dados; organização e representação da informação) e da tecnologia (ex.: sistemas de informação; agregação de dados), a saber: i) quais os tipos de sistemas de informação utilizados por essas instituições?

ii) quais os recursos tecnológicos utilizados para a publicação dos acervos? iii) quais licenças,

regras de catalogação, padrões de metadados e linguagens documentárias são explicitados na divulgação do acervo? iv) quais as formas de exposição dos itens dos acervos? v) quais as formas de extração dos dados do acervo? vi) em qual formato os itens do acervo estão disponibilizados? e vii) qual o tamanho do acervo?

A construção das categorias levou em consideração o objetivo da pesquisa e seu potencial uso estratégico como desdobramento dos resultados atuais, a saber, que esses dados sirvam de insumo para a construção de um de serviço de informação para coleta, agregação, busca e recuperação dos objetos culturais brasileiros de forma integrada. Ou seja, para além dos elementos analíticos já descritos aqui, são de interesse da pesquisa compreender qual o tamanho em termos quantitativos dos objetos culturais digitais já disponibilizados na web pelas instituições pesquisadas, de que forma esses acervos podem ser visualizados, os tipos de mídia que estão disponíveis, as possíveis formas de extração de dados, quando existirem, além do tipo de sistema de recuperação da informação que está sendo utilizado. Entende-se que esses fatores permitirão um diagnóstico situacional dos acervos digitais das instituições, permitindo que os resultados possam ser usados para refletir qual a melhor maneira de se aproveitar esses objetos digitais para um serviço de informação que tenha por objetivo agregar os acervos. Perguntas relativas a se os softwares atuais devem ser migrados para outras soluções ou não, uma estimativa inicial sobre o tamanho dos servidores web para hospedar os objetos digitais, a possibilidade de reaproveitar esses objetos ou não, entre outras, podem ser analisadas em pesquisas posteriores com os resultados aqui apresentados.

A partir de tais questões e usando-se do critério semântico para alinhamento temático, as categorias de análise foram definidas e são descritas como se segue.

Tipo de SRI: visa especificar o tipo de sistema de recuperação da informação com foco a destacar de que maneira o objeto digital é acessível. Foram usadas as categorias página estática (HTML), quando os objetos digitais foram identificados dentro de uma página web como imagens, vídeos ou documentos textuais e não se identificou um sistema específico para a geração desta página web, sistema de repositório digital, quando se identificou qual sistema específico com características de repositório digital (sistemas que apresentaram organização por coleções, metadados explícitos, navegação facetada e campo de busca específico) era utilizado para dar acesso aos objetos digitais, sistema de gerenciamento de conteúdo (CMS), quando se identificou que havia um sistema de gerenciamento de conteúdo que gerava as páginas web nas quais os objetos foram identificados, mas que não possuía a estrutura de um repositório digital , arquivo de dados, quando se identificou que os objetos digitais estavam inseridos dentro de arquivos de dados, como por exemplo, um conjunto de fotografias disponível em um catálogo em PDF. Software utilizado: a tecnologia utilizada pelo sistema de recuperação da informação. Foram propostas as categorias sistema de gerenciamento de conteúdo, repositório digital, arquivo de dados, página estática html e não identificados. Licença sobre os dados: a licença de publicação dos objetos digitais. Organização e representação da informação: nessa categoria foram considerados, quando identificados, os padrões de metadados, as linguagens documentárias e as regras de catalogação utilizadas pelos acervos. Visualização do acervo: constitui da forma como o

conjunto de itens é apresentado no sítio web das entidades vinculadas, foram identificadas 3 categorias: Coleções, em que os itens estão claramente organizados em coleções; Exposição, em que os itens estão organizados no formato de exposições, onde se percebe que apenas alguns itens são exibidos com informações contextuais de apoio (páginas HTML com parte do acervo por exemplo); Hierárquica, em que os itens estão organizados de forma hierárquica (como em listas de registros ou vídeos, ou ainda conjuntos de pastas por exemplo); Extração dos dados: formas de exportação dos dados dos sistemas de informação. Tipo de mídia: o formato das mídias, se representam texto, imagem, vídeo ou áudio. Quantidade de itens no acervo: o total de itens identificados em cada sítio web

- A exploração do material

A segunda fase partiu das categorias determinadas na fase anterior, para as quais se somaram as inferências realizadas ao corpus central da pesquisa, que culminou num segundo instrumento de coleta de dados destinado a descrever o conteúdo de cada fonte documental à luz das categorias previamente determinadas. Desse modo, os dados obtidos a partir dessa descrição foram revisados e normalizados a partir da revisitação de todas as fontes documentais, verificando a adequação a cada categoria de análise. Novas categorias foram então identificadas e agregadas ao processo de análise, garantindo, assim, a sistematização coerente dos dados referente aos acervos de cada instituição investigada.

- O tratamento dos resultados

A última fase marcou o desenvolvimento da síntese gráfica de vários resultados quantitativos obtidos a partir do instrumento de coleta. A partir do tratamento e da organização dos dados, tornou-se possível obter um corpus teórico conceitual frente aos objetos investigados e tirar conclusões teóricas e metodológicas acerca do processo de construção e publicação de acervos digitais em rede pelas instituições de cultura brasileiras.

Apresentamos, a seguir, alguns resultados dos dados coletados.

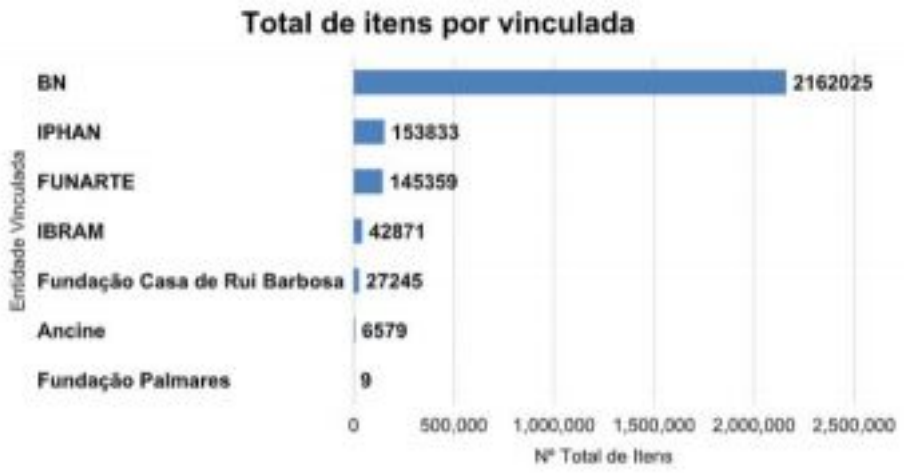


Figura 01. Total de itens identificados por instituição cultural vinculada

Figura 02. Distribuição relativa dos links de acervo por estrutura de sistema de informação

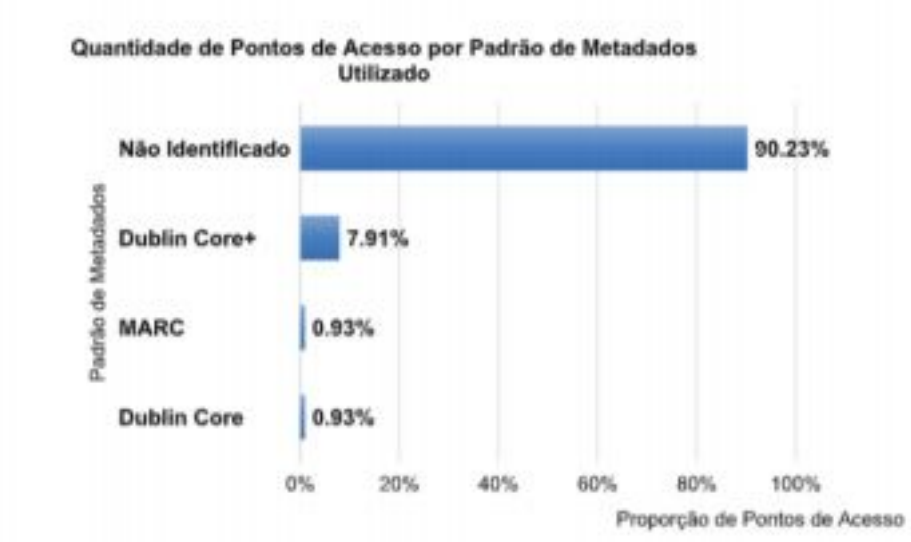


Figura 03. Distribuição relativa dos links de acervo por padrão de metadados

Figura 04. Distribuição relativa dos links de acervo por formas de extração de dados

Um relatório completo detalhado sobre os dados coletados está disponível em:

https://unbbr
my.sharepoint.com/:w:/g/personal/daltonmartins_unb_br/EbOT8tFg94ZCjegMQVcmDNMB urmbD33GimWCWW4dutaq-A?e=vYehST

De forma geral os resultados encontrados apontam algumas questões importantes:

1. dificuldade de se encontrar documentação sobre os acervos das vinculadas. Como apresentam os resultados analisados das variáveis de Linguagens Documentárias, Regras de Catalogação, Padrão de Metadados e Licenças, o que significa um potencial comprometimento da interpretação da organização da informação desses acervos. Para agregar os dados é fundamental o esclarecimento da documentação dos pontos de acesso, para entender a correspondência conceitual dos metadados utilizados entre as entidades vinculadas por exemplo.
2. grande quantidade de itens sem condição de coleta formal, ou seja, poucos pontos de acesso 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 (Web Scraping) que consistem em identificar como os dados estão estruturados através de suas páginas web e desenvolver scripts de coleta para cada modo de disponibilização dos objetos.

Fase 2 – Duração: 4 meses - Coleta amostral de dados: consiste na coleta de uma amostra de dados para formação de uma base de dados relacional de controle do projeto para testes e avaliação da informação

Denominado “Coleta Amostral dos Dados das Entidades Vinculadas à Secretaria Especial de Cultura”, esta etapa prevê o estudo e o desenvolvimento de formas de acesso e recuperação dos dados dos acervos identificados nos portais web das entidades vinculadas. Como referência, a primeira fase intitulada “Mapeamento sistemático das formas de organização da informação dos repositórios digitais de cultura”, cita o processo de identificação dos links de potenciais acervos dos portais web das entidades vinculadas, links dos quais foram selecionados uma amostra para a coleta de dados.

Como objetivo geral a pesquisa realizada nesta etapa do projeto busca identificar a viabilidade da coleta de dados dos acervos das entidades vinculadas à secretaria especial da cultura através da coleta de uma amostra dos acervos, uma vez que os acervos potenciais dessas entidades vinculadas foram mapeados na etapa anterior do projeto.

Como objetivos específicos espera-se, 1 - Prospectar a viabilidade da coleta de dados dos acervos das entidades vinculadas através das tecnologias de coleta de dados mapeados em cada caso; 2 - Evidenciar as eventuais dificuldades encontradas no processo de coleta de dados, e as estratégias de coleta mais expressivas de modo a dar indícios da complexidade de realizar uma coleta de todos os acervos mapeados;
Como resultados dessa coleta amostral espera-se conhecer melhor quais as limitações existentes para acesso aos dados, bem como propor encaminhamentos para uma estratégia de acesso, recuperação e reuso almejando a entender caminhos para a interoperabilidade

dos acervos. Com o diagnóstico dos metadados disponíveis resultantes da coleta será possível ainda subsidiar o próximo produto previsto no projeto, que é a proposta de um padrão conceitual de metadados.

É importante ressaltar que este produto tem uma perspectiva metodológica exploratória, onde se espera compreender a atual situação dos acervos identificados no primeiro relatório e identificar o potencial de coleta e reuso desses objetos digitais já disponíveis para novas estratégias de agregação e interoperabilidade entre as diferentes organizações culturais envolvidas na presente pesquisa. Parte-se do princípio, a ser validado ao longo deste projeto, de que já existiu um grande esforço de digitalização e publicação na web de objetos digitais culturais e que seria potencialmente proveitoso a sociedade a sua reutilização a partir da aplicação de princípios de organização e representação da informação. Logo, seria possível ampliar o acesso e a reutilização de objetos culturais sem ter que necessariamente digitalizar os objetos, mas sim reaproveitando e reestruturando aquilo que já foi feito pelas organizações culturais.

Para atender ao objetivo de identificar a viabilidade da coleta de dados dos acervos das entidades vinculadas à secretaria especial da cultura, essa pesquisa aplica uma metodologia dividida em duas etapas principais: a primeira etapa envolve a seleção dos links dos acervos para compor a amostra e a segunda etapa indica as diretrizes de coleta dos links selecionados.

- Seleção de uma amostra dos links de acervos das entidades vinculadas à secretaria especial de cultura

O trabalho proposto para essa segunda etapa projeto é uma coleta de parte dos acervos mapeados na primeira etapa, desse modo vale relembrar que o mapeamento realizado anteriormente identificou ao todo 217 links que redirecionam para potenciais acervos das entidades vinculadas à secretaria especial de cultura: FUNARTE, Ancine, Fundação Palmares, Biblioteca Nacional, Fundação Casa de Rui Barbosa e IBRAM.

Desses links mapeados, para o atual momento do projeto foi selecionada uma amostra que de acordo com a representatividade dos acervos, levando em conta àqueles acervos que atendem aos seguintes critérios:

- quantidade de itens, em que foram selecionados para a amostra links que representam a maior parte dos itens publicados pela instituição em seu site; • forma de coleta de dados, em que foram selecionados para amostra links cujas formas de coleta de dados contemplem a maior variação possível, possibilitando que os testes de coleta aqui realizados permitam generalizar as características de coleta para os demais links mapeados;
- ferramenta utilizada, também foi levada em conta a variabilidade das ferramentas utilizadas para publicação do acervo, dado que ferramentas como SophiA, DSpace e Tainacan são por exemplo bastante difundidas entre as entidades analisadas, ao mesmo tempo que o Flickr e o Fotoweb 7 aparecem em casos específicos. Assim essa amostra buscou representar essa variabilidade considerando as proporções de

representatividade de cada ferramenta.

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 construída uma amostra com 16 links para acervos das entidades vinculadas, como apresenta a tabela 2 abaixo.

Tabela 2 - Amostra de links de acervos selecionados para coleta dos dados

Entidade Vinculada	URL	Ferramenta	Extração dos Dados
Biblioteca Nacional	http://bndigital.bn.gov.br/acervodigital/	SophiA	Raspagem de Dados
Biblioteca Nacional	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
Fundação Casa de Rui Barbosa	http://rubi.casaruibarbosa.gov.br/	DSpace	OAI-PHM
Fundação Casa de Rui Barbosa	http://iconografia.casaruibarbosa.gov.br/fotoweb/default.fwx	Fotoweb 7	Raspagem de Dados
Fundação Casa de Rui Barbosa	http://acervos.casaruibarbosa.gov.br/index.html	SophiA	Raspagem de Dados
Fundação Cultural Palmares	https://www.flickr.com/photos/culturanegra/	Flickr	Raspagem de Dados
Fundação Cultural Palmares	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.gov.br/	WordPress + Tainacan	API
IPHAN	http://acervodigital.iphan.gov.br/xmlui/	DSpace	Raspagem de Dados
IPHAN	https://pergamum.iphan.gov.br/biblioteca/index.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: Dados da pesquisa, 2020.

Assim, com essa amostra de 16 links selecionados, conclui-se o primeiro processo que permite o encaminhamento da próxima etapa prevista, a coleta dos dados dos acervos disponíveis nesses links, de modo que o tópico abordado a seguir apresenta as diretrizes de coleta dos dados desses acervos selecionados.

Coleta dos dados dos acervos selecionados

Relembrando o processo realizado na etapa anterior deste projeto, ao levantar os links para acervos nos sites das entidades, esses links foram diagnosticados na perspectiva de descrição das características dos acervos. Uma das características importantes para a atual etapa foi a
forma de coleta dos dados, essa informação (presente na coluna “Extração dos dados” da

tabela 1) guiará a forma como os dados serão coletados para cada link de acervo selecionado.

Vale ressaltar que ao se aprofundar na técnica de coleta dos dados este estudo pode revelar outras formas de extração de dados, como por exemplo repositórios que permitem a exportação de dados através de mais de um formato. Quando isso ocorrer as características das diferentes formas de coleta serão descritas para cada link selecionado na amostra.

Assim, a metodologia para coleta aplicada neste estudo envolve de maneira generalizada três grandes etapas, também apresentadas na figura 5 abaixo:

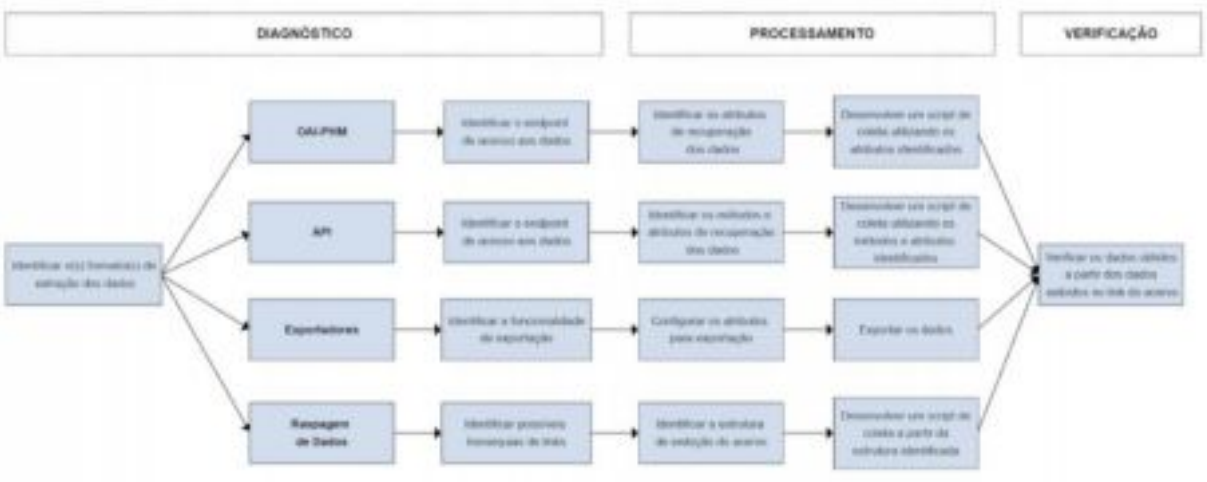


Figura 05. Etapas da metodologia de coleta.

- Diagnóstico, que compõe as atividades de identificação da forma de coleta de dados, e ainda a identificação dos pontos de coleta dos dados: no caso de OAI-PMH e API identificar o(s) link(s) do *endpoint* de acesso; para exportadores, identificar a funcionalidade de exportação; e para raspagem de dados identificar possíveis hierarquias de links.
- Processamento, que envolve o processamento dos dados, quanto à identificação dos metadados, valores e objetos, bem como os métodos e atributos disponíveis nos diferentes formatos de coleta dos dados, no caso de raspagem de dados essa etapa contempla a identificação da estrutura de exibição do acervo.
    - Ainda no contexto de processamento são desenvolvidos scripts de coleta com base nos procedimentos de diagnóstico executados para os formatos de coleta através de OAI-PMH, API e Raspagem de Dados, para os exportadores são utilizadas as funcionalidades de exportação através da interface web do próprio acervo.
    - Vale ressaltar que para os acervos obtidos através da raspagem de dados, foram coletados um máximo de 100 registros para cada link selecionado na amostra. Isso se deve pelo fato da onerosidade de tempo para coleta de itens através da raspagem de dados, já que é necessário requisitar cada página onde os registros do acervo aparecem, e o tempo de resposta de requisição varia podendo levar entre segundos a minutos para recuperar um único registro,

dessa forma a título de otimização do tempo de coleta para a produção deste relatório a coleta de registros foi limitada ao máximo de 100 registros.
    - Outro ponto em comum das atividades de processamento realizadas, com exceção à coleta por exportadores, foi o uso da linguagem de programação Python para obter os e armazenar os resultados, além dessa linguagem se apresentar de forma robusta para a automatização de aplicações que envolvam o ambiente web, é uma expertise já desenvolvida pelo contexto do grupo de pesquisa que desenvolve este projeto.
    - Ainda derivado da linguagem Python, o também é comum entre os processos de coleta que necessitaram de desenvolvimento de scripts o uso da biblioteca Pandas, que é utilizada para análise de dados, e para este projeto foi muito utilizada no contexto de armazenamento dos dados utilizando o método DataFrame, que permite estruturar os dados em um formato bi-dimensional de colunas e linhas, possibilitando também sua exportação para diversos outros formatos, como por exemplo, CSV, SQL (Single Query Language), Excel, entre outros.
- Verificação, uma vez que os dados foram coletados, foi realizado um processo de verificação dos dados para cada saída, de forma que 20 itens escolhidos de forma aleatória da coleta obtida foram comparados com os dados apresentados no acervo on-line.

Para os dados obtidos através de scripts (OAI-PMH, API e raspagem de dados), o processo de validação dos dados, ocorreu utilizando o método sample da biblioteca pandas (aproveitando que os mesmos já se encontram no formato DataFrame), para apresentar os dados armazenados de 20 registros aleatórios, que foram buscados na interface web do acervo e confirmada a correspondência dos dados. O mesmo processo pode ser aplicado para arquivos exportados, ao se desenvolver um script para leitura dos arquivos e apresentar os registros aleatórios.

Um relatório detalhado com os resultados quantitativos dessa etapa se encontra aqui: https://unbbr my.sharepoint.com/:w:/g/personal/daltonmartins_unb_br/EcIOa7y2hC9EvwOjxJze0BIBzFrk NyVIT6XIb0ss38HA4w?e=6oSMfb

Foram produzidos 15 scripts para a realização das atividades técnicas dessa etapa que se encontram aqui: https://github.com/tainacan/data_science/tree/master/FAPESP/coleta_amostral

Os dados coletados se encontram aqui: https://drive.google.com/drive/folders/1_7rLpmpecR0QfplI3xOcTKUI301EW28b.

Fase 3 – Duração: 6 meses - Análise de modelos conceituais semânticos para interoperabilidade entre os acervos: consiste na análise dos diferentes modelos conceituais e ontologias mapeadas como referências para o projeto para escolha do modelo a ser utilizado. Inclui a proposta de mapeamento semântico dos modelos de dados dos repositórios

para o modelo conceitual adotado. Esta etapa utilizará os dados amostrais coletados da etapa anterior para teste e validação do modelo

Essa etapa encontra-se em execução, ainda não finalizada. A etapa foi afetada em sua dinâmica de execução pela necessidade de reformulação das estratégias de encontro e reuniões online pela equipe do projeto por conta da pandemia. É etapa para a qual não foi destinada um bolsista específico e teria de ser construída pela equipe de pesquisadores do projeto. Inicialmente, previámos uma série de encontros presenciais para a discussão do tema, realização de seminário interno e revisão bibliográfica em conjunto. Vale frisar que, para a boa realização dessa etapa, era necessário concluir a coleta dos dados e análise do que foi identificado pelos acervos culturais. Conforme já mencionado, os dados e os modelos conceituais identificados na documentação existente dos acervos culturais é bastante pobre em termos de representação, adoção de modelos de metadados, de vocabulários controlados e regras de catalogação. A pesquisa aponta, algo ainda a ser aprofundado em discussão em andamento, para a adoção do modelo de metadados Dublin Core simplificado, sem aplicação de vocabulários controlados específicos e regras de catalogação padronizadas para a junção mínima e mapeamento de todos os dados coletados para o padrão.

Vale ressaltar que, ao longo do período dessa fase, foram estudados os seguintes padrões identificados na literatura científica da área como sendo os mais utilizados e recomendados para conteúdos digitais culturais de instituições de memória:

- Padrões para estrutura de dados: conjunto de elementos de metadados ou esquemas de categorias que formam um registro de informação (e.g.; MARC; Dublin Core; VRA Core; LIDO; MPEG-7).
- Padrões para valores dos dados: linguagens documentárias, vocabulários e ontologias de domínio usadas para preencher os dados nos elementos de metadados (e.g.: Library of Congress Subject Heading; Union List of Artist Names - ULAN; CRMdig).
- Padrões para conteúdo dos dados: regras e códigos de catalogação que podem ser orientados por modelos conceituais (e.g: FRBR; FRAD; FRSAD; EDM; CIDOC-CRM; Linked.art) em formatações, sintaxes e relacionamentos para os valores de dados usados para preencher os elementos de metadados (e.g.: RDA; Cataloging Cultural Objects - CCO; Descriptive Cataloging of Rare Materials - DCRM).
- Padrões para comunicação de dados: padrões de metadados expressados em uma linguagem de representação legível para a máquina (e.g.: MARC21; Dublin Core RDF/XML; LIDO XML Schema; VRA Core 4.0 XML).

Espera-se concluir essa etapa de mapeamento e análise do modelo no início de 2021 para o prosseguimento das próximas etapas.

Descrição e avaliação do apoio institucional recebido no período

Durante o período do projeto, vale ressaltar que a Universidade de Brasília entrou em processo de suspensão de suas atividades presenciais por conta da pandemia do

Coronavírus.

Desse modo toda possibilidade uso da infraestrutura institucional ficou comprometida, considerando laboratórios de informática, apoio técnico, suporte logístico e organizacional.

Plano de atividades para o próximo período

Para o próximo período do projeto, de dezembro de 2020 a novembro de 2021, o plano de atividades consiste:

- **Dezembro de 2020 a Janeiro de 2021 –** concluir a fase 3, finalizando a análise dos modelos de dados encontrados e o mapeamento para o modelo semântico mínimo identificado.
- **Dezembro de 2020 a Fevereiro de 2021** - Fase 4 **–** Duração: 3 meses - Coleta dos dados e implementação da base de dados semântica: consiste na coleta massiva dos dados disponíveis em cada repositório digital, no mapeamento e curadoria desses dados para o modelo semântico adotado, bem como na escolha da tecnologia de base de dados de grafos a ser utilizada pelo projeto. Essa etapa usará os scripts já produzidos na fase 2 do projeto, simplificando a etapa técnica de desenvolvimento de programação. Além disso, considerando que ainda não temos acesso a infraestrutura presencial da UnB, foi solicitado mudança no plano financeiro do projeto, para remanejar parte do recurso que estava destinado a logística para compra de um computador de alta performance para realizar a etapa de coleta e acesso aos dados, considerando que devem ser coletados por volta de 3 milhões de registros nessa fase. O computador encontra-se em fase de compra e deve estar disponível para uso no início de janeiro de 2021.
- **Março de 2021 a Agosto de 2021** - Fase 5 – Duração: 6 meses Desenvolvimento e customização do repositório digital para busca integrada: consiste na customização de uma solução livre de repositório digital, preferencialmente a solução Tainacan, para dar acesso a base de dados de grafos semânticos contendo os dados agregados dos repositórios digitais coletados. Envolve a modelagem do design gráfico da solução, a implementação das técnicas de busca, análise de desempenho das consultas a banco de dados e dos diferentes modos de exibição dos dados, além do teste de usabilidade do sistema com usuários.
- **Setembro de 2021 a Novembro de 2021 - Fase 6 – Duração: 3 meses** - Produção de relatórios e artigos científicos do projeto: consiste na documentação sistematizada em relatórios técnicos e artigos científicos relatando os principais resultados e descobertas realizadas no projeto

Participação em evento científico

CERTIFICADO

Certificamos que Dalton Lopes Martins
participou na qualidade de convidado da Live Difusão Digital: Ibram, Tainacan e Acervos em Rede,
realizada no dia 04 de junho de 2020 pelo Instituto Brasileiro de Museus - Ibram,
ação que integra o Projeto Compartilhando.

Presidente do Ibram

ComPartilhando

Realização: ibram

MUSEU [redacted]
AFRO-BRASIL-SUL

design
escola
arte

DECLARAÇÃO

Declaro para os devidos fins que o(a). professor **Dalton Martins**, CPF: 26465082802, participou como **Conferencista** no **Webinar N. 6 "O DESAFIO DO USO DE NOVAS TECNOLOGIAS NA PROMOÇÃO E CONSTRUÇÃO DE NOVOS SABERES"** realizado no dia **27 de agosto de 2020 às 19 horas por web conferência,** totalizando **carga-horária de 1h30min.**

Esta atividade foi desenvolvida pelo Projeto Museu Virtual Afro-Brasil-Sul, ação:

Webinar: Museu Afro-Brasil-Sul - Resgatando e Registrando Memórias, liderado

Prof. Dra. Rosemar Gomes Lemos.

Pelotas, 03 de agosto de 2020

ROSEMAR GOMES LEMOS
Líder do Grupo de Pesquisa CNPQ - Design, Escola e Arte - DEA - Desvendando patrimônios,

Lista das publicações resultantes do auxílio no período a que se refere o Relatório Científico

1. CARMO, Danielle do, MARTINS, Dalton Lopes (2020). Acervos Culturais Brasileiros no Repositório Wikimedia Commons: um estudo sobre o reuso e a visualização de mídias referentes a coleções de museus do Instituto Brasileiro de Museus. **Ciência Da Informação**, ISSN Eletrônico: 1518-8353. **Situação: aceito para publicação.**
2. SIQUEIRA, Joyce, MARTINS, Dalton Lopes (2020). Workflow de agregação de dados: processos para criação de uma interface de busca integrada do patrimônio cultural. **Ciência Da Informação**, ISSN Eletrônico: 1518-8353. **Situação: aceito para publicação.**

Os trabalhos listados encontram-se ao final do relatório, como anexo.

Lista dos trabalhos preparados ou submetidos

1. SIQUEIRA, Joyce, MARTINS, Dalton Lopes. WORKFLOW MODELS FOR AGGREGATING CULTURAL HERITAGE DATA ON THE WEB: A SYSTEMATIC LITERATURE REVIEW. **Situação: artigo concluído e já traduzido. Em preparação para submissão para o Journal of the Association for Information Science and Technology (JASIST).**

2. MARTINS, Luciana Conrado, MARTINS, Dalton Lopes. O desafio da formação de educadores museais e a cultura digital: perspectivas profissionais no século XXI. **Situação: Capítulo de livro submetido e em avaliação para Série de Livros Educação Museal: conceitos, história e políticas - Volume 5 do Museu Histórico Nacional.**

3. MARTINS, Luciana Conrado, MARTINS, Dalton Lopes. Experimentações sócio-técnicas para organização e difusão de coleções digitais universitárias: o caso do projeto Tainacan. **Situação: artigo submetido e em avaliação para revista CPC - Centro de Preservação Cultural da Universidade de São Paulo – ISSN 1980-4466**

4. MARTINS, Dalton Lopes, CARMO, Danielle do, SIQUEIRA, Joyce, LEMOS, Daniela Lucas de Silva (2021). Elementos para a construção de uma política de qualidade de dados para agregação de acervos culturais digitais: os casos da Digital Public Library of America.Inc e da Europeana Foundation. Situação: submetido para apresentação no evento **EAI DIONE 2021- 2nd EAI International Conference on Data and Information in Online Environments – Situação: aguardando resultado final em 10/12/2020.**

Os trabalhos listados encontram-se ao final do relatório, como anexo.

Apêndice 1 - Relatório de bolsista

Bolsas de Treinamento Técnico

1. Nome do bolsista

Vinícius Nunes Medeiros

2. Informação sobre o nível e período de usufruto da Bolsa.

TT5 – 40 hora semanais

3. Descrição das atividades do bolsista no projeto de pesquisa

Mês 13 a 15 – **Coleta dos dados e implementação da base de dados semântica:** consiste na coleta massiva dos dados disponíveis em cada repositório digital, no mapeamento e curadoria desses dados para o modelo semântico adotado, bem como na escolha da tecnologia de base de dados de grafos a ser utilizada pelo projeto

Esta etapa equivale, durante o período de vigência da bolsa, aos meses de junho, julho e agosto de 2020. Segue abaixo a descrição das atividades realizadas.

Para atender ao objetivo de identificar a viabilidade da coleta de dados dos acervos das entidades vinculadas à Secretaria Especial da Cultura, essa pesquisa aplica uma metodologia dividida em duas etapas principais: a primeira etapa envolve a seleção dos links dos acervos para compor a amostra, apresentada na seção 3.1, e a segunda etapa indica as diretrizes de coleta dos links selecionados, apresentada na seção 3.2. Estruturar essas etapas foi a principal tarefa realizada pelo bolsista no período.

3.1 Seleção de uma amostra dos links de acervos das entidades vinculadas à secretaria especial de cultura

O trabalho proposto para essa etapa projeto é uma coleta de parte dos acervos mapeados na primeira etapa, desse modo vale ressaltar que o mapeamento realizado identificou ao todo 217 links que redirecionam para potenciais acervos das entidades vinculadas à secretaria especial de cultura: FUNARTE, Ancine, Fundação Palmares, Biblioteca Nacional, Fundação Casa de Rui Barbosa e IBRAM.

Desses links mapeados, para o atual momento do projeto foi selecionada uma amostra que de acordo com a representatividade dos acervos, levando em conta àqueles acervos que atendem aos seguintes critérios:

- **quantidade de itens**, em que foram selecionados para a amostra links que representam a maior parte dos itens publicados pela instituição em seu site;
- **forma de coleta de dados**, em que foram selecionados para amostra links cujas formas de coleta de dados contemplem a maior variação possível, possibilitando que os testes de coleta aqui realizados permitam generalizar as características de coleta para os demais links mapeados;
- **ferramenta utilizada**, também foi levada em conta a variabilidade das ferramentas utilizadas para publicação do acervo, dado que ferramentas como SophiA, DSpace e Tainacan são por exemplo bastante difundidas entre as entidades analisadas, ao mesmo tempo que o Flickr e o Fotoweb 7 aparecem em casos específicos. Assim essa amostra buscou representar essa variabilidade considerando as proporções de representatividade de cada ferramenta.

Dessa forma levando em conta a representatividade da estrutura geral apresentada pelos resultados do diagnóstico realizado no âmbito do projeto, e a partir dos critérios apresentados no parágrafo anterior, foi construída uma amostra com 16 links para acervos das entidades vinculadas, como apresenta a tabela 1 abaixo.

Tabela 1 - Amostra de links de acervos selecionados para coleta dos dados

Entidade Vinculada	URL	Ferramenta	Extração dos Dados
Biblioteca Nacional	http://bndigital.bn.gov.b r/acervodigital/	SophiA	Raspagem de Dados
Biblioteca Nacional	http://acervo.bn.gov.br/ sophia_web/	SophiA	Raspagem de Dados
FUNARTE	http://sbrittod.funarte.g ov.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
Fundação Casa de Rui Barbosa	http://rubi.casaruibarbo sa.gov.br/	DSpace	OAI-PHM
Fundação Casa de Rui Barbosa	http://iconografia.casar uibarbosa.gov.br/fotowe b/default.fwx	Fotoweb 7	Raspagem de Dados
Fundação Casa de Rui Barbosa	http://acervos.casaruiba rbosa.gov.br/index.html	SophiA	Raspagem de Dados
Fundação Cultural Palmares	https://www.flickr.com/ photos/culturanegra/	Flickr	Raspagem de Dados
Fundação Cultural Palmares	http://www.palmares.gov.br/?page_id=50190	Wordpress	Raspagem de Dados
IBRAM	http://museudainconfidencia.acervos.museus.go v.br/	WordPress + Tainacan	API

IBRAM	http://museudearqueologiadeitaipu.museus.gov. br/	WordPress + Tainacan	API
IPHAN	http://acervodigital.ipha n.gov.br/xmlui/	DSpace	Raspagem de Dados
IPHAN	https://pergamum.iphan .gov.br/biblioteca/index. 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: Dados da pesquisa, 2020.

3.2 Coleta dos dados dos acervos selecionados

Uma das características importantes para a atual etapa foi a forma de coleta dos dados, essa informação (presente na coluna "Extração dos dados" da tabela 1) guiará a forma como os dados serão coletados para cada link de acervo selecionado.

Vale ressaltar que ao se aprofundar na técnica de coleta dos dados este estudo pode revelar outras formas de extração de dados, como, por exemplo, repositórios que permitem a exportação de dados através de mais de um formato. Quando isso ocorrer as características das diferentes formas de coleta serão descritas para cada link selecionado na amostra.

Assim, a metodologia para coleta aplicada neste estudo envolve de maneira generalizada três grandes etapas, também apresentadas na figura 1 abaixo:

- **Diagnóstico**, que compõe as atividades de identificação da forma de coleta de dados, e ainda a identificação dos pontos de coleta dos dados: no caso de OAI-PMH e API identificar o(s) link(s) do *endpoint* de acesso; para exportadores, identificar a funcionalidade de exportação; e para raspagem de dados identificar possíveis hierarquias de links.
- **Processamento**, que envolve o processamento dos dados, quanto à identificação dos metadados, valores e objetos, bem como os métodos e atributos disponíveis nos diferentes formatos de coleta dos dados, no caso de raspagem de dados essa etapa contempla a identificação da estrutura de exibição do acervo.

Ainda no contexto de processamento são desenvolvidos scripts de coleta com base nos os

procedimentos de diagnóstico executados para os formatos de coleta através de OAI-PMH, API e Raspagem de Dados, para os exportadores são utilizadas as funcionalidades de exportação através da interface web do próprio acervo.

Vale ressaltar que para os acervos obtidos através da raspagem de dados, foram coletados um máximo de 100 registros para cada link selecionado na amostra. Isso se deve pelo fato da onerosidade de tempo para coleta de itens através da raspagem de dados, já que é necessário requisitar cada página onde os registros do acervo aparecem, e o tempo de resposta de requisição varia podendo levar entre segundos a minutos para recuperar um único registro, dessa forma a título de otimização do tempo de coleta para a produção deste relatório a coleta de registros foi limitada ao máximo de 100 registros.

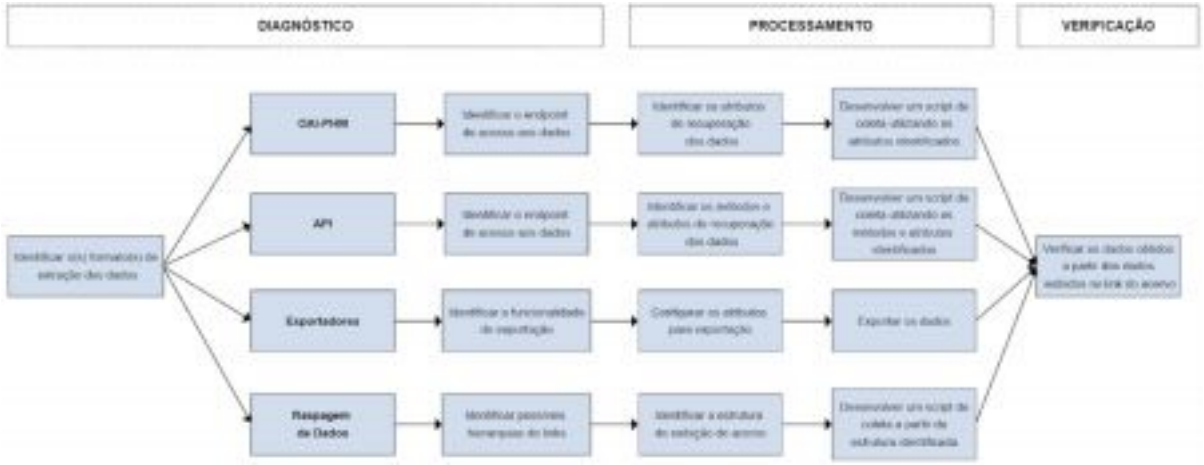
Outro ponto em comum das atividades de processamento realizadas, com exceção à coleta por exportadores, foi o uso da linguagem de programação Python para obter os e armazenar os resultados, além dessa linguagem se apresentar de forma robusta para a automatização de aplicações que envolvam o ambiente web, é uma expertise já desenvolvida pelo contexto do grupo de pesquisa que desenvolve este projeto.

Ainda derivado da linguagem Python, o também é comum entre os processos de coleta que necessitaram de desenvolvimento de scripts o uso da biblioteca *Pandas*, que é utilizada para análise de dados, e para este projeto foi muito utilizada no contexto de armazenamento dos dados utilizando o método *DataFrame*, que permite estruturar os dados em um formato bi dimensional de colunas e linhas, possibilitando também sua exportação para diversos outros formatos, como por exemplo, CSV, SQL (Single Query Language), Excel, entre outros.

- **Verificação**, uma vez que os dados foram coletados, foi realizado um processo de verificação dos dados para cada saída, de forma que 20 itens escolhidos de forma aleatória da coleta obtida foram comparados com os dados apresentados no acervo on-line.

Para os dados obtidos através de scripts (OAI-PMH, API e raspagem de dados), o processo de validação dos dados, ocorreu utilizando o método *sample* da biblioteca *pandas* (aproveitando que os mesmos já se encontram no formato *DataFrame)*, para apresentar os dados armazenados de 20 registros aleatórios, que foram buscados na interface web do acervo e confirmada a correspondência dos dados. O mesmo processo pode ser aplicado para arquivos exportados, ao se desenvolver um script para leitura dos arquivos e apresentar os registros aleatórios.

Diagrama 1 - Etapas gerais de coleta dos dados



Fonte: Dados da pesquisa, 2020.

A partir dessa coleta piloto, os dados completos dos repositórios serão coletados até fevereiro de 2021, conforme cronograma original do projeto.

Mês 16 a 21 – **Desenvolvimento e customização do repositório digital para busca integrada:** consiste na customização de uma solução livre de repositório digital, preferencialmente a solução Tainacan, para dar acesso a base de dados de grafos semânticos contendo os dados agregados dos repositórios digitais coletados. Envolve a modelagem do design gráfico da solução, a implementação das técnicas de busca, análise de desempenho das consultas a banco de dados e dos diferentes modos de exibição dos dados, além do teste de usabilidade do sistema com usuários;

Esta etapa equivale, durante o período de vigência da bolsa, aos meses de junho, setembro de 2020 a fevereiro de 2021. Segue abaixo a descrição das atividades realizadas. Nessa etapa, ainda em fase inicial, foi realizada a atividade de integração do repositório digital Tainacan com a solução ElasticSearch, para criação de um índice de busca integrado com os dados piloto coletados do projeto.

4. Informar e justificar caso tenham ocorrido mudanças e, eventualmente, os ajustes realizados nas atividades de pesquisa do bolsista, em relação ao Plano de Atividades

A modificação mais expressiva dessa fase é a constatação de que os dados coletados possuem modelos de metadados bastante simples, não permitindo com que os mesmos sejam automaticamente mapeados e expressados para modelos conceituais semânticos mais elaborados. Dessa forma, optou-se por utilizar um banco de dados relacional para coleta e armazenamento dos dados e não um banco de dados de grafos, como era originalmente previsto. Além disso, considerando a carga completa de dados, que deve chegar a um número aproximado de 3 milhões de registros, optou-se por utilizarmos a ferramenta ElasticSearch para construção do índice de busca e conexão com o repositório digital Tainacan.

5. Avaliação do impacto das atividades do bolsista sobre o andamento do projeto

Entende-se que o trabalho realizado cumpriu o esperado e entrega ao projeto uma etapa estruturante, permitindo com que os dados possam ser coletados em seus diferentes formatos de acesso e exportação. Com o trabalho realizado até aqui, é possível demonstrar que todos os dados publicados dos acervos culturais podem ser coletados, com níveis de expressividade, automação e dificuldade técnica bastante diferentes, mas ainda assim passíveis de recuperação para a construção de um índice de busca integrado

6. Juntar o histórico escolar atualizado do bolsista

Informamos que o bolsista já é formado e seu histórico se encontra junto ao processo de seleção da bolsa.

7. Apreciação do desempenho do bolsista

Considero o desempenho do bolsista adequado e conforme as expectativas. O mesmo demonstrou bom conhecimento técnico, capacidade de solução de problemas, pesquisa, colaboração com o grupo de pesquisadores, iniciativa na proposição de caminhos técnicos e boa organização na implementação, documentação e análise do trabalho realizado.

Apêndice 2 - Relatório de bolsista

Bolsa de Treinamento Técnico

1. Nome do bolsista

Robson Batista da Silva

2. Informação sobre o nível e período de usufruto da Bolsa.

Treinamento Técnico IV (TT IV)

3. Descrição das atividades do bolsista no projeto de pesquisa

Mês 1 e 2 – Análise conceitual dos repositórios digitais

Período: meses de julho a agosto de 2020.

A pesquisa foi realizada nos acervos culturais de instituições vinculadas à Secretaria Especial da Cultura (2013c), elencada especialmente por sua relevância e abrangência nacional, a saber: o Instituto do Patrimônio Histórico e Artístico Nacional – Iphan; o Instituto Brasileiro de Museus – Ibram; a Agência Nacional do Cinema – Ancine; a Fundação Casa de Rui Barbosa – FCRB; a Fundação Cultural Palmares – FCP; a Fundação Nacional das Artes – Funarte; e a Fundação Biblioteca Nacional - FBN.

A etapa de coleta e análise dos dados contou com o método de pesquisa conhecido como análise de conteúdo (Bardin, 2016), que adota como estratégia um conjunto de métodos e técnicas com abordagens quali-quantitativas visando a criação de categorias de análise para compreensão mais abrangente do fenômeno investigado de modo a facilitar a extração, a análise e a interpretação mais concisa dos dados presentes nos materiais elencados no

estudo. Desse modo, a pesquisa se organizou de forma cronológica em torno de três fases: i) pré-análise; ii) exploração do material; e iii) tratamento dos resultados, a inferência e a interpretação.

- A pré-análise: definição do corpus central e das categorias de análise

A fase inicial partiu do objetivo da pesquisa, que foi a base para a construção do instrumento de coleta de dados. Consistiu em três ações baseadas na construção do corpus central da pesquisa (elementos que a análise vai fazer falar), na identificação das hipóteses e objetivos da pesquisa e na elaboração dos indicadores de análise que permitam a interpretação dos resultados. Os indicadores se tornaram unidades comparáveis de categorização para análise temática a partir do contato inicial com fontes bibliográficas e documentais para a obtenção de impressões e orientações acerca de seus conteúdos.

A primeira ação, executada de forma manual, se deu a partir da identificação dos endereços eletrônicos das instituições vinculadas à Secretaria Especial da Cultura. A partir desses endereços, tornou-se possível explorar o ambiente digital e identificar fontes documentais em potencial inerentes aos conjuntos de itens presentes nos acervos de cada instituição vinculada (Tabela 1). Torna-se válido ressaltar que a pesquisa considerou acervo como todo conjunto de itens (ou documentos) organizado em possíveis tipos de SRIs, tais como repositórios e bibliotecas digitais, sistemas de gestão de conteúdo e itens sistematizados em páginas web dispersas. Para este último, considerou que potencialmente poderiam fazer parte do acervo da instituição vinculada e que, em uma possível agregação e busca integrada, seria interessante que houvesse a recuperação desses itens. Sobretudo, deseja-se conhecer aqui os objetos culturais que já foram digitalizados e estão disponibilizados ao público de alguma maneira acessível pela web, sendo que foram eliminados da pesquisa documentos administrativos ou ligados a gestão da instituição.

Na tabela 1, a seguir, apresenta-se as os sítios web de cada instituição cultural e as fontes documentais que foram identificadas nessa etapa de pré-análise e que serão objeto da análise de conteúdo proposta na presente pesquisa. Na coluna "fontes documentais", pode-se observar o total de endereços web que representam conjuntos de objetos culturais publicados na web para cada instituição. Vale ressaltar aqui que um endereço web pode ser um endereço de um repositório digital que agrega milhares de objetos culturais ou apenas uma página web com uma dezena de objetos culturais disponíveis para visitação como imagem. O foco da pesquisa é exatamente analisar essa variabilidade nas estratégias de publicação de objetos culturais e como isso pode impactar na ideia de construção de uma ferramenta de busca e recuperação da informação de forma integrada.

Tabela 1 - *Corpus* central da pesquisa - fontes documentais por instituição vinculada

Instituição Vinculada	Endereço eletrônico (*link* principal)	Nº Fontes documentais
Agência Nacional do Cinema	https://www.ancine.gov.br/	3
Fundação Biblioteca Nacional	https://www.bn.gov.br/	38
Fundação Casa de Rui Barbosa	http://www.casaruibarbosa.gov.br/	22
Fundação Cultural Palmares	http://www.palmares.gov.br/	1
Fundação Nacional das Artes	https://www.funarte.gov.br/	10
Instituto Brasileiro de Museus	https://www.museus.gov.br/museus-ibram/	28
Instituto do Patrimônio Histórico e Artístico Nacional	http://portal.iphan.gov.br/	113

A segunda ação da etapa de pré-análise consiste na identificação das hipóteses e objetivos da pesquisa. Tem-se por hipótese, baseado na experiência dos pesquisadores em lidar cotidianamente com as instituições culturais brasileiras, que os acervos culturais digitalizados
são publicados na web de forma ainda bastante precária, seja tanto em termos das tecnologias para acesso e navegação nos documentos quanto nas práticas de gestão da informação estabelecidas, o que envolve os usos de padrões de metadados, vocabulários controlados e regras de catalogação, para citar alguns exemplos. Logo, entende-se que a hipótese estabelecida para esses acervos, apesar de digitalizados, é de que estão disponíveis de forma a dificultar sua agregação, busca e recuperação pelos seus usuários em potencial. Desse modo, tem-se como objetivo da presente pesquisa apresentar um mapeamento sistemático das formas de organização e representação da informação que estão sendo aplicadas aos dados pertencentes aos acervos das instituições vinculadas à Secretaria Especial da Cultura.

A terceira ação se baseou em princípios advogados por Bardin (2016) quanto ao uso de categorias para procedimentos de análise qualitativa e quantitativa. Nesse sentido, o procedimento de categorização se orientou em questões fundamentais com garantia literária em áreas ligadas aos campos da informação (ex.: política de qualidade de dados; organização e representação da informação) e da tecnologia (ex.: sistemas de informação; agregação de dados), a saber: i) quais os tipos de sistemas de informação utilizados por

essas instituições? ii) quais os recursos tecnológicos utilizados para a publicação dos acervos? iii) quais licenças, regras de catalogação, padrões de metadados e linguagens documentárias são explicitados na divulgação do acervo? iv) quais as formas de exposição dos itens dos acervos? v) quais as formas de extração dos dados do acervo? vi) em qual formato os itens do acervo estão disponibilizados? e vii) qual o tamanho do acervo?

A construção das categorias levou em consideração o objetivo da pesquisa e seu potencial uso estratégico como desdobramento dos resultados atuais, a saber, que esses dados sirvam de insumo para a construção de um de serviço de informação para coleta, agregação, busca e recuperação dos objetos culturais brasileiros de forma integrada. Ou seja, para além dos elementos analíticos já descritos aqui, são de interesse da pesquisa compreender qual o tamanho em termos quantitativos dos objetos culturais digitais já disponibilizados na web pelas instituições pesquisadas, de que forma esses acervos podem ser visualizados, os tipos de mídia que estão disponíveis, as possíveis formas de extração de dados, quando existirem, além do tipo de sistema de recuperação da informação que está sendo utilizado. Entende-se que esses fatores permitirão um diagnóstico situacional dos acervos digitais das instituições, permitindo que os resultados possam ser usados para refletir qual a melhor maneira de se aproveitar esses objetos digitais para um serviço de informação que tenha por objetivo agregar os acervos. Perguntas relativas a se os softwares atuais devem ser migrados para outras soluções ou não, uma estimativa inicial sobre o tamanho dos servidores web para hospedar os objetos digitais, a possibilidade de reaproveitar esses objetos ou não, entre outras, podem ser analisadas em pesquisas posteriores com os resultados aqui apresentados.

A partir de tais questões e usando-se do critério semântico para alinhamento temático, as categorias de análise foram definidas e são descritas como se segue.

Tipo de SRI: visa especificar o tipo de sistema de recuperação da informação com foco a destacar de que maneira o objeto digital é acessível. Foram usadas as categorias página estática (HTML), quando os objetos digitais foram identificados dentro de uma página web como imagens, vídeos ou documentos textuais e não se identificou um sistema específico para a geração desta página web, sistema de repositório digital, quando se identificou qual sistema específico com características de repositório digital (sistemas que apresentaram organização por coleções, metadados explícitos, navegação facetada e campo de busca específico) era utilizado para dar acesso aos objetos digitais, sistema de gerenciamento de conteúdo (CMS), quando se identificou que havia um sistema de gerenciamento de conteúdo que gerava as páginas web nas quais os objetos foram identificados, mas que não possuía a estrutura de um repositório digital , arquivo de dados, quando se identificou que os objetos digitais estavam inseridos dentro de arquivos de dados, como por exemplo, um conjunto de fotografias disponível em um catálogo em PDF. Software utilizado: a tecnologia utilizada pelo sistema de recuperação da informação. Foram propostas as categorias sistema de gerenciamento de conteúdo, repositório digital, arquivo de dados, página estática html e não identificados. Licença sobre os dados: a licença de publicação dos objetos digitais. Organização e representação da informação: nessa categoria foram considerados, quando identificados, os padrões de metadados, as linguagens documentárias

e as regras de catalogação utilizadas pelos acervos. Visualização do acervo: constitui da forma como o conjunto de itens é apresentado no sítio web das entidades vinculadas, foram identificadas 3 categorias: Coleções, em que os itens estão claramente organizados em coleções; Exposição, em que os itens estão organizados no formato de exposições, onde se percebe que apenas alguns itens são exibidos com informações contextuais de apoio (páginas HTML com parte do acervo por exemplo); Hierárquica, em que os itens estão organizados de forma hierárquica (como em listas de registros ou vídeos, ou ainda conjuntos de pastas por exemplo); Extração dos dados: formas de exportação dos dados dos sistemas de informação. Tipo de mídia: o formato das mídias, se representam texto, imagem, vídeo ou áudio. Quantidade de itens no acervo: o total de itens identificados em cada sítio web

- A exploração do material

A segunda fase partiu das categorias determinadas na fase anterior, para as quais se somaram as inferências realizadas ao corpus central da pesquisa, que culminou num segundo instrumento de coleta de dados destinado a descrever o conteúdo de cada fonte documental à luz das categorias previamente determinadas. Desse modo, os dados obtidos a partir dessa descrição foram revisados e normalizados a partir da revisitação de todas as fontes documentais, verificando a adequação a cada categoria de análise. Novas categorias foram então identificadas e agregadas ao processo de análise, garantindo, assim, a sistematização coerente dos dados referente aos acervos de cada instituição investigada.

- O tratamento dos resultados

A última fase marcou o desenvolvimento da síntese gráfica de vários resultados quantitativos obtidos a partir do instrumento de coleta. A partir do tratamento e da organização dos dados, tornou-se possível obter um corpus teórico conceitual frente aos objetos investigados e tirar conclusões teóricas e metodológicas acerca do processo de construção e publicação de acervos digitais em rede pelas instituições de cultura brasileiras. Os resultados quantitativos e gráficos encontram-se disponíveis em: https://unbbr my.sharepoint.com/:w:/g/personal/daltonmartins_unb_br/EbOT8tFg94ZCjegMQVcmDNMBurmbD33GimWCWW4dutaq-A?e=9Znj55

Mês 3 a 6 – Desenvolvimento dos scripts de coleta de dados e armazenamento em base de dados relacional

Período: meses de setembro a dezembro de 2020.

Foram desenvolvidos 18 scripts para extração dos dados e armazenamento em base de dados. Os scripts foram feitos em linguagem Python e estão disponíveis e licenciados com código aberto na plataforma Github: https://github.com/tainacan/data_science/tree/master/FAPESP/scripts_extracao.

Os dados extraídos para análise e teste estão disponíveis aqui: https://drive.google.com/drive/folders/1_7rLpmpecR0QfplI3xOcTKUI301EW28b.

4. **Informar e justificar caso tenham ocorrido mudanças e, eventualmente, os ajustes realizados nas atividades de pesquisa do bolsista, em relação ao Plano de Atividades**

Não houve modificação expressiva no plano de trabalho do bolsista. O projeto transcorreu como esperado.

5. **Avaliação do impacto das atividades do bolsista sobre o andamento do projeto**

Entende-se que o trabalho realizado cumpriu o esperado e entrega ao projeto uma etapa estruturante, permitindo com que os dados possam ser coletados em seus diferentes formatos de acesso e exportação, além da fundamental análise conceitual e metodológica realizada sob orientação da coordenação do projeto. Com o trabalho realizado até aqui, é possível demonstrar que todos os dados publicados dos acervos culturais podem ser coletados, com níveis de expressividade, automação e dificuldade técnica bastante diferentes, mas ainda assim passíveis de recuperação para a construção de um índice de busca integrado

6. **Juntar o histórico escolar atualizado do bolsista**

Informamos que o bolsista já é formado e seu histórico se encontra junto ao processo de seleção da bolsa.

7. **Apreciação do desempenho do bolsista**

Considero o desempenho do bolsista adequado e conforme as expectativas. O mesmo demonstrou bom conhecimento técnico, capacidade de solução de problemas, pesquisa, colaboração com o grupo de pesquisadores, iniciativa na proposição de caminhos técnicos e boa organização na implementação, documentação e análise do trabalho realizado.

Anexo A

Texto completo dos artigos produzidos relacionados e/ou derivados das atividades do projeto

Acervos Culturais Brasileiros no Repositório Wikimedia Commons: um estudo sobre o reuso e visualização de mídias referentes a coleções de museus do Instituto Brasileiro de Museus

Resumo

O presente artigo apresenta um estudo webométrico acerca de mídias referentes a coleções de acervos de museus brasileiros no repositório de mídias Wikimedia Commons, mais especificamente sobre mídias de coleções de nove museus geridos pelo Instituto Brasileiro de Museus (Ibram). Utilizando as ferramentas GLAMourous e GLAMourous 2 foi possível obter dados relativos ao reuso de mídias da Wikimedia Commons em outras plataformas wikis da Fundação Wikimedia, como os projetos Wikipédia de diferentes idiomas, a Wikidata e outros.

Também foi possível obter números relativos a quantidade de visualizações desses acervos no ano de 2019. Dessa forma a investigação realizada, assim como os dados obtidos, podem ajudar instituições culturais guardiãs de acervos, como os museus, a entender o que acontece com seus acervos digitais uma vez que são disponibilizados nas plataformas wiki, fornecendo dados importantes para a gestão de suas coleções, principalmente em relação ao acesso e reapropriação desses acervos por parte dos usuários.

Palavras-chave: Wikimedia Commons. Acervos culturais digitais. Museus do Ibram. Wikidata. Instituições culturais. GLAM.

Brazilian Cultural Collections in the Wikimedia Commons Repository: a study on the reuse and visualization of media related to museum collections of the Brazilian Institute of Museums

Abstract

This article presents a webometric study about media referring to collections of Brazilian museums in the Wikimedia Commons media repository, more specifically about media from collections of nine museums managed by the Brazilian Museum Institute (Ibram). Using the GLAMourous and GLAMourous 2 tools, it was possible to obtain data regarding the media reuse of Wikimedia Commons in other Wikimedia Foundation wiki platforms, such as Wikipedia projects in different languages, Wikidata and others. It was also possible to obtain numbers related to the views of these collections in 2019. In this way, the research carried out, as well as the data obtained, ways to help cultural institutions guarding collections, such as museums, to understand what happens with their digital collections. since they are made available on wiki platforms, providing important data for the management of their collections, mainly in relation to access and reappropriation of these collections by users.

***Keywords**: Wikimedia Commons. Digital cultural collections. Museums of Ibram. Wikidata. Cultural institutions. GLAM.*

Colecciones culturales brasileñas en el repositorio de Wikimedia Commons: un estudio sobre la reutilización y visualización de medios relacionados con colecciones de museos del Instituto Brasileño de Museos

Resumen

Este artículo presenta un estudio webométrico sobre los medios que se refieren a colecciones de museos brasileños en el repositorio de medios de Wikimedia Commons, más específicamente sobre medios de colecciones de nueve museos administrados por el Instituto Brasileño de Museos (Ibram). Utilizando las herramientas GLAMourous y GLAMourous 2, fue posible obtener datos sobre la reutilización de medios de Wikimedia Commons en otras plataformas de wikis de la Fundación Wikimedia, como proyectos de Wikipedia en diferentes idiomas, Wikidata y otros. También fue posible obtener números relacionados con el número de vistas de estas colecciones en 2019. De esta manera, la investigación llevada a cabo, así como los datos obtenidos, pueden ayudar a las instituciones culturales que custodian colecciones, como los museos, a comprender lo que sucede con sus colecciones digitales. ya

que están disponibles en plataformas wiki, proporcionando datos importantes para la gestión de sus colecciones, principalmente en relación con el acceso y la reapropiación de estas colecciones por parte de los usuarios.

***Palabras clave**: Wikimedia Commons. Colecciones culturales digitales. Museos de Ibram. Wikidata. Instituciones culturales. GLAM..*

INTRODUÇÃO

Com o surgimento das novas tecnologias da informação e comunicação, mais especificamente com o advento da internet, pôde-se observar a emergência de novos espaços sociais baseados no digital e em novos tipos de fluxos informacionais. Esses espaços, ocupados com objetivos e propósitos diversos, oferecem aos seus usuários meios de consumo e compartilhamento de informações, mecanismos de interação com as informações e dispositivos que promovem a conexão e comunicação direta entre usuários. Como exemplo desses meios sociais digitais que emergiram principalmente com o advento da Web 2.0, podemos citar plataformas de conteúdos gerados pelo usuário como a rede social Facebook, a rede de microblogs Twitter, o streaming de vídeo YouTube, a enciclopédia livre Wikipédia. Esses espaços sociotécnicos além de serem entendidos como meios potenciais de socialização de informação e conteúdo, também podem ser entendidos como fontes de informação que permitem a coleta de dados sobre a circulação e apropriação de objetos digitais, e nos fornece informações sobre a forma como são disponibilizados, contextualizados e descritos.

Nesse sentido, a webometria pode ser uma importante aliada em estudos que pretendem investigar fenômenos baseados na web. Bjorneborn e Ingwersen definiram a webometria como o "estudo de fenômenos da web baseados em técnicas quantitativas e recorrendo a métodos infométricos" (2004). Para Thelwall (2009, p.1) essa definição foi importante por atribuir à webometria características de método infométrico, pois dessa forma os autores posicionaram a webometria como um campo da Ciência da Informação, mas propõe uma definição mais abrangente ao dizer que a webometria é "o estudo de conteúdos baseados na web com métodos essencialmente quantitativos que não são específicos para um campo de estudo" (THELWALL, 2009, p.6). Vanti (2002, p.161) posiciona a webometria, a bibliometria e a cienciometria como subcampos da infometria e diz que todas "têm funções semelhantes, mas, ao mesmo tempo, cada uma delas propõe medir a difusão do conhecimento científico e o fluxo da informação sob enfoques diversos" (VANTI, 2002). Entre os estudos webométricos estariam incluídos aqueles que adotam métodos de rastreamento das ações dos usuários online, que por meio de ferramentas de análise, podem captar informações que nos permitem medir aspectos de uso de recursos disponíveis na web (THELWALL, 2009 p. 89).

Da mesma forma que os estudos bibliométricos permitem uma melhor administração das coleções no contexto das bibliotecas, os estudos webométricos podem dar subsídios que auxiliem as instituições culturais, como os museus, a administrar os recursos, entre eles objetos de acervos culturais digitais, e entender o seu uso pelos usuários que os acessam. Dessa forma, a presente investigação explora ferramentas de coleta de dados de informações e apresenta possibilidades de análise e mensuração de aspectos relacionados ao acesso e a reutilização de acervos brasileiros que estão disponíveis na web por meio do repositório de mídias Wikimedia Commons.

O repositório de mídias online Wikimedia Commons armazena e disponibiliza, de forma gratuita, diversos tipos de arquivos de mídias que são compartilhadas de forma coletiva

e colaborativa por meio da ação de usuários voluntários. É um projeto da fundação sem fins lucrativos Wikimedia, que mantém na web outros projetos baseados em conteúdo gerado por usuário como a Wikipédia, o Wikitionaty, o Wikibooks, o Wikisource, o Wikinews, o Wikiversite, o Wikiquote, o Wikidata e outros.

O Wikimedia Commons, lançado no início de setembro de 2004, disponibiliza atualmente mais de 60 milhões de arquivos de mídias. Essas mídias são disponibilizadas sob licenças individuais que permitem a cópia, a reapropriação e a modificação, de acordo com termos especificados. Os conteúdos disponibilizados por meio do repositório Wikimedia Commons estão sob a licença *Creative Commons Attribution / Share-Alike* (WIKIMEDIA COMMONS, 2019). Segundo dados fornecidos pelo projeto, em abril de 2020, a plataforma registra um total de 43,020 usuários ativos e 168 *bots* que auxiliam na edição do conteúdo (WIKIMEDIA COMMONS, 2020a).

Ao explorar o repositório Wikimedia Commons é possível observar que estão disponíveis diferentes tipos mídias de diversas temáticas. Entre as mídias é possível identificar conteúdos referentes a itens de acervos oriundos de coleções de diversas instituições culturais como museus, arquivos e galerias localizadas em diversas partes do mundo. Essas mídias podem ter sido disponibilizadas tanto por usuários que realizaram o upload desse material de modo espontâneo, ou em alguns casos, por meio de parcerias estabelecidas entre usuários e instituições culturais. Segundo os autores Stinson, Fauconnier e Wyatt (2018, p.17) há um histórico de colaboração entre a comunidade Wikimedia e as instituições culturais, reunidas sob o termo guarda-chuva GLAM[1]. Essas instituições identificaram nos projetos da Fundação Wikimedia em especial no Wikimedia Commons, na enciclopédia online Wikipédia e no repositório de dados estruturados Wikidata, meios potenciais de difusão de seus acervos na internet. Para se ter uma noção do dos efeitos dessa prática, Zeinstra (2013) aponta que no ano 2013 o conteúdo GLAM reunia cerca mais de dois milhões de objetos digitais no repositório de mídias, o que corresponderia um total de 13,14% do conteúdo da Wikimedia Commons (ZEINSTRA, 2013).

> Em outras palavras, um em cada oito arquivos na Wikimedia Commons é disponibilizado através de alguma colaboração com GLAM's. Tanto a GLAM coloca suas coleções em domínio público ou abre a licença de suas coleções online e voluntários realizam upload, ou alguma colaboração ativa entre GLAM's e a comunidade Wikimedia como Wiki Loves Monuments cria conteúdo GLAM. Dessa forma, voluntários colocam esses objetos de mídia em artigos da Wikipédia criando um incrível aumento em sua visibilidade. Instituições contribuintes enxergam o potencial da Wikimedia como um canal de distribuição. (ZEINSTRA, 2013)

Villaespesa e Navarete (2019) chamam a atenção que na utilização de mecanismos de buscas como o Google e o Google Imagens, ou por meio da busca de voz de assistentes virtuais como a Siri e a Alexa é possível identificar uma exposição privilegiada de conteúdos da Wikimedia Commons, Wikidata e Wikipédia. Essa exposição privilegiada de conteúdos na ordem de apresentação de resultados se daria devido "aos ambientes estruturados fornecidos pelas plataformas Wiki que influenciam fortemente os resultados das pesquisas" (VILLAESPESA; NAVARETTE, 2019).

Além de se apresentarem como potenciais aliados na difusão das informações referentes aos acervos das instituições culturais, os projetos da Wikimedia constituem-se como espaços sociotécnicos que promovem práticas de curadoria coletiva e que podem nos

fornecer dados relevantes sobre aspectos relacionados ao acesso e a reutilização das coleções de acervos culturais compartilhados nesses ambientes. Nesse sentido a Wikimedia Foundation, por meio de uma página na Wikimedia Outreach sobre os projetos com GLAM's (WIKIMEDIA OUTREACH, 2020) indica uma série de ferramentas desenvolvidas com o objetivo de captar e os expor dados sobre as mídias digitais de coleções culturais em ambientes como a Wikimedia Commons.

O objetivo da presente pesquisa foi explorar como esses dados podem ser extraídos e analisados, evidenciando seu potencial como fonte de informação sistematizada e ampliar a própria sensibilização da área da Ciência da Informação para os fenômenos culturais que se dão nesses ambientes e o quanto eles podem indicar novas práticas sociais de gestão e curadoria da informação que devem ser levadas em consideração em pesquisas na área. O foco da pesquisa é demonstrar a presença dos acervos de uma amostra de instituições culturais na Wikimedia e como essa presença pode ser percebida a partir de diferentes tipos de indicadores e informações disponibilizadas pela plataforma. Como amostra utilizaremos mídias de coleções que tenham como origem acervos de 30 museus que estão sob a gerência do Instituto Brasileiro de Museus (Ibram).

METODOLOGIA

A presente pesquisa se apresenta como exploratória e descritiva, de natureza quantitativa, e para que fosse possível coletar os dados para análise foram utilizadas duas ferramentas, a GLAMourous e a GLAMorous 2. Ambas as ferramentas foram desenvolvidas pelo wikimedista[2] Magnus Manske e se encontram listadas para uso em uma página de ferramentas voltadas para os GLAM's no Wikimedia Outreach (WIKIMEDIA OUTREACH, 2020). Com essas ferramentas é possível obter dados de visualização das páginas dos arquivos das mídias do Wikimedia Commons com base nas categorias. Dessa forma, depois de identificar os museus sob gestão do Ibram, exploramos as categorias referentes às coleções dos museus na Wikimedia Commons. Uma vez identificadas, foi possível utilizá-las nos mecanismos de busca das ferramentas selecionadas e estabelecer parâmetros. A ferramenta GLAMourous permite o rastreamento de uso de imagens que estão em uma categoria commons, e por meio dela foi possível obter dados relativos ao reuso de mídias do Wikimedia Commons em páginas dos diversos projetos da Fundação Wikimedia. É importante destacar que essa possibilidade de rastreamento do uso de objetos digitais pode ter diversas aplicações em pesquisas que tenham por objetivo compreender as formas de apropriação social e reuso da informação, sobretudo ressaltando que o contexto em que elas são utilizadas pode ser dessa forma analisado, evidenciando significados fundamentais para a compreensão dos fluxos informacionais constituídos nesses ambientes. Desde o momento de publicação de um objeto digital por uma instituição, passando pelo momento em que esse objeto é escolhido para ilustrar um conceito em um verbete até a sua visualização por um usuário em um verbete específico na Wikipédia, diversas camadas podem compreendidas por meio dessas informações registradas e apoiar em compreensões do fenômeno social de uso da informação que ainda hoje carecem de entendimento e fortalecimento metodológico. Já a ferramenta GLAMorous 2 complementa a ferramenta GLAMorous, apresentando funcionamento semelhante e, além de dados de reuso, por meio dela é possível obter dados de visualização dos arquivos. Com essas ferramentas foi possível realizar análises webométricas que serão apresentados nas próximas seções do presente artigo.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

A reutilização de imagens de coleções culturais na Wikimedia Commons em outros projetos da Fundação Wikimedia

Ao identificar 30 museus que estão sob a administração direta do Ibram, foi possível localizar categorias de mídias que correspondem às suas coleções no Wikimedia Commons por meio de buscas na plataforma[3]. Dessa forma foram identificadas categorias referentes a coleções de nove museus que estão elencadas no Quadro 1.

Quadro 1 – Categorias referentes às coleções dos museus no Wikimedia Commons

#	Museu	Categoria
1	Museu da Inconfidência	Collections of the Museu da Inconfidência
2	Museu da República	Collections of the Museu da República
3	Museu Histórico Nacional	Collections of the Museu Histórico Nacional
4	Museu Imperial	Collections of the Museu Imperial
5	Museu Nacional de Belas Artes	Collections of the Museu Nacional de Belas Artes
6	Museu do Açude (equipamento dos Museus Castro Maya)	Media contributed by the Museu Casa de Benjamin Constant
7	Museu Casa de Benjamin Constant	
8	Museu Regional de São João del Rei	Collections of the Museu Regional de São João del-Rei
		Collections of the Museus Castro Maya
9	Museu Victor Meirelles	Collections of the Museu Victor Meirelles

Fonte: Dados da pesquisa, 2020.

Identificadas as categorias referentes a coleções de museus, foi possível utilizá-las para realizar buscas na ferramenta GLAMorous e coletar estatísticas atuais de reuso de imagens da Wikimedia Commons em páginas da Wikipédia em português, em Wikipédias de outros idiomas e em páginas de outros projetos da Fundação Wikimedia, como o Wikibooks, Wikiquotes e o Wikidata. Além disso foi possível obter a quantidade de imagens encontradas em cada categoria, o total de imagens distintas utilizadas, o total de uso das imagens das categorias, assim como a porcentagem do total de imagens da categoria que estão sendo reutilizadas, como pode ser visto na Tabela 1 e na Tabela 2.

Tabela 1 – Reuso de imagens nos projetos da Fundação Wikimedia

Coleção Arquivos na categoria	Wikipédia português	Wikipédia outros idiomas	Wikibooks	Wikiquote	Wikidata
Museu da Inconfidênci a		Museu da República Museu			

Nacional de Belas Artes

Museus
Castro Maya

Museu Casa de Benjamin Constant

Museu
Regional de
7 8 10 0 0 0

9 33 75 1 1 0 17 2 4 0 0 1

46 12 25 5 0 3 31 4 0 0 0 0

6 0 0 0 0 0
São João del-Rei[4]

Museu Victor
Meirelles

Museu Histórico Nacional

Museu Imperial
4 5 12 0 0 3 471 13 19 0 0 4 38 7 52 0 0 0

Fonte: Dados da pesquisa, 2020[5].

Tabela 2 – Quantidade de imagens reutilizadas nos projetos da Fundação Wikimedia.

Coleção Arquivos na categoria	Total de imagens distintas	Total de uso de imagens	Imagens da categoria (%)
Meirelles			
Museu da Inconfidência			
Museu Histórico Nacional			
Museu da República		7 6 18 85,71 9 4 110 44,44 17 7 7 17,65	
Museu Nacional de Belas Artes			
Museus Castro Maya		46 17 45 36,96 31 4 4 12,9 6 0 0 0 4 3 20 75 471	
Museu Casa de Benjamin Constant			
Museu Regional de São João del-Rei		12 36 2,55	
Museu Victor			
Museu Imperial	38	8 59	21,05

Fonte: Dados da pesquisa, 2020.

Com base nos dados obtidos, é possível observar que a coleção de museu que mais apresenta imagens reutilizadas na ilustração páginas da Wikipédia em português é a categoria referente ao Museu da República que apresenta 33 ocorrências de imagens presentes em artigos. Essa categoria também é a que ilustra mais Wikipédias de outros idiomas, apresentando assim 75 ocorrências em páginas de Wikipédias em outros idiomas, 1 ocorrência de imagem no Wikibooks e 1 no Wikiquote. Assim foi possível constatar que das 9 imagens encontradas na categoria correspondente à coleção do Museu da República, 4 estão sendo reutilizadas em
110 páginas de projetos da Fundação Wikimedia, o que corresponde um total de uso de 44,4% das imagens disponíveis na categoria.

A categoria referente à coleção do Museu Histórico Nacional apresentou o maior número de arquivos, com 471 imagens encontradas na categoria. Essas imagens foram reutilizadas 13 vezes em páginas da Wikipédia em português e 19 vezes em 7 páginas de Wikipédias de outros idiomas. Das categorias selecionadas para a pesquisa foi a que apresentou uma maior ocorrência de imagens ilustrando itens no Wikidata com 4 ocorrências. Das 471 imagens disponibilizadas na categoria, há 9 imagens sendo reutilizadas em 36 ocorrências, o que constitui 2,55% de reuso do total de imagens disponíveis na categoria.

A coleção referente ao Museu Castro Maya é a categoria que apresenta uma maior quantidade de imagens distintas utilizadas. A categoria disponibiliza um total de 46 imagens, das quais 17 imagens apresentaram ocorrência em 12 páginas da Wikipédia em português, em 25 páginas de Wikipédias de outros idiomas, em 5 páginas do Wikibooks e em 3 itens do Wikidata, somando assim 45 ocorrências de utilização de imagens, o que corresponde a um reuso de 36, 96% de todas as imagens disponíveis na categoria.

Apesar de apresentar somente 7 imagens na categoria referente ao Museu da Inconfidência, essa coleção é a que apresenta a maior porcentagem de ocorrência de usos das imagens disponíveis com o uso de 6 imagens distintas em 8 páginas da Wikipédia em português e em 10 artigos de outros idiomas, resultando na ocorrência de um total de 8 reusos de imagens. A reutilização das imagens das categorias referente às coleções de museus na Wikipédia

A reutilização das imagens das categorias referente às coleções de museus na Wikipédia

A Wikipédia é o projeto de enciclopédia online de edição coletiva e colaborativa da Fundação Wikimedia. Devido à sua proposta multilíngue, atualmente existem versões da Wikipédia em 303 idiomas. Cada idioma apresenta-se como um projeto Wikipédia independente do outro, ou seja, os conteúdos não são automaticamente traduzidos e reproduzidos, sendo originalmente criados e geridos pelos usuários de cada comunidade linguística e apresentam seu próprio domínio. A Wikipédia, em seus variados idiomas, tem seus conteúdos baseados na licença aberta Creative Commons *Attribution-ShareAlike*. Atualmente[6]a Wikipédia lusófona alcança uma marca de 1.029.760 artigos produzidos, 6.110 usuários ativos (WIKIPÉDIA, 2020).

Para entender melhor como se dá a reutilização dessas imagens tanto em páginas da Wikipédia lusófona, quanto em outros idiomas, buscamos coletar dados sobre a imagem da categoria referente ao Museu da República que foi reutilizada com maior frequência[7]. Por meio da ferramenta GLAMorous verificamos que a imagem com maior utilização é uma fotografia do presidente brasileiro Prudente de Morais, esse arquivo intitulado "Prudentedemorais.jpg" aparece com uma ocorrência 95 reutilizações.

Figura 1 – Retrato de Prudente de Morais na Wikimedia Commons

Fonte: Wikimedia Commons, 2020[8]

Ao observar a imagem no contexto da plataforma Wikimedia Common é possível verificar que a página da imagem oferece meios de download e compartilhamento da imagem e também apresenta uma série de sessões intituladas “Informações do arquivos” que fornece uma série de informações sobre a imagem como “Descrição”, “Data”, “Permissão”[9]e “Outras versões”. Além disso pode ser observada outras seções como “Histórico do arquivo” que mostra o histórico de substituição de arquivos no Wikimedia Commons, a seção “Uso de arquivos no Commons” que dá informações sobre o uso de versões no Commons, e a seção “Uso de arquivos em outras wikis” que gera uma lista de projetos wiki nos quais a imagem está sendo utilizada, e por fim a seção “Metadado” que apresenta metadados e dados específicos da mídia reprodutora da imagem, nesse caso específico à fotografia que resultou na imagem digitalizada.

Ao consultar a listagem na seção “Uso de arquivos de outras wikis” é possível observar em quais outros idiomas de Wikipédias o retrato de Prudente de Morais está sendo utilizado e em quais páginas específicas. Essas informações também podem ser obtidas por meio de consultas à ferramenta GLAMourous. Dessa forma foi possível verificar a presença do retrato de Prudente de Morais em artigos de Wikipédias em 44 idiomas distintos e também em uma página da Wikiquote em português. Também foi possível verificar que a mesma imagem está presente em 30 artigos da Wikipédia em português. No quadro abaixo é possível verificar a listagem desses artigos.

Quadro 2 – Lista de artigos que reutilizam a imagem o retrato de Prudente de Morais

pt.wikipedia	Prudente de Morais
	Rodrigues Alves
	Presidente Prudente
	Lista de eleições presidenciais no Brasil
	Presidente do Brasil
	Bernardino José de Campos Júnior
	Manuel José Alves Barbosa
	Eleição presidencial no Brasil em 1891

Amaro Cavalcanti

Manuel Vitorino

Joaquim Murtinho

Antônio Olinto dos Santos Pires

Antônio Gonçalves Ferreira

Dionísio Evangelista de Castro Cerqueira

Alberto Torres

Carlos Machado de Bittencourt

Carlos Augusto de Carvalho

Lista de presidentes do Senado Federal do Brasil

João Tomás de Cantuária

Jerônimo Rodrigues de Morais Jardim

Sebastião Eurico Gonçalves de Lacerda

Francisco de Paula Argolo

Bernardo Vasques

Elisiário José Barbosa

Eleição presidencial no Brasil em 1894

Lista de presidentes do Brasil

Governo Prudente de Morais

Lista de senadores do Brasil da 23.ª legislatura

Lista de senadores do Brasil da 22.ª legislatura

Fonte: Dados da pesquisa, 2020.

Percebe-se que a imagem é reutilizada sobretudo para ilustrar verbetes de personagens históricos que tiveram relação ou eventos importantes relacionados a Prudente de Morais. Além disso, alguns eventos históricos aparecem na lista, listas de senadores em legislaturas específicas, eleições presidenciais e o próprio verbete de seu governo. A imagem do acervo do museu é colocada em contexto a partir da textualidade dos verbetes e das relações simbólicas que estabelece com outras imagens e elementos textuais. Estudar essas relações disponibiliza uma camada informacional de compreensão do reuso da informação a ser desenvolvida em pesquisas posteriores, mas algo a destacar como potencial para a área.

A reutilização de imagens no repositório de dados estruturados Wikidata

Outro aspecto interessante a ser verificado é a reutilização das imagens de coleções do universo de museus e instituições culturais na ilustração de itens do Wikidata. O Wikidata é o projeto de repositório de dados estruturados da Fundação Wikimedia e tem "como objetivo a criação de uma base de conhecimento livre sobre o mundo que pode ser lida e editada por humanos e máquinas" (WIKIDATA, 2019). Segundo dados coletados em abril de 2020 na

página do
projeto, o Wikidata registrava 83.779.497 milhões de itens de dados estruturados em seu banco, 26.160 usuários ativos no último mês e 308 *bots* que auxiliam na edição. Os dados publicados no Wikidata também são disponibilizados de forma aberta sob a licença Creative Commons *Public Domain Dedication* 1.0. As regras de criação e gerenciamento do conteúdo são de responsabilidade da comunidade de editores, que devem seguir os mesmos princípios e políticas de colaboração norteiam os demais projetos Wikimedia.

Para visualizar e entender a reutilização de imagens da Wikimedia Commons na Wikidata selecionamos para a análise a coleção que apresentou, na Tabela 1, o número maior de ocorrências de uso de imagens no Wikidata, que foi o caso do Museu Histórico Nacional. Usando a ferramenta GLAMorous, e explorando os resultados detalhados retornado por meio de busca com o nome da categoria, foi possível verificar quais imagens foram utilizadas nas 4 ocorrências detectadas e também o código identificador do item no Wikidata, como pode ser visto no quadro a seguir.

Quadro 3 – Imagens reutilizadas na Wikidata

Nome do arquivo	ID Wikidata
Conselheiro Francisco de Carvalho Soares Brandão, da coleção Museu Histórico Nacional.jpg	da coleção Museu Histórico Nacional.jpg Fonte: Dados da pesquisa, 2020. Q62091980 Q62091982
Beatriz Reinall, da coleção Museu Histórico Nacional.jpg	Q62091983 Q62091547
Caminho na floresta, da coleção Museu Histórico Nacional.jpg	
Viscondessa do Bom Conselho,	

Para visualizar como se dá essa reutilização no contexto da plataforma Wikidata, realizamos buscas na plataforma por meio do identificador referente ao arquivo de imagem intitulada “Conselheiro Francisco de Carvalho Soares Brandão, da coleção Museu Histórico Nacional.jpg”. Desse modo foi possível observar que a mídia foi inserida na propriedade imagens de item do Wikidata que apresenta dados estruturados sobre a obra cuja a original analógica está sob guarda no acervo do Museu Histórico Nacional.

Figura 2 – Retrato de Prudente de Morais na Wikimedia Commons.

Fonte: Wikimedia Commons, 2020[10]

Ao verificar a reutilização do arquivo intitulado "Beatriz Reinall, da coleção Museu Histórico Nacional.jpg" foi possível identificar que essa imagem também ilustra um item na Wikidata que apresenta dados estruturados sobre a obra de arte retratada. Essa mesma situação se apresenta no caso das outras duas imagens identificadas que também ilustram itens sobre as obras originais. Com essas informações é possível inferir que foi realizado um trabalho de descrição de obras do Museu Histórico Nacional na Wikidata. Em uma investigação realizada em 2019, os autores Carmo e Martins (2019, p. 12) identificaram que o Museu Histórico Nacional representava o segundo maior acervo oriundo de instituições brasileiras com itens de pinturas descritas no Wikidata, apresentando 503 itens. Esses dados indicam que a instituição estabeleceu em algum momento, ou ainda estabelece, práticas que adotam as plataformas wikis da Fundação Wikimedia como meios estratégicos de disseminação de acervos.

A visualização de arquivos de mídia do Wikimedia Commons

Utilizando a ferramenta GLAMorous 2 foi possível obter dados acerca da visualização dos arquivos de mídias disponibilizados na Wikimedia Commons que estão em categorias referentes a coleções de museus geridos pelo Ibram. Essas visualizações são contabilizadas com base no acesso a páginas dos mais diversos projetos da Fundação Wikimedia, como o Wikidata, Wikibooks e Wikiquote e também como as Wikipédias dos mais variados idiomas que apresentam imagens do Wikimedia Commons reutilizadas em suas páginas. O recorte utilizado alcança dados de visualização relativos aos 12 meses do ano de 2019, de janeiro a dezembro. No quadro abaixo podemos verificar o total de visualizações de arquivos de mídia no ano de 2019 assim como a média mensal, também podemos observar o total de visualizações somente na Wikipédia em português e o total de visualizações em outras Wikipédias e projetos.

Tabela 2 – Reuso de imagens nos projetos da Fundação Wikimedia

Coleção	**Total de visitas - 2019**	**Média mensal**	**Total de visualizações pt wikipédia**	**Total de visualizações outras wikis**
Museu da Inconfidência			Museu da	

República	345.624,50	3.225.730	921.764	21.873	1.822,75
Museu Nacional de Belas Artes	20.904	969	882.608	73.550,67	795.007	87.601
Museus Castro Maya	6.216 518,00	6.216	0	0 - 0	0	69.089	5.757,42
Museu Casa de Benjamin Constant	44.791	24.298	578.668	48.222,33	45.643	45.643
Museu Regional de São João del-Rei	675.728	56.310,67	635.523	40.205	4.147.494
Museu Victor Meirelles	650.276	54.189,67	174.333	475.943
Museu Histórico Nacional					
Museu Imperial					

Fonte: Dados da pesquisa, 2020.

Com base nos dados coletados, é possível observar alguns números que nos chamam especial atenção. Tomando como exemplo os dados de visualização da coleção referente ao Museu da República, que com apenas 9 arquivos na categoria, foi a que gerou o número de visualizações de arquivos de mídias mais expressivo no ano de 2019, alcançando a marca de 4.147.494 milhões de visualizações. Ao recorrer a opção "Detalhes de uso do arquivo" na ferramenta GLAMorous 2 é possível observar a imagem identificada como "Prudentedemorais.jpg"[11], e já anteriormente identificada como a mídia que mais apresenta reusos dentro da coleção, se apresenta sozinha como a responsável por 3.096.852 do total de visualizações das mídias que estão na categoria da coleção do museu. Seguida da imagem intitulada "GetulioVargasPijamaRevolver.jpg", que apresentou no ano de 2019 um total de 992.687 visualizações. Também foi possível constatar que 3.225.730 milhões de visualizações das mídias da coleção do Museu da República foram somente em Wikipédias em português e 675.728 mil em Wikipédias de outros idiomas e em outros projetos da Wikimedia. A título de ilustração, para que se possa problematizar o fenômeno social em questão, no primeiro semestre do ano de 2019, o Museu da República teve em torno de 105.000 visitantes presenciais, segundo dados fornecidos pelo próprio Instituto Brasileiro de Museus (MINISTÉRIO DA CIDADANIA, 2019). Se tomarmos por partido a comparação com os números digitais, temos que o número de pessoas que visualizaram objetos do acervo presencialmente representa em torno de 2,5% dos que visualizaram os objetos digitais. Sem dúvida, e é fundamental destacar esse ponto, não se quer com isso fazer uma equivalência das experiências, sabendo que as visitas presenciais são mediadas de diversos outros elementos culturais que precisam ser reconhecidos e que possuem efeitos importantes na apropriação da visita. Mas, compreender esse espaço de socialização e sua dimensão simbólica como estratégia complementar para o Museu, é algo de enorme importância nos tempos atuais. Há um espaço de trabalho que pode ser aprimorado e que apresenta grande potencial de socialização e produção de valor social pelo museu.

Outro dado interessante foi a quantidade de visualizações de mídias referente a categoria do Museu Imperial que apresentou um maior número de visualizações em outros idiomas da Wikipédia, 475.943 mil, do que na Wikipédia em português com 174.333 mil visualizações. Ao verificar esses dados na aba "Uso global do arquivo" da ferramenta foi possível constatar que só a Wikipédia em inglês apresenta um total de 319.974 mil visualizações e faz 14 usos de 7 diferentes mídias da coleção do Museu Imperial em suas

páginas[12] no ano 2019.

Já as mídias da categoria referente ao Museu Benjamin Constant chamam atenção por não apresentarem visualizações em outros projetos da Wikimedia ou em Wikipédias de outros idiomas, somente na Wikipédia em português com 6.216 mil visualizações no ano de 2019, representando a coleção que menos obteve visualizações dentro do universo investigado.

CONCLUSÃO

O presente artigo apresentou resultados de uma investigação que buscou explorar possibilidades de coleta dados que permitissem a realização de análises webométricas que possibilitam mensurar aspectos relacionados ao acesso e à reutilização de mídias de acervos brasileiros disponíveis na internet, por meio do repositório de mídias Wikimedia Commons. Dessa forma, verificamos a presença de coleções de nove dos 30 museus sob responsabilidade do Ibram entre as categorias do Wikimedia Commons. Identificadas essas categorias, com o auxílio de ferramentas, obtivemos e apresentamos números relativos à reutilização dessas mídias tanto em páginas da Wikipédia tanto em português e em outros idiomas, assim como em outros projetos da Fundação Wikimedia. Assim pudemos observar que o número de mídias disponíveis não afeta diretamente o número de reutilizações ou de visualizações, já que o Museu da República se destacou tanto em número de reusos quanto em números de visualizações com apenas 9 arquivos na categoria referente a suas coleções, apesar do Museu Histórico Nacional apresentar o maior número de arquivos da amostra, 471, não obteve números relevantes nos dados de reuso ou de visualização dos mesmos. Também foi possível observar que uma mesma imagem de uma determinada coleção é utilizada diversas vezes gerando uma porcentagem baixa do total de uso dos arquivos, outra situação que se repetiu foi a baixíssima reutilização de mídias para ilustrar outros projetos da Fundação Wikimedia como o Wikidata. Já a presença de mídia e a quantidade de visualizações nas Wikipédias de diferentes idiomas foram dados que nos chamaram atenção, com ênfase no Museu da República e mais especificamente em um retrato do presidente brasileiro Prudente de Morais que aparece ilustrando 30 artigos da Wikipédia em português e 45 artigos de Wikipédias de outros idiomas. Outro dado curioso de ser mencionado foi o fato da Museu Imperial apresentar um maior número de visualizações na Wikipédia em inglês do que na Wikipédia em português no ano de 2019. Por fim, outro dado relevante foi o número de visualizações obtidos pelos arquivos na categoria referente a coleções do Museu da República, que concentrou em apenas uma imagem mais de 3 milhões de visualizações, somente no ano de 2019. Foi possível perceber evidências que apontam o quanto esse público digital do acervo é expressivo em relação ao público presencial, deixando elementos que podem ser explorados em pesquisas posteriores, procurando qualificar e entender que público é esse e os impactos que esse tipo de visitação gera em relação ao trabalho e modos de funcionamento das ações do museu.

Por meio dos dados obtidos e das análises realizadas foi possível observar que a plataforma Wikimedia Commons e outros projetos da Fundação Wikimedia, como a Wikipédia e Wikidata são espaços sociotécnicos que além de se apresentarem como meios potenciais de socialização de acervos e também se configuram como fonte de informação valiosa sobre a
manipulação e uso das informações disponibilizadas por meio deles. Esses dados podem ajudar instituições culturais guardiãs de acervos, como os museus, a entender melhor o que acontece com seus acervos digitais uma vez que são disponibilizados nesses espaços, fornecendo dados importantes para a gestão das coleções de seu acervo digital,

principalmente em relação ao acesso e reapropriação desses acervos por parte dos usuários.

REFERÊNCIAS

BJÖRNEBORN, L., INGWERSEN, P. Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, p.1216–1227, 2004, doi:10.1002/asi.20077. Disponível em: <https://doi.org/10.1002/asi.20077> Acesso em 2 abr. 2020.

CARMO, D; MARTINS, D.L. A presença dos museus brasileiros na ecologia informacional da Fundação Wikimedia: estudo de caso do projeto Sum of all Paintings. *Encontro Nacional de Pesquisa em Ciência da Informação,* n. XX ENANCIB, 2019.

INSTITUTO BRASILEIRO DE MUSEUS. *Museus Ibram*. Disponível em: <https://www.museus.gov.br/os-museus/museus-ibram/> Acesso em 5 abr. 2020.

MANSKE, M. *GLAMourous*. 2020. Disponível em: <https://tools.wmflabs.org/glamtools/glamorous.php>. Acesso em 18 abr. 2020.

MANSKE, M. *GLAMorous* 2. 2020. Disponível em: <https://tools.wmflabs.org/glamtools/glamorous/>. Acesso em 18 abr. 2020.

MINISTÉRIO DA CIDADANIA/SECRETARIA ESPECIAL DE CULTURA (BRASIL). *Plano Nacional de Cultura. Museus brasileiros apresentam aumento no número de visitantes.* [S. l.], 19 ago. 2019. Disponível em: <http://pnc.cultura.gov.br/2019/08/19/museus-brasileiros-apresentam-aumento-no-numero de-visitantes/>. Acesso em: 30 abr. 2020.

STINSON, A.D.; FAUCONNIER, S; WYATT, L. Stepping Beyond Libraries: The Changing Orientation in Global GLAM-Wiki. *JLIS.it* 9, 3 set. 2018. Disponível em: <https://www.jlis.it/article/download/12480/113>. Acesso em: 18 jul. 2019.

THELWALL, M. *Introduction to Webometrics: Quantitative Web research for the social sciences*. New York: Morgan & Claypool, 2009.

VANTI, N.A.P. Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. *Ci. Inf.*, Brasília , v. 31, n. 2, p. 369-379, 2002. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000200016&lng=en&nrm=iso>. Acesso em: 18 abr. 2020.

WIKIMEDIA COMMONS. *Special: Statistics*, 2020a. Disponível em: <https://commons.wikimedia.org/wiki/Special:Statistics>. Acesso em: 18 abr. 2020.

WIKIMEDIA COMMONS. *Commons: Welcome*, 2020b. Disponível em: <https://commons.wikimedia.org/wiki/Commons:Welcome>. Acesso em: 15 abr. 2020.

WIKIMEDIA OUTREACH. *GLAM/Resources/Tools*. 2020. Disponível em: <https://outreach.wikimedia.org/wiki/GLAM/Resources/Tools>. Acesso em 20 mar. 2020.

WIKIPEDIA. *Página Principal*. 2019. Disponível em: <https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:P%C3%A1gina_principal>. Acesso em 18 out. 2020.

ZEINSTRA, M. *Report on Requirements for Usage and Reuse Statistics for GLAM content. Kennisland,* 2013. Disponível em: <https://www.kl.nl/wp- content/uploads/2014/09/report on-requirements-for-usage-and-reuse-statistics-for-glam-content.pdf>. Acesso em: 20 set. 2019.

Workflow de agregação de dados: processos para criação de uma interface de busca integrada do patrimônio cultural

Submetido em: 28/04/2020. Aprovado em: dd/mm/yyyy. Publicado em: dd/mm/yyyy .

Resumo

Nos últimos anos diferentes instituições culturais vêm envidando esforços para difundir a cultura por meio da construção de uma interface única de busca, que integre seus objetos digitais e facilite a recuperação de dados para os usuários leigos. Contudo, integrar dados culturais não é uma tarefa trivial, pois estes dados são diversos e singulares, necessitando assim, de uma variedade de etapas entre a coleta e a apresentação dos dados. Dessa forma, esta pesquisa visa localizar *workflows* de agregação e discutir as etapas propostas. Para tal, realizou-se pesquisa descritiva e bibliográfica, de natureza qualitativa, em bases de dados acadêmicas e na literatura cinzenta. Como resultado, apresenta-se oito projetos: American Art Collaborative, DigitalNZ, D-NET Software, Europeana, Mexicana, Parthenos Aggregator, TROVE e UNLV's Linked Data Project. A análise do conjunto de *workflows* resultou em oito diferentes etapas a serem executadas: 1. Extrair, 2. Estruturar, 3. Transformar, 4. Reconciliar, 5. Armazenar, 6. Publicar, 7. Expor e 8. Possibilitar novas aplicações. Além disso, também é visível a necessidade de maior detalhamento das etapas, a fim de que seja possível replicar o *workflow*, e usufruir de seus benefícios em outras instituições.

Palavras-chave: Agregação de dados. Busca integrada. Patrimônio cultural. Recuperação de Informação. *Workflow*.

Data aggregation workflow: processes for creating an integrated search interface for cultural heritage

Abstract

Institutions have been making efforts to spread culture through the construction of a unique search interface. The interface integrates digital objects and facilitates data retrieval for lay users. However, integrating cultural data is not a trivial task. Cultural data is diverse and unique, thus requiring a variety of steps between data collection and presentation. Thus, this

research aims to locate aggregation workflows, and to discuss the proposed steps. To this end, descriptive and bibliographic research, of a qualitative nature, was carried out in academic databases and in the gray literature. As a result, eight aggregation workflows proposed by: American Art Collaborative, DigitalNZ, D-NET Software, Europeana, Mexicana, Parthenos Aggregator, TROVE and UNLV's Linked Data Project. For each of these institutions, the workflow and the details of the stages were presented, when possible. The analysis of the set of workflows resulted in eight different steps to be performed: 1. extract, 2. structure, 3. transform, 4. reconcile, 5. store, 6. expose, 7. publish and 8. enable new applications. In addition, the need for more detailed stages is also viable, so it's possible to replicate the workflow, and enjoy its benefits in other institutions.

***Keywords**: Data aggregation. Integrated search. Cultural heritage. Information Retrieval. Workflow.*

Flujo de trabajo de agregación de datos: procesos para crear una interfaz de búsqueda integrada para el patrimonio cultural

Resumen

En los últimos años, diferentes instituciones culturales han estado haciendo esfuerzos para difundir la cultura a través de la construcción de una interfaz de búsqueda única, que integra sus objetos digitales y facilita la recuperación de datos para los usuarios legos. Sin embargo, la integración de datos culturales no es una tarea trivial, ya que estos datos son diversos y únicos, por lo que requieren una variedad de pasos entre la recopilación y presentación de datos. Por lo tanto, esta investigación tiene como objetivo localizar los flujos de trabajo de agregación y discutir los pasos propuestos. Para ello, se realizó una investigación descriptiva y bibliográfica, de carácter cualitativo, en bases de datos académicas y en la literatura gris. Como resultado, se presentan ocho proyectos: American Art Collaborative, DigitalNZ, D-NET Software, Europeana, Mexicana, Parthenos Aggregator, TROVE y UNLV's Linked Data Project. Presentó como resultado ocho pasos diferentes a realizar: 1. extraer, 2. estructura, 3. transformar, 4. conciliar, 5. almacenar, 6. exponer, 7. publicar y 8. habilitar nuevas aplicaciones. Además, también es visible la necesidad de etapas más detalladas, de modo que sea posible replicar el flujo de trabajo y disfrutar de sus beneficios en otras instituciones.

***Palabras clave**: Agregación de datos. Búsqueda integrada Patrimonio cultural. Recuperación de información. Flujo de trabajo.*

INTRODUÇÃO

Instituições culturais estão, a cada dia, se reinventando e inovando suas formas de interagir com público, com destaque, a disponibilização de objetos digitais e informações sobre esses objetos por meio de metadados em sites e/ou repositórios institucionais, como um meio para exercer sua prática comunicacional, assim como, difundir seus acervos digitalizados.

Essa realidade fez explodir, no Brasil e no mundo, a quantidade de objetos na rede,

resultando em uma nova problemática: como permitir que os usuários, principalmente os leigos, encontrem o objeto de seu interesse, em meio a tanta oferta e diferentes mecanismos de busca?

De forma ampla, a resposta a esta pergunta foi oferecer uma interface de busca integrada, que agrega um conjunto específico de bancos de dados, capaz de recuperar, mais facilmente, o objeto desejado. Com esse intento, surgiram as ferramentas de busca federada, que realizam uma pesquisa simultânea em diversas fontes, apresentando os resultados em uma lista única. Nos anos 2000, algumas bibliotecas adotaram a pesquisa federada, porém, com o tempo, tornou-se evidente uma série de problemas, tais como, lentidão nos tempos de resposta, resultados duplicados e a impossibilidade de refinamento dos resultados (BRIGHAM et al., 2016; PAVÃO; CAREGNATO, 2015).

Dessa forma, difundir a cultura por meio da oferta de uma interface de busca integrada, com uma navegação eficiente ainda é um objetivo fortemente almejado e, falando especificamente de Brasil, algo que ainda não foi realizado em ampla escala e que poderia contribuir, de forma significativa, para outras formas de socialização da cultura brasileira.

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. Pois, apesar de diversos padrões de metadados, modelos conceituais e regras de catalogação, tais como CIDOC-CRM, EDM, LRM, entre outros, existirem para a área da cultura, os mesmos nem sempre são consensuados e se encontram aplicados em níveis muito diferentes de interiorização pelas instituições. Cabe ressaltar, a título de explicitação, que se considera neste trabalho que agregação de dados envolve a agregação de metadados mais a agregação dos objetos digitais descritos por esses metadados.

Dessa forma, para efetivação da integração dos dados, são necessárias uma variedade de etapas entre a coleta e a apresentação dos dados para os usuários. Com esse objetivo, 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, fornece acesso a 58 milhões de objetos digitais, com sofisticadas ferramentas de pesquisa e filtro, além de coleções temáticas, exposições, galerias e blogs (EUROPEANA PRO, 2020). A Europeana é um caso mundialmente conhecido, faz parte dos resultados deste estudo, contudo, outras instituições também realizam pesquisa na área e oferecem soluções para agregação de dados, considerando diferentes realidades.

Dessa forma, o objetivo deste estudo é localizar *workflows* de agregação de dados culturais, para realizar uma análise qualitativa das etapas escolhidas por cada instituição, por meio de pesquisa de caráter descritivo e bibliográfico, de natureza qualitativa, realizada em bases de dados acadêmicas e na literatura cinzenta.

Ao final, foram localizados oito *workflows* que são apresentados na seção de Resultados, assim como a descrição das etapas. Este artigo está assim dividido: seção 2, Metodologia, seção 3, Análise e Discussão dos Resultados, e por último, na seção 4, as Conclusões.

METODOLOGIA

Pesquisa de caráter descritivo e bibliográfico, de natureza qualitativa, realizada em bases de dados acadêmicas e na literatura cinzenta, com o intuito de encontrar *workflows* de

agregação de dados culturais.

As buscas foram realizadas no Google, Google Acadêmico, EBSCOhost e BRAPCI, utilizando as palavras: *"pipeline", "architecture", "aggregation", "metadata ingest", "metadata aggregation", "aggregative data infrastructures"* e suas versões em português. Além destes, também foram realizadas pesquisas por meio de projetos de agregação de dados conhecidos, como: *"Europeana", "Mexicana", "Digital Public Library of America", "Trove" e "DigitalNZ"*. Cabe dizer que estes projetos foram escolhidos por serem os principais agregadores de dados culturais nas diferentes regiões do mundo conforme apresentado por Navarrete (2016).

Optou-se pelo Google para localizar *workflows* na literatura cinzenta, a BRAPCI, por ser específica da área de Ciência da Informação no Brasil e a EBSCOhost pela produção científica internacional, tornando a pesquisa mais ampla.

ANÁLISE E DISCUSSÃO DOS RESULTADOS

As pesquisas resultaram em sete instituições, listadas no Quadro 1, cujos *workflows* são apresentados e detalhados nesta seção.

Quadro 1 - Instituições, países de origem e seus projetos

N.	Instituição	País	Projeto
01	American Art Collaborative	EUA	AAC
02	Biblioteca Nacional	Austrália	Trove
03	Biblioteca Nacional	Nova Zelândia	DigitalNZ
04	Fundação Europeana	União Europeia	Europeana
05	Instituto de Ciência e Tecnologia da Informação		Aggregator
06	Instituto de Ciência e Tecnologia da Informação	Itália	D-NET Software
		Itália	Parthenos
07	Secretaria de Cultura	México	Repositório Mexicana
08	Universidade de Nevada	EUA	UNLV's Linked Data Project

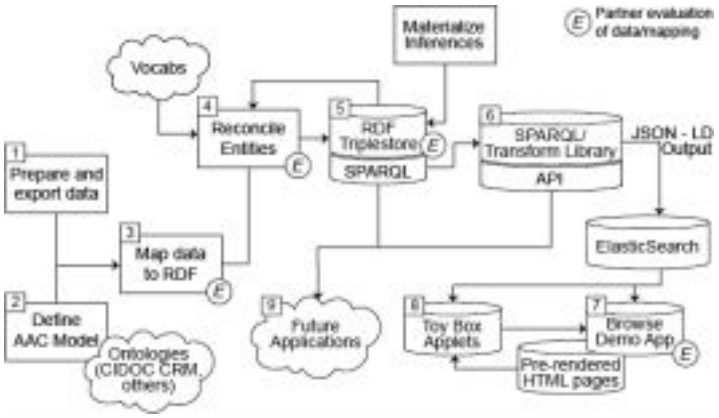
Fonte: elaborado pelos autores

American Art Collaborative - AAC Pipeline

A American Art Collaborative – AAC, é um consórcio de 14 instituições de arte que visam investigar e começar a construir uma massa crítica de *Linked Open Data* – LOD. Para

Fink (2018), LOD se trata de um método para publicar dados estruturados na web de forma que as informações sejam interconectadas e, assim, tornadas amplamente úteis. A Figura 1, apresenta o *workflow* proposto pela AAC.

Figura 1 - *Workflow* de agregação da AAC



Fonte: Fink (2018, p. 32). Adaptada.

O *workflow* prevê nove etapas. A Etapa 1. "*Prepare and export data*", em tradução livre, "Preparar e exportar os dados", visa fornecer dados principais e dados adicionais, considerados úteis pelos parceiros, que, na prática, exportam dados brutos de seus sistemas de origem e os carregaram em um repositório GitHub compartilhado. A etapa 2. "*Define AAC Model*" ou "Definir o Modelo AAC", se trata de um conjunto geral de orientações sobre ontologias que podem ser adotadas e reutilizadas, tendo por objetivo constituir um modelo conceitual único para agregar os dados das diferentes instituições. A etapa 3. "*Map data to RDF*" ou "Mapear dos dados para RDF", visa mapear os dados das instituições parceiras para um modelo de destino e fornecer flexibilidade para usuários adicionais, assim, usando o modelo de destino e a ferramenta de integração de dados Karma, os dados de cada parceiro são convertidos em RDF. A etapa 4. "*Reconcilie Entities*" ou "Reconciliar entidades" visa mapear entidades individuais para IDs comuns, considerando, sempre que possível, vocabulários públicos padronizados. A etapa 5. "*RDF Triple Store - SPARQL*" ou "Armazenar dados RDF em um Triple Store - SPARQL", visa armazenar e fornecer acesso aos dados vinculados enriquecidos pela reconciliação de entidades e permite consultas SPARQL. A etapa 6. "*SPARQL/Transform library - API*" ou "SPARQL/API para transformação de Bibliotecas", visa compilar um conjunto de consultas claras, reutilizáveis e focadas em entidades que navegam no gráfico de dados vinculados para fornecer documentos JSON-LD para desenvolvedores e humanistas digitais. A etapa 7. "*Browse Demo App*" ou "Aplicações via browsers", visa apresentar uma ilustração inicial suficientemente rica de dados vinculados para os usuários. A etapa 8. "*Toy Box Applets*" ou "Ampliação das aplicações via browser", visa
ampliar as possibilidades e ideias do AAC, por meio de um aplicativo de navegação. A etapa 9. "*Future Applications*" ou "Produzir aplicações futuras" visa continuar a explorar casos de uso e aplicativos para dados vinculados por meio de contribuições de dados de parceiros (FINK, 2018).

Biblioteca Nacional da Austrália – Trove

A Biblioteca Nacional da Austrália desenvolveu o Trove, que objetiva disponibilizar recursos culturais relacionados à Austrália. O Trove oferece um mecanismo de busca integrada em bibliotecas, museus, arquivos e outras organizações de pesquisa, além de um conjunto de serviços (TROVE HELP CENTRE, 2020). A Figura 2 apresenta o *workflow* proposto.

Figura 2 - *Workflow* de agregação da Trove

Fonte: National Library of Australia (2010). Adaptada.

O Trove mantém um site com informações relevantes do projeto, no entanto, ainda que apresente o *workflow* de agregação, não traz, nas fontes pesquisadas, documentação que explique cada etapa do *workflow*.

Contudo, na análise realizada pela pesquisa, percebe-se um grande uso do protocolo OAI-PMH para coleta de dados dos provedores, bem como a criação de diferentes índices para indexação ágil e recuperação da informação utilizando a tecnologia Apache Lucene, que é um software voltado para busca e indexação de documentos de alta escalabilidade e aplicado em projetos que exigem processamento de dados massivos. A arquitetura se vale de diversas camadas, mas devido à falta de documentação explícita, somente se pode inferir como as
camadas se relacionam, sem condições de uma análise crítica da solução para eventual replicação.

Biblioteca Nacional da Nova Zelândia – DigitalNZ

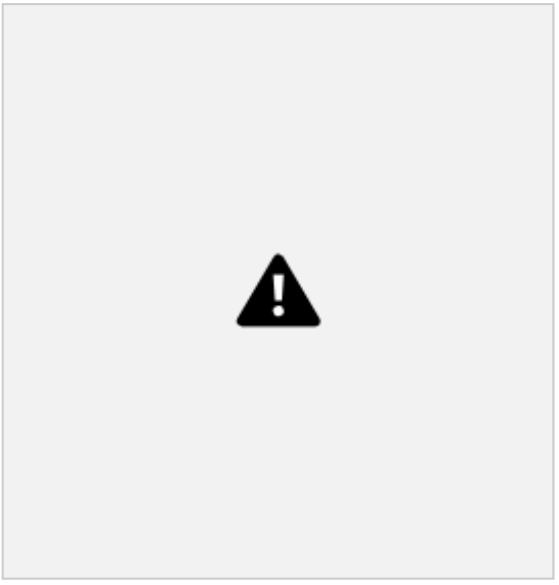
A Biblioteca Nacional da Nova Zelândia junto a Rede do povo Aotearoa Kaharoa desenvolveu, no início de 2006, o DigitalNZ, que utiliza o software Supplejack para agregação de dados (DIGITAL NEW ZEALAND, 2019). A Figura 3, apresenta uma ilustração que demonstra como a agregação da DigitalNZ acontece.

Figura 3 - *Workflow* de agregação proposto pela DigitalNZ

Fonte: Digital New Zealand (2018). Adaptada.

A Figura 3 mostra a ferramenta Supplejack como ferramenta central para agregação dos dados, dessa forma, a Figura 4, apresenta a arquitetura da plataforma Supplejack.

Figura 4 - Arquitetura da plataforma Supplejack



Fonte: Supplejack (2020). Adaptada.

A arquitetura é composta por: *Manager* ou Gerenciador, que apresenta uma interface para o usuário controlar as atividades do software; *Worker* ou Trabalhador, que realiza as atividades de coleta, enriquecimento e verificação de links; *API*, um *wrapper* público para pesquisar o repositório de índice e metadados; *Common* ou Comum, são ajudantes compartilhados entre o *Worker* e o *Manager* (SUPPLEJACK, 2020).

Como apresentado na Figura 4, o Supplejack depende da integração com um índice de pesquisa, o padrão é Solr, e um repositório de metadados, o padrão é MongoDB. O Apache SOLR trata-se de tecnologia voltada para pesquisa e indexação de documentos massivos, e MongoDB, um banco de dados do tipo NoSQL, também utilizado em projetos contemporâneos que envolvem novas arquiteturas para processamento de dados massivos baseados em informação semiestruturada ou mesmo desestruturada.

Fundação Europeana – Europeana

A Fundação Europeana desenvolveu a Europeana, que reuniu mais de 55 milhões de objetos digitais das coleções on-line de mais de 3.500 galerias, bibliotecas, museus, coleções audiovisuais e arquivos de toda a Europa (SCHOLZ, 2019). A Figura 5 apresenta seu *workflow* de agregação.

Figura 5 - *Workflow* de agregação de dados proposto pela Europeana

Fonte: Kollia et al. (2012, p. 70). Adaptada.

A primeira etapa, "*Harvesting, Delivery*" ou "Colheita, Entrega", se refere a coleta de metadados de provedores de conteúdo, por meio de protocolos de entrega, como o OAI-PMH, HTTP e FTP. A segunda etapa, "*Schema Mapping*", ou "Mapeamento de esquema", alinha os metadados coletados a um modelo de referência comum. Nesta etapa, uma interface gráfica colabora com o mapeamento, por meio de uma linguagem de mapeamento compreensível por máquinas. A terceira etapa, "*Value Mapping*" ou "Mapeamento de valores", foca no alinhamento e transformação dos termos constantes nos metadados coletados para arquivo de autoridade ou fonte externa, ou seja, permite a normalização de datas, localizações, países, idiomas, dentre outros. A quarta etapa, "*Revision, Annotation*", ou "Revisão, Anotação", permite adicionar anotações para atribuir metadados não disponíveis no contexto original, e por último, na quinta etapa, "*Semantic Enrichment*" ou "Enriquecimento semântico", foca na transformação dos dados em um modelo semântico, extração e identificação de recursos e implantação em RDF (SCHOLZ, 2019).

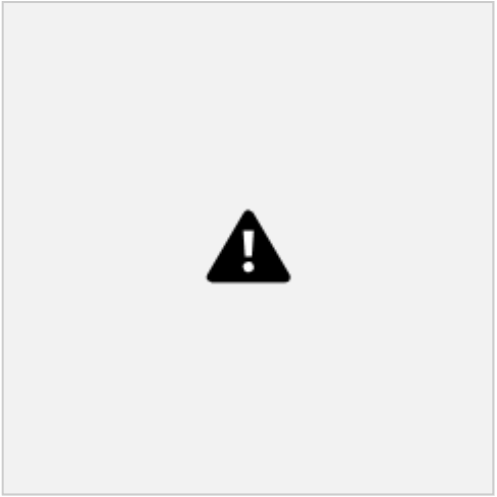
A documentação da Europeana não entra em detalhes tecnológicos específicos, não ficando claro que ferramentas e tecnologias são utilizadas em cada etapa, como as mesmas são parametrizadas e que esforços foram desenvolvidos para a integração dos serviços. Novamente, há dificuldade de se encontrar evidências que facilitem a replicação ou mesmo adaptação de soluções em outros contextos.

Instituto de Ciência e Tecnologia da Informação - D-NET

O Istituto di Scienza e Tecnologie dell'Informazione desenvolveu o D-NET, uma estrutura orientada a serviços, de uso geral, na qual os designers podem construir infraestruturas agregadas autônomas, robustas, escaláveis e personalizadas, de maneira econômica. Oferece serviços de gerenciamento de dados capazes de fornecer acesso a diferentes tipos de fontes de dados externas, armazenar e processar objetos de informações de qualquer modelo de dados, convertê-los em formatos comuns e expor objetos de informações a aplicativos de terceiros por meio de vários acessos padrão (MANGHI et al., 2014).

Neste estudo, categorizamos D-NET e o Supplejack em um mesmo nicho, visto ambos serem softwares que tem por objetivo fornecer um serviço completo para agregação de dados. A Figura 6 apresenta a arquitetura do software.

Figura 6 - Infraestrutura D-NET Software Toolkit



Fonte: Manghi et al. (2014, p. 327)

O D-NET subdivide sua arquitetura em 4 camadas principais, *Manipulation, Storage, Provision* e *Mediation*. Os serviços de manipulação, *Manipulation*, foram projetados para executar o enriquecimento, a validação, o espelhamento e geração de estatísticas. Os serviços de armazenamento, *Storage*, como o próprio nome diz, propiciam o armazenamento de objetos, agrupando tecnologias conhecidas, de código-fonte aberto, como índices de texto

completo, bancos de dados relacionais, repositórios de documentos, etc. Os serviços de fornecimento de dados, *Provision*, fazem interface com aplicativos externos, como por exemplo, portais para uso dos usuários finais ou serviços de terceiros. Além do acesso aleatório, o D-NET suporta as APIs: OAI-PMH *publisher service* e OAI-ORE *publisher service*. Os serviços de mediação, "*Mediation"*, visam buscar dados de fontes externas e importá-los para a infraestrutura agregada, como por exemplo, objetos em conformidade com um determinado recurso de modelo de dados (BARDI, MANGHI E ZOPPI, 2012).

Conforme supracitado, o Supplejack foi utilizado pela DigitalNZ. Nesse estudo, citamos o Parthenos Aggregator, que utiliza do D-NET.

Instituto de Ciência e Tecnologia da Informação - Parthenos Aggregator

As infraestruturas de humanidades digitais (DHIs) apoiam pesquisadores no campo das ciências humanas, oferecendo um ambiente digital, no qual podem encontrar e usar ferramentas e dados de pesquisa para conduzir suas atividades. Há um número crescente de DHIs e para integrá-los, à Comissão Europeia lançou o projeto *Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies* - PARTHENOS (FROZINI et al, 2018). A Figura 7 apresenta o *workflow* proposto.

Figura 7 - *Workflow* de agregação e provisão Parthenos Aggregator

Fonte: Frosini et al. (2018, p. 40). Adaptada.

O *workflow* se divide em dois fluxos de trabalho: agregação e provisionamento. No fluxo de agregação, a etapa "*Collection*" ou "Coleta" visa lidar com a coleta de metadados por meio de diferentes protocolos de acesso: OAI-PMH, FTP(S), SFTP, HTTP(S), RESTful. A etapa "*Transformation*" ou "Transformação" visa mapear os metadados a uma ontologia única. A "*Metadata Cleaner*" ou "Limpeza de metadado", se trata do serviço que harmoniza valores
em registros de metadados com base em um conjunto de tesauros. Após esta etapa, inicia-se o processo de inspeção, na etapa "*Metadata Record Inspector*", ou "Inspeção de registro de metadados", na qual uma GUI da Web integrado ao D-NET, fornece dados aos curadores com

uma visão geral das informações, possibilitando pesquisas e navegação entre os registros para verificar a correção da fase de transformação (por exemplo, metadados sem mapeamento, erros ou inconsistências semânticas) e a fase de limpeza. Uma vez positivamente verificado, os registros podem ser exportados publicamente na "*OAI-PMH publisher service*", o serviço oferece Interfaces OAI-PMH para aplicativos de terceiros que desejam acessar metadados. "*Index service*" o serviço orienta a alimentação de Índices Solr e também é responsável por transformando os registros de metadados agregados em documentos Solr (FROSINI et al., 2018).

Secretaria de Cultura do México – Mexicana

A Secretaria de Cultura do México desenvolveu a Mexicana, um Repositório do Patrimônio Cultural do México, livre e aberto, que tem o objetivo principal de difundir e vincular os acervos do patrimônio cultural do México (SECRETARÍA DE CULTURA, 2018). Seu *workflow* está apresentado na Figura 8.

Figura 8 - *Workflow* de agregação Mexicana

Fonte: Secretaria de Cultura (2018). Adaptada.

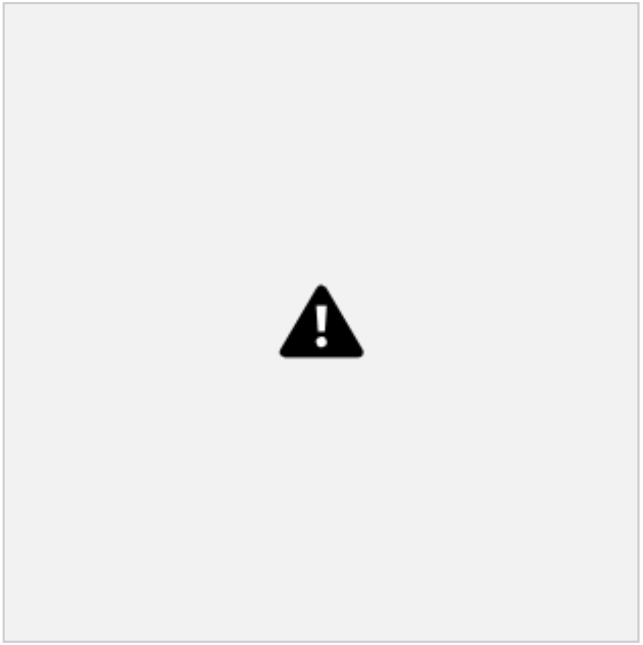
A Secretaria de Cultura do México desenvolveu documento explicativo sobre o projeto Mexicana, no qual consta o *workflow* apresentado, que é dividido em *Back End* e *Front End. No Back End,* a Etapa 1. *"Extractores*" ou "Extratores", se trata dos componentes de código responsáveis pela cópia de dados locais ou remotos, realizando uma reestruturação mínima dos dados. Nesta etapa, se considera a criação de extratores para diferentes formatos e a criação de interfaces para facilitar a extensão da funcionalidade. A Etapa 2. "*Cosechadores*" ou "Coletores", trata-se de componente de código configurável que permite gerenciar os extratores
e o mapeador dos dados (próxima etapa), de acordo com regras estabelecidas, tais como: execução sob demanda e por periodicidade. A Etapa 3. "*Mapeador*", permite configurar e executar regras de mapeamento definidas entre os dados gerados pelos extratores e o esquema

de dados unificados do sistema. A Etapa 4. "*Almacenamiento*" ou "Armazenamento", como o próprio nome diz, visa armazenar os dados coletados. Nessa etapa, no *workflow*, há dois itens além do armazenamento, o "*Índices de búsqueda*" ou "Índices de pesquisa", no qual o componente explora os serviços do ElasticSearch para gerar índices de pesquisa de metadados de objetos incorporados ao armazenamento do sistema e a "*Réplica para API*", que replica os objetos digitais para fornecer seu acesso rápido através do APIs de exposição. A Etapa 5. "API", reforça o uso servidor ElasticSearch para indexação e recuperação de metadados, assim como a "API de Búsqueda" ou "API de Exposição" trata sobre formatos para exibição de objetos digitais e seus metadados. No *Front end,* etapa voltada para visualização dos dados, trata-se do "Buscador" ou "Pesquisa", que trata sobre pesquisa desagregada de objetos digitais inserido no sistema através da exploração da API; "*Exhibiciones*" ou "Exposições", que permite a configuração de coleções de objetos digitais agrupados por tema ou evento específico; "*Detalle de objetos*" ou "Detalhe do objeto", que permite visualizar o arquivo detalhes dos objetos digitais inseridos no sistema e no metadados associados; o "*EndPoint SPARQL*", que permite a execução de consultas SPARQL para o modelo de dados do sistema unificado através da estrutura semântica definida pelo Modelo de Dados e por fim, "*Administración*" ou "Administração", que permite a administração de usuário e fluxo de trabalho para o configuração e execução das coletas para fontes de dados.

Universidade de Nevada - UNLV's Linked Data Project

A Universidade de Nevada, por meio da equipe do departamento de Coleções Digitais das Bibliotecas da Universidade, reuniu esforços para encontrar maneiras de tornar mais eficiente a descoberta e uso das informações, iniciando assim estudos para adoção do Linked Open Data – LOD, culminando no desenvolvimento do *UNLV's Linked Data Project* (SOUTHWICK, 2015). A Figura 9, apresenta o *workflow*.

Figura 9 - *Workflow* de agregação de dados proposto pela Universidade de Nevada



Fonte: Southwick (2015, p. 13). Adaptada.

Southwick (2015) afirma que para o desenvolvimento do projeto optou-se pela adoção de tecnologias com código aberto, sem qualquer adaptação ou desenvolvimento, ou seja, "no estado em que se encontram".

O projeto foi dividido em cinco etapas*: Planning, Designing, Implementing, "Publishing LOD" e "Consuming LOD"*, ou seja, Planejamento, Concepção, Implementação, "Publicando LOD" e "Consumindo LOD". O Planejamento é composto por duas revisões de literatura e retrata o período de estudo necessário ao desenvolvimento do protótipo.

A segunda etapa, Concepção, se desdobra em três atividades: "*Evaluate e select technologies*" ou "Avaliar e selecionar as tecnologias", na qual a equipe do projeto avaliou várias tecnologias e selecionou seis para aplicação no protótipo. A autora destaca que, embora as tecnologias selecionadas tenham funcionado bem, não significa que são as únicas ou as melhores; "*Select data model"* ou "Selecionar modelo de dados", que diz respeito à seleção da ontologia utilizada, a partir da investigação de modelos de dados utilizados por outras instituições; e "*Definition of local URIs strutucture*" ou "Definir a estrutura das URIs", fase na qual optou-se por criar URIs apenas para "coisas" que ainda não receberam URIs de outros provedores de dados. Se posteriormente forem descobertas URIs diferentes atribuídas à mesma "coisa" à qual já atribuímos um URI, elas são adicionadas ao conjunto de triplas que indicam a equivalência entre URIs.

A terceira fase, a Implementação, é composta por seis etapas: "*Clean and export metadata*" ou "Limpar e exportar metadados": o CONTENTdm é utilizado como sistema de gerenciamento de conteúdo e este exporta metadados em formato de planilha delimitada por tabulação, que pode ser importado para o OpenRefine. A limpeza consiste em cumprir rigorosamente os termos usados nas coleções que adotam um determinado vocabulário controlado e criar vocabulários controlados locais consistentes; "*Prepare metadata for transformation*" ou "Preparar metadados para transformação", fase de preparação dos metadados para gerar LOD, para tal, utilizou-se funções OpenRefine, como: remover espaços em branco; separar tipos diferentes de dados e separar valores agrupados em um único campo; "*Reconcilie with controlled vocabularies*" ou "Reconciliar com vocabulários controlados", trata-se da reconciliação feita usando a extensão "OpenRefine RDF", do OpenRefine e "*Generate RDF files*" ou "Gerar arquivos RDF", como o próprio nome diz, trata-se da geração de arquivos RDF que serão utilizados nas próximas etapas. As etapas "*Create new URIs for local controlled vocabularies*" e "*Implement mapping*" são a efetivação das etapas "*Definition of local URIs structure*" e "*Select data model*", da fase da Concepção.

A quarta fase, a Publicação, é composta por uma única atividade, "*Publish Linked Open Data*" ou "Publicar no Linked Open Data – LOD", na qual, o arquivo RDF é publicado à comunidade. A última fase, Consumo, também é composta por uma única atividade, "*Apply visualization tools*" ou "Aplicar ferramentas de visualização", na qual foram realizados três experimentos com ferramentas de visualização para arquivos RDF: com Pivot Viewer que foi útil para visualizar imagens de maneira muito dinâmica, pois é baseado em consultas SPARQL; com RelFinder, que visa encontrar relacionamentos entre as "coisas" e; também com RelFinder, mas com o conceito de relacionamento expandido, considerando relacionamentos que duas "coisas" tinham uma ou mais "coisas" em comum.

As interfaces de busca integradas estão disponíveis na rede e o Quadro 2, apresenta o link para cada uma delas.

Quadro 2 – Links das interfaces de busca integradas

Projeto	**Sites**
AAC	https://americanart.si.edu/search
Trove	https://trove.nla.gov.au/
DigitalNZ	https://digitalnz.org/
Europeana	https://www.europeana.eu/pt/collections
D-NET Software - Parthenos Aggregator	http://www.parthenos-project.eu/portal
Repositório Mexicana	https://mexicana.cultura.gob.mx/
UNLV's Linked Data Project	https://www.library.unlv.edu/linked-data

Fonte: elaborado pelos autores

Considerando a análise de todos os *workflows* apresentados acima, foram encontradas oito fases para agregação, sendo elas: extração, estruturação, transformação, reconciliação, armazenamento, exposição, publicação e novas aplicações. De forma sintética, estas etapas significam:

1. Extrair: extração dos dados em sua forma bruta, que podem estar, por exemplo, em pdf, em planilhas eletrônicas, documentos de texto, XML (*eXtensible Markup Language*), em bancos de dados relacionais, dentre outras opções.
2. Estruturar: selecionar vocabulários controlados pré-existentes e ontologias para aplicação nos dados.
3. Transformar: realizar a normalização, limpeza e correção sintática dos dados. 4. Reconciliar: enriquecer os metadados por meio de outros dados existentes na web. 5. Armazenar: se trata da escolha de onde os dados coletados serão armazenados. 6. Publicar: se trata da interface única de busca integrada.
7. Expor: disponibilizar os dados agregados por meio de API, que exponham os dados em formato RDF, OAI-PMH ou JSON.
8. Possibilitar novas aplicações: a partir dos arquivos disponibilizados na etapa ‘Expor’, considerar que novas aplicações podem ser criadas.

O Quadro 3 mostra um resumo individual, considerando a presença (X) ou não (-) de cada etapa, para visualização geral.

Quadro 3 - Etapas dos *Workflows* de Agregação, panorama individual

Projeto/ Etapas	**Extrair**	**Estruturar**	**Transformar**	**Reconciliar**	**Armazenar**	**Publicar**	**Expor**	**Novas aplicações**

AAC X X - X X X X X Trove [1] X - - - X X - -	
DigitalN Z	somente a partir da
European a [2]	X - - - X X - - X X - X X² X² - - X X X X X X X X
D-NET Software	X X X X X X X X
Partheno s Aggregat or	
Repositór io Mexican a	X X - X X X X - X X X X X X X X
UNLV's Linked Data Project	

[1] Dados observados visualização dos workflows

[2] Itens não explícitos no workflow, mas identificados na documentação

Fonte: elaborado pelos autores

O Quadro 4 apresenta quais etapas foram consideradas para constar como “presente” no Quadro 03. Este Quadro também nos ajuda a visualizar quais são os nomes mais usados para nomear cada etapa, para colaborar com pesquisas futuras.

A documentação na qual os *workflows* estão inseridos apresentam alguns dados que não constam do fluxograma. Além disso, percebe-se pouca preocupação com a qualidade dos dados inseridos, ou seja, os dados coletados na etapa de extração, havendo pouca menção a etapas tradicionais de projetos de análise de dados, envolvendo limpeza, tratamento e normalização de dados.

Além das etapas, as publicações apresentam algumas ferramentas de softwares utilização para execução do *workflow*. De forma geral, os *workflows* são genéricos demais e não apresentam o fluxo real de processos necessários, contrariando assim, um dos princípios básicos de um *workflow*, que é a possibilidade de ser replicado. Além disso, percebe-se a necessidade de um conhecimento técnico avançado e extremamente especializado para compreensão de todas as etapas.

CONCLUSÕES

A análise dos diferentes *workflows* de agregação de dados permitirá aos pesquisadores compreender quais etapas estão sendo executadas, quais estão sendo postas em segundo plano e quais precisam ser incluídas. Esse conhecimento estruturado pode auxiliar na compreensão de etapas que devem ser resolvidas do ponto de vista da criação de um serviço de agregação de dados culturais. Além disso, é importante compreender que não consenso nem na quantidade de etapas, no seu nome e nem nas tecnologias utilizadas, demonstrando o quanto esse tema parece ainda em estágio inicial de pesquisa ou mesmo revelando que as soluções são altamente customizadas, exigindo soluções locais para problemas específicos.

É importante destacar que a grande maioria menciona soluções para processamento de dados massivos e construídas para lidar com projetos de *big data*. São mencionados o Apache Lucene, Apache Solr, ElasticSearch, Hadoop, MapReduce e MongoDB. Não fica claro a forma
como essas tecnologias são utilizadas, a maneira como são integradas e a documentação se mostra bastante deficitária de detalhes e discussões alongadas sobre o tema. No entanto, é importante perceber que já há na discussão sobre a agregação de dados culturais a presença dessas tecnologias de forma determinante. É importante reconhecer que esse é um tema ainda novo para a Ciência da Informação e que esforços de pesquisa e desenvolvimento devem ser feitos para que se compreendam as possíveis aplicações dessas tecnologias, dado que as mesmas não apenas novas técnicas, mas representam novas formas de se pensar nos dados e em um ecossistema completo de serviços analíticos.

Também é possível notar a baixa densidade dos trabalhos apresentados, sendo as discussões feitas de forma bastante genérica. O trabalho mais detalhado identificado diz respeito a iniciativa menos automatizada, relacionada ao trabalho da Universidade de Nevada (SOUTHWICK, 2015), que fez intensivo da ferramenta OpenRefine como estratégia de coleta, tratamento e organização dos dados. Apesar da importância da pesquisa, a mesma demonstra que todo o fluxo de trabalho deveria ser feito novamente para cada novo registro publicado, inviabilizando sua adoção para solução para serviços que exigem atualização automática dos índices de busca e recuperação da informação. Fica evidente, a partir dos resultados desta pesquisa, o quanto ainda é necessário se compreender como esses fluxos de agregação devem funcionar e como podem ser utilizados para a criação de serviços informacionais de agregação de dados. Vale ressaltar que serviços dessa ordem representam grandes contribuições da área da Ciência da Informação a sociedade brasileira, assim como tem sido com serviços como a Biblioteca Digital de Teses e Dissertações (BDTD) criada pelo IBICT e a própria BRAPCI, no caso específico da comunidade da Ciência da Informação.

Como trabalho futuro, pretende-se realizar pesquisas direcionadas a cada etapa do *workflow*, buscando ampliar a compreensão de como as etapas são realizadas, seus detalhes operacionais, técnicos e informacionais.

Quadro 4 – Nomenclatura original das etapas na literatura revisada

Projeto/ Etapas	Extrair	Estruturar	Transformar Reconciliar	Armazenar Publicar	Expor	Novas aplicações

AAC	*Prepare and export Define Model*	*AAC*	*- Reconcili e Entities RDF*	*Triple Store - SPARQL*	*Browse Demo APP e Toy Box*	*Applets SPARQL/ Transfor m Interface - -*	*Library API Future Applicati ons*	
Trove [1]	*NLA Harvest*	*- - - MySQL e Unit Database*	*Seachabl e Trove Manager e Common - - - API API - -*	*User*				
DigitalN Z		*Harvesti ng Delivery Schema*	*Mapping - Value Mapping X [2] X [2] - -*					
European a D-NET Software		*Manipula tion Mediatio n*	*Manipula tion Manipula tion*	*Storage Provision*	*Provision Provision*			
Partheno s	Aggregat or	*Collectio n Transfor*	*mation Metadata Record*	*Inspector Metadata Cleaner*	*Index Service X [2] OAI*	*publisher service PMH*	*PMH publisher service OAI*	
Repositór io	Mexican a	*Extractor es e*	*Cosecha dores Mapeado r*	*- - Almacien amento*	*Buscador e Exhibicio nes*	*API de Búsqued a*	-	
UNLV's Linked Data Project	os somente a partir da visualizaç ão do workflow [2] Item não explicito no	workflow *Clean and export Impleme nt mapping Prepare matadata for transfor mation Reconcili e with*	*controlle d vocabula ries e Create new URIs for local controlle d vocabula ries Publish LOD*	*Publish LOD Apply visualizat ion tools Publish LOD*				

[1] Dados observad

Fonte: elaborado pelos autores

REFERÊNCIAS

BARDI, Alessia; MANGHI, Paolo; ZOPPI, Franco. Aggregative data infrastructures for the cultural heritage. In: Dodero J.M., Palomo-Duarte M., Karampiperis P. (eds) *Metadata and Semantics Research*. MTSR 2012. Communications in Computer and Information Science, vol 343. Springer, Berlin, Heidelberg, 2012. p. 239-251. DOI: <https://doi.org/10.1007/978-3-642-35233-1_2>

BRIGHAM, Tara J.; FARRELL, Ann M.; OSTERHAUS TRZASKO, Leah C.; ATTWOOD, Carol Ann; WENTZ, Mark W.; ARP, Kelly A. Web-Scale Discovery Service: Is It Right for Your Library? Mayo Clinic Libraries Experience. *Journal of Hospital Librarianship*, 16(1), 25–39, 2016. DOI: <https://doi.org/10.1080/15323269.2016.1118280>

DIGITAL NEW ZEALAND. This is Digital New Zealand. *YouTube.* 20 dez 2018. Disponível em: <https://www.youtube.com/watch?v=UWbIDwsaA4o>. Acesso em: 18 abr. 2020.

DIGITAL NEW ZEALAND. Our History. 2019. Disponível em: <https://digitalnz.org/about/our-history>. Acesso em: 18 abr. 2020.

EUROPEANA PRO. Brief History. 2020. Disponível: <https://pro.europeana.eu/about us/mission#brief-history>. Acesso em: 27 abr. 2020.

FINK, Eleanor E. Overview and Recommendations for Good Practices. *American Art Collaborative. Linked Open Data Initiative,* 2018. Disponível em: <http://americanartcollaborative.org/wp content/uploads/2018/03/AAC_LOD_Overview_Recommendations.pdf>. Acesso em: 15 abr. 2020.

FROSINI, Luca et al. An Aggregation Framework for Digital Humanities Infrastructures: the parthenos experience. *Scientific Research and Information Technology*. v. 8, n. 1, p. 33-50, 2018. DOI: <http://dx.doi.org/10.2423/i22394303v8n1p33>.

KOLLIA, Ilianna, TZOUVARAS, Vassilis, DROSOPOULOS, Nasos and STAMOU, Giorgos. A Systemic Approach for Effective SemanticAccess to Cultural Content. *Semantic Web*, v. 3, n. 1, p. 65-83, 2012. DOI: <http://dx.doi.org/10.3233/SW-2012-0051>.

MANGHI, Paolo et al. The D-NET software toolkit: A framework for the realization, maintenance,and operation of aggregative infrastructures. *Program: electronic library and information systems*, v. 48, n. 4, p. 322–354, 2014. DOI: <https://doi.org/10.1108/PROG-08-2013-0045>.

NATIONAL LIBRARY OF AUSTRALIA. Trove Help Center. Trove System Architecture Diagram. 2010. Disponível em: <https://www.nla.gov.au/trove/marketing/Trove%20architecture%20diagram.pdf>. Acesso em: 18 abr. 2020.

NAVARRETE, Trilce. Europeana as online cultural information service: study report. [S.l.]: Europeana, 2016. Disponível em: <https://pro.europeana.eu/files/Europeana_Professional/Publications/europeana-benchmark report-sep-2016.pdf>. Acesso em: 27 abr. 2020.

PAVÃO, Caterina Marta Groposo; CAREGNATO, Sônia Elisa. Serviços de descoberta em rede: a experiência do modelo Google para os usuários de bibliotecas universitárias. Em Questão, [s.l.], v. 21, no. 3, p. 130, 2015. Disponível em: <https://seer.ufrgs.br/EmQuestao/article/view/58410/36046>. Acesso em: 27 abr. 2020.

SCHOLZ, Henning. Europeana Publishing Guide v1.8. p. 1–32, 2019. Disponível em: <https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Publishing%20Guide%20v1.8.pdf>. Acesso: em 18 abr. 2020.

SECRETARÍA DE CULTURA. Mexicana Repositorio del Patrimonio Cultural de México. *Dirección General de Tecnologías de la Información y Comunicaciones, Agenda Digital De Cultura*. Colonia Cuauhtémoc. Ciudad de México. 2018. Disponível em: <https://mexicana.cultura.gob.mx/work/models/repositorio/Resource/126/2/images/documentacion.pdf>. Acesso em: 18 abr. 2020.

SOUTHWICK, Silvia B. A guide for transforming digital collections metadata into linked

data using open source technologies. *Journal of Library Metadata*, v. 15, n. 1, p. 1-35, 2015. DOI: <http://dx.doi.org/10.1080/19386389.2015.1007009>.

TROVE HELP CENTRE. About Trove. 2020. Disponível em: <https://help.nla.gov.au/trove/using-trove/getting-to-know-us>. Acesso em: 18 abr. 2020.

SUPPLEJACK. Architecture. Documentation (Version 0.1). 2020. Disponível em: <http://digitalnz.github.io/supplejack/architecture.html>. Acesso em: 18 abr. 2020.

WORKFLOW MODELS FOR AGGREGATING CULTURAL HERITAGE DATA ON THE WEB: A SYSTEMATIC LITERATURE REVIEW

1. INTRODUCTION

Institutions responsible for the custody of cultural heritage collections in Brazil and especially abroad have invested in the digitalization and dissemination of their digital collections on the Web. This movement has resulted in a large amount of available content, which brings challenges to the discovery of digital resources to deal with cultural heritage. This trend has made it essential to provide users with an efficient way to find digital objects, by offering integrated search interfaces. Internationally there are different initiatives of this kind, being Europeana the best known. In Brazil, there is still no widely applied solution, as national initiatives are generally restricted to small projects in local institutions.

Although there are a variety of solutions, it is consensual that getting together cultural data brings additional difficulties, due to the diversity of types of digital objects. We can state that there is no simple solution, but conversely, there are very challenging software engineering and data modeling problems (AVAZPOUR; GRUNDY; ZHU, 2019).

To understand the data aggregation models of cultural collections proposed worldwide aiming at their customization and development for the Brazilian reality, we proposed a Systematic Literature Review (SLR), with the objective of answering five research questions: What are the projects? What are the steps foreseen for data aggregation? What are technologies used? Are the steps performed manually, automatically or semi-automatically?

and Which perform semantic search?

In addition to this introduction, the article is divided as follows: in section 2, Method, we present the protocol for the realization of the SLR; in section 3, Systematic Literature Review (SLR), we present the Planning, Execution and Results; in section 4, Discussion, we discuss the results and, finally, in section 5 we bring the conclusions.

2. METHOD

This study unfolds as post-positivist, mixed approach research, using the Systematic Literature Review (SLR) method, which is a form of secondary study, i.e., it reviews primary studies related to research issues, using a well-defined methodology, in order to identify, analyze and interpret the available evidence in an impartial and replicable manner (KITCHENHAM; CHARTERS, 2007).

This SLR uses the guidelines proposed by Kitchenham (2004), which defines and describes three steps: Planning, Execution and Results.

3. SYSTEMATIC LITERATURE REVIEW (SLR)

3.1. Planning

Planning contemplates the identification of need for review, the elaboration of the research questions, the choice of the databases, as well as the creation of the search string and the elaboration of the inclusion and exclusion criteria.

Identification of need for review

The review aims to provide technical and conceptual subsidies for researchers involved in the Brazilian effort to build a tool for the aggregation of metadata of digital objects in the area of culture. We expect this research to better understand how such a challenge is faced by other initiatives and facilitate the production of a local solution.

Research questions

The SLR is aimed at answering four questions:

Q01. What are the projects?

Q02. What are the planned steps?

- Present a brief history of the project.
- Present the aggregation workflow.
- Present the detailing of the steps foreseen in the workflow.

Q03. Which technologies are used?

Q04. Are the steps performed manually, automatically or

semi-automatically? Q05. Which perform semantic search?

Databases

We selected three important databases: the first, Networked Digital Library of Theses and Dissertations (NDLTD) <http://www.ndltd.org/>, as an international organization for the dissemination of theses and dissertations, and the others, Scopus <https://www.scopus.com/> and Web of Science (WoS) <https://www.webofknowledge.com/>, for its recognized quantity and variety of publications and areas of knowledge.

Search string

The definition of the search string brought an additional challenge, as the subject is relatively new, besides being very polysemic, due to its high degree of interdisciplinarity. Therefore, there is no consensus on which term(s) best represent it.

Thus, at first we tried the theme "discovery services in scale on the web" (SIQUEIRA; MARTINS, 2019); in the sequence, understanding that the term did not meet the search for data aggregators, we started to identify, in the gray literature, large projects, national and international, for data aggregation (SIQUEIRA; MARTINS, 2020).

Both studies proved essential for us to understand and expand the search terms, resulting in the search string that follows, duly formatted according to the specificities of the search engine of each database. In NDLTD, we performed the search in all available fields,

while in Scopus and WoS we performed search in the title, abstract and keywords fields.

- NDLTD: ((aggregat* OR integration) AND (infrastructure OR workflow OR flux OR flow) AND (metadata) NOT ("real time" OR real-time OR realtime OR "big data")). •
  Scopus: TITLE-ABS-KEY ((aggregat* OR integration) AND (infrastructure OR workflow OR flux OR flow) AND (metadata) AND NOT ("real time" OR real-time OR realtime OR "big data")).
- WoS: ts=((aggregat* OR integration) AND (infrastructure OR workflow OR flux OR flow) AND (metadata) NO ("real time" OR real-time OR realtime OR "big data")).

Selection criteria

The selection criteria were divided into inclusion criteria (IC) and exclusion criteria (EC), as follows.

Inclusion criteria (IC)

IC01. Portuguese, English or Spanish languages.

IC02. Period from 2010 to 2020.

IC03. Present solutions for the aggregation of data focused on cultural

heritage. IC04. Present the aggregation workflow.

Exclusion criteria (EC)

EC01. Duplicate publications.

EC02. Inaccessible publications.

EC03. Out of scope.

EC03.1. Related to real time data processing.
EC03.2. Related to ubiquitous computing.

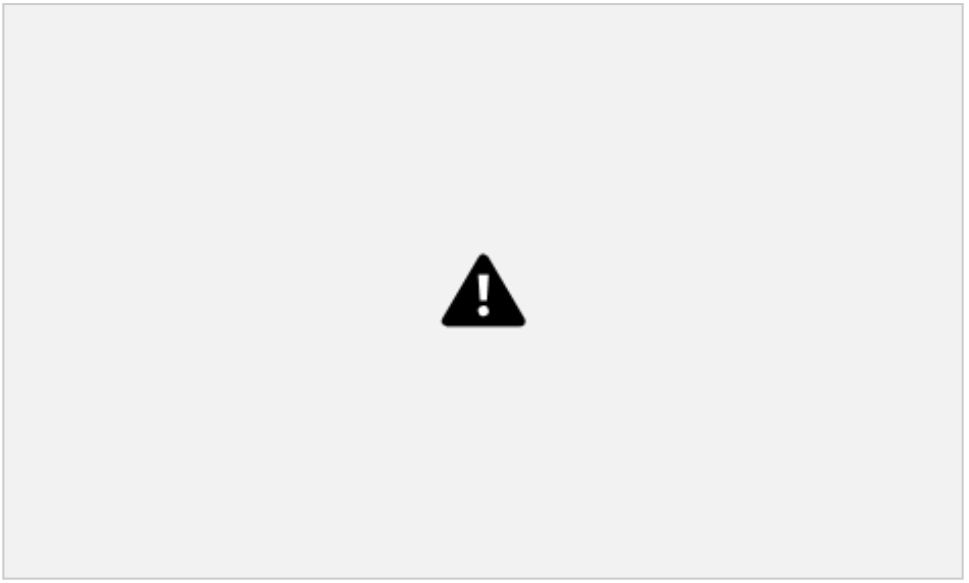
In order to explain some of the criteria we chose, IC04 aimed at publications presenting the workflow, which are conceptually directed to steps that can be replicated. Thus,

they would be more likely to offer answers to the research questions. In EC02, we excluded the publications unavailable in the databases to which we have access, also considering those paid by University of Brasilia. In EC03 we excluded publications with no relation to the proposed theme. In addition to these publications, we also considered out of scope research presenting solutions for real-time data processing (EC03.1), since for cultural heritage, the inclusion of data from catalogs does not require this type of approach. And those related to ubiquitous computing (EC03.2), which, although they deal with data aggregation, have specific purposes, apart from our interest.

3.2 Execution

The execution of the search string was performed on May 15th, 2020, returning a total of 1476 publications, 72 at NDLTD, 939 at Scopus and 465 at WoS. We applied the selection criteria as follows: IC01 and IC02 inclusion criteria were applied with the filters of the search tools themselves. We carried out the inclusion criteria IC03, IC04, and the exclusion criteria EC03, EC03.1 and EC03.2 in two steps, being the first one reading the abstracts and, in case of doubt, reading the complete document. Figure 01 shows the numbers relating to the execution of the string.

Figure 01. Numbers related to the execution of the search string



Source: authors (2020)

The 16 selected documents subsidize the answers to the research questions, presented in the Results.

3.3 Results

We answered the five questions we proposed: “Q01. What are the projects?”; “Q02. What are the planned steps?”; “Q03. Which technologies are used?”; “Q04. Are the steps performed manually, automatically or semi-automatically?” e “Q05. Which ones perform semantic search?”

Q01. What are the projects?

The documents resulted in the discovery of 12 projects, listed in Chart 01.

Chart 01. List of projects selected by the SLR, sorted in ascending order, by year of publication

1. Connecting Archaeology and Architecture in Europeana Project - CARARE Project (PAPATHEODOROU et al., 2011; GAVRILIS, DALLAS e ANGELIS, 2013);
2. EUscreen Project (OOMEN e TZOUVARAS, 2012; OOMEN, TZOUVARAS e HYYPPAA, 2013);
3. Heritage of the People's Europe Project - HOPE Project (BARDI, MANGHI e ZOPPI, 2012, 2014; MANGHI et al., 2014);
4. Natural Europe Project (MAKRIS et al., 2013);
5. European Film Gateway Project – EFG Project (ARTINI et al., 2013);
6. Open Cultural Digital Content Infrastructure (STATHOPOULOU et al., 2014);
7. Meta-Share (PIPERIDIS et al., 2014);
8. Europeana network of Ancient Greek and Latin Epigraphy Project – EAGLE Project (MANNOCCI et al., 2014);
9. Digital Public Library of America – DPLA (MATIENZO e RUDERSDORF, 2014); 10. European Collected Library of Artistic Performance Content Aggregator – ECLAP (BELLINI et al., 2015);
11. ARIADNE Project (RONZINO, FELICETTI e DI GIORGIO, 2016);
12. Data Aggregation Lab Software Tool – DAL (FREIRE, 2019).

Source: authors (2020)

We answered question 02, "What are the planned steps?" with a brief history of each

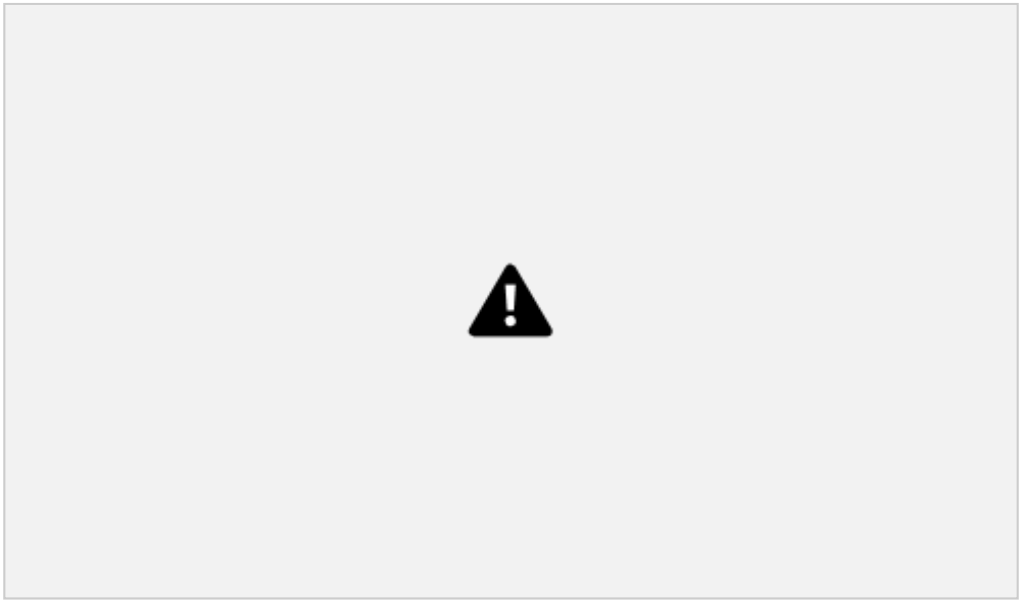
project listed in Chart 1, the aggregation workflow and the details of the planned steps, whenever possible.

Question 02. What are the planned steps?

CARARE Project

The CARARE best practices network was created to increase the quantity and quality of the digital content of Europeana's archaeological and architectural heritage (PAPATHEODOROU et al., 2011). The CARARE architecture, shown in Figure 02, presents a central repository called the Monument Repository (MoRe), where all metadata is stored and enriched before being mapped to Europeana Data Model (EDM) and provided to Europeana. MoRe provides services such as: geographic information normalization, metadata integrity monitoring, semantic enrichment, etc. (PAPATHEODOROU et al., 2011).

Figure 02. Architecture of CARARE system



Source: Papatheodorou et al. (2011, p. 420)

The authors do not provide details about each step, showing only the workflow diagram for analysis. Gavrilis, Dallas & Angelis (2013) provide more information about the project, especially about the MoRe repository, as depicted in Figure 03.

Figure 03. Architecture of the MoRe repository

Source: Gavrilis, Dallas e Angelis (2013, p. 3)

Submission Information Packages (SIP) are sources of shipment information, which in the Ingestion service layer, are manipulated following precise specifications regarding their content and structure. Content owners initially map internal metadata in records compatible with the CARARE scheme, using a transformation tool, such as MINT or Repox; in the Indexing service layer (not in the workflow). MoRe, through its own indexing mechanism, defines which parts of the metadata will be indexed and the structure of the SQL database to which they will be indexed. The core architecture, Fedora-Commons and Index Database, consists of a layer of services that receive information packages, pre-process them and store the resulting data streams in a Fedora installation. The indexes of these data flows are stored in a MySQL database and kept in synchronization with Fedora; in the Workflow engine layer, simple cleaning and filling services are performed. Complex services such as adding relationships, removing duplicate records, etc. are executed later; the Quality monitoring layer ensures that data owners can effectively control the status of ingested metadata objects, with information quality measured by collection or even by submission package; the Curation layer allows for monitoring and information objects and performs evaluation tasks, including enrichment, aimed at improving content quality. The actions performed in this step include: element and attribute cleaning; de-duplication; attribute filling; relationship addition and spatial transformation; the Preservation layer is responsible for maintaining the metadata of the records provided to the repository, enabling its revision, version and validation.

EUscreen Project

The EUscreen project offers access to thousands of audiovisual heritage items, gathering clips of social, cultural, political and economic events. It also allows to explore television programs focused on everyday experience and acts as a domain aggregator for Europeana (EUSCREEN, 2020). The workflow consists of four phases, as shown in Figure 04.

Figure 04. Ingestion workflow of Content of the EUscreen Project

Source: Oomen; Tzouvaras (2012, p. 2)

The article does not detail the steps. In more recent work, Oomen; Tzouvaras; Hyyppaä (2013) present a similar workflow, shown in Figure 05.

Figure 05. Ingestion workflow of content of EUscreen Project

Source: Oomen; Tzouvaras; Hyyppaä (2013, p. 481)

This work brings the ingestion of metadata, its transformation into a common reference scheme, enrichment and finally publication as linked data. The authors do not detail the steps, but include the tool used for ingestion, transformation and enrichment, the Mint Platform. The publication brings information about the front end, which consists of "serverless" APIs, in which a proxy system handles communication with the back-end services. Finally, internal and external linking to EUscreen content was performed and the resulting repository became accessible through a SPARQL query endpoint.

HOPE Project

The D-NET Software Toolkit is presented as an ideal candidate for the creation of sustainable, extensible, scalable and dynamic data aggregation infrastructures for cultural heritage. Figure 06 presents HOPE's aggregation workflow using D-NET (BARD, MANGHI and ZOPPI, 2012).

Figure 06. Aggregation workflow of HOPE Project

Source: Bard, Manghi e Zoppi (2012, p. 248)

The use of cross-walk (or mappings) solves the structural and semantic heterogeneity of metadata records. As Figure 06 depicts, there are three phases, the intra-profile, performed by mapping metadata records from all data sources of the same profile to metadata records conforming to a given standard data model; the cross-profile, performed by providing stable mapping of these metadata records to HOPE data model records and the Europeana cross-walk, to transform into EDM and store.

The services can collect records via OAI-PMH, FTP, SFTP, HTTP and HTTPS (Harvester Service). An instance of the Tranformator Service is created for each data source to transform input records based on mappings defined with the help of the content provider. After the records are transformed into the Profile Metadata Format (PMF), they are processed by another instance of the Tranformator Service configured with the cross-over from the PMF to the Common Metadata Format (CMF) of HOPE. At this point the flow goes to the Metadata Cleaner Service, which applies the semantic transformation of values (provider vocabulary terms in common vocabulary terms). The clean records are delivered in the index,

for portal use and transformed in the EDM to be collected by OAI-PMH by Europeana. The conversion of metadata records is completed by services solving problems of granularity (Metadata (un)packaging Service) XML records, i.e. performing conversions from one to many and from many to one.

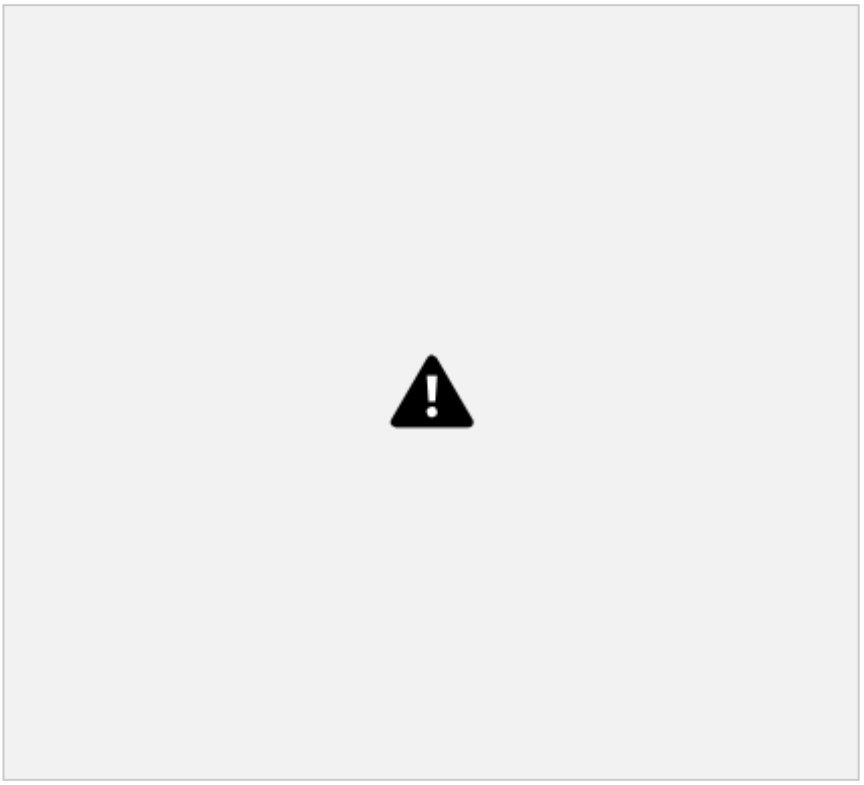
The information is stored in Metadata Store Service and is accessible via Full-Text Index Service. The Curator Services, which provides tools to find mapping errors and semantic inconsistencies and data enrichment, i.e. the Content Checker Service and the Record Tagging Service have been implemented. The last two services are not part of the workflow.

Natural Europe Project

The Natural Europe project offers a coordinated solution that aims to improve the availability and relevance of environmental cultural content, in a multilingual and multicultural context. The content is collected from six Natural History Museums (NHMs) from Europe (MAKRIS et al., 2013).

Figure 07 presents and structure of the Federation of Digital Cultural Libraries of Natural Europe, together with its tools and services, which consist of: The Natural Europe Cultural Environment (NECE), i.e. the infrastructure and toolset deployed in each NHM, allowing its curators to publish, describe, manage and disseminate the Cultural Heritage Objects (CHOs) and the Natural Europe Cultural Heritage Infrastructure (NECHI), interconnecting the NHM digital libraries and further exposing their metadata records to Europeana.eu.

Figure 07. The Natural Europe Cultural Digital Libraries Federation

Architecture 

Fonte: Makris et al. (2013, p. 3)

NECE consists of the Multimedia Authoring Tool (MMAT), a multilingual web-based management system that facilitates the enrichment of CHO metadata: CHO Management, responsible for creating, recovering, updating and deleting CHOs, records/collections and users; Multimedia Manipulation, which manages all the functionality regarding multimedia files in the system, including thumbnail generation and extraction of metadata from media files, which are used for the creation and enrichment of CHO records; Concurrency Management, which provides the basic functionality for simultaneous access to data in the repository. It ensures that there are no consistency issues when multiple users try to access the same resource; Vocabulary Management, which allows access to taxonomic terms, vocabulary and authority

files. This information resides on Vocabulary Server, providing indexing and search capabilities; Persistency Management, which manages the upload/recovery of information packages to and from the CHO Repository; Graphical User Interface, which is responsible for user interaction, information presentation and communication with the server; and CHO Repository, which is responsible for ingestion, maintenance and dissemination of content and metadata and adopts the OAIS Reference Model.

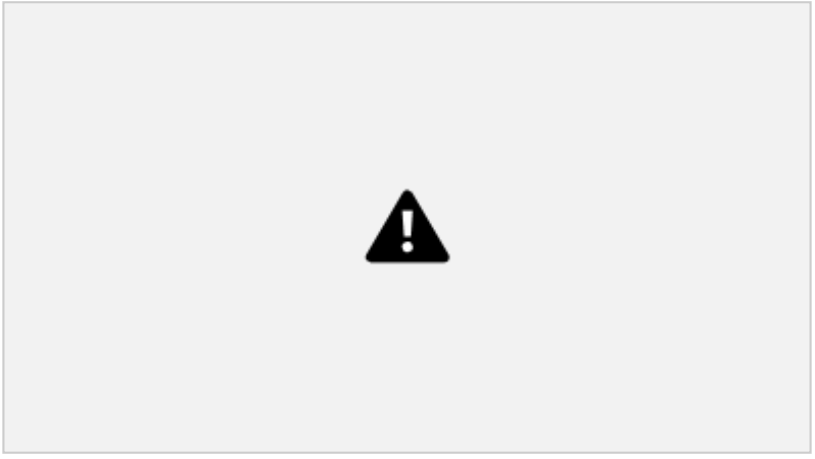
Natural Europe CHO Application Profile is a super-set of the Europeana Semantic Elements (ESSE) metadata format and was developed through an interactive process involving NHM domain experts and technical project partners, guided by the needs and requirements of stakeholders and the application domain.

NECHI was based on ARIADNE technologies and services. Natural Europe Harvester uses OAI-PMH to collect the metadata of the OAI-PMH destinations and publishes it in a central repository through the publication service that integrates the services: Publish, which publishes the collected metadata in a central repository; Transformation, which converts metadata from the Natural Europe CHO Application Profile format to the Europeana specification format; Identification, which provides persistent digital identifiers for resources in the ARIADNE infrastructure; Metadata validation, which provides syntactic and semantic validation of metadata instances against predefined application profiles. Finally, Search Widgets supports a simple, facetted or connected search.

EFG Project

Film archives containing collections of film-related digital material have been created in many European countries. EFG has designed a common data model for film information, in which file data models can be optimally mapped. It realizes a data infrastructure based on the D-NET software toolkit, which has been expanded with advanced tools for data curatorship. Figure 08 present the workflow.

Figure 08. Phases of the EFG data ingestion workflow



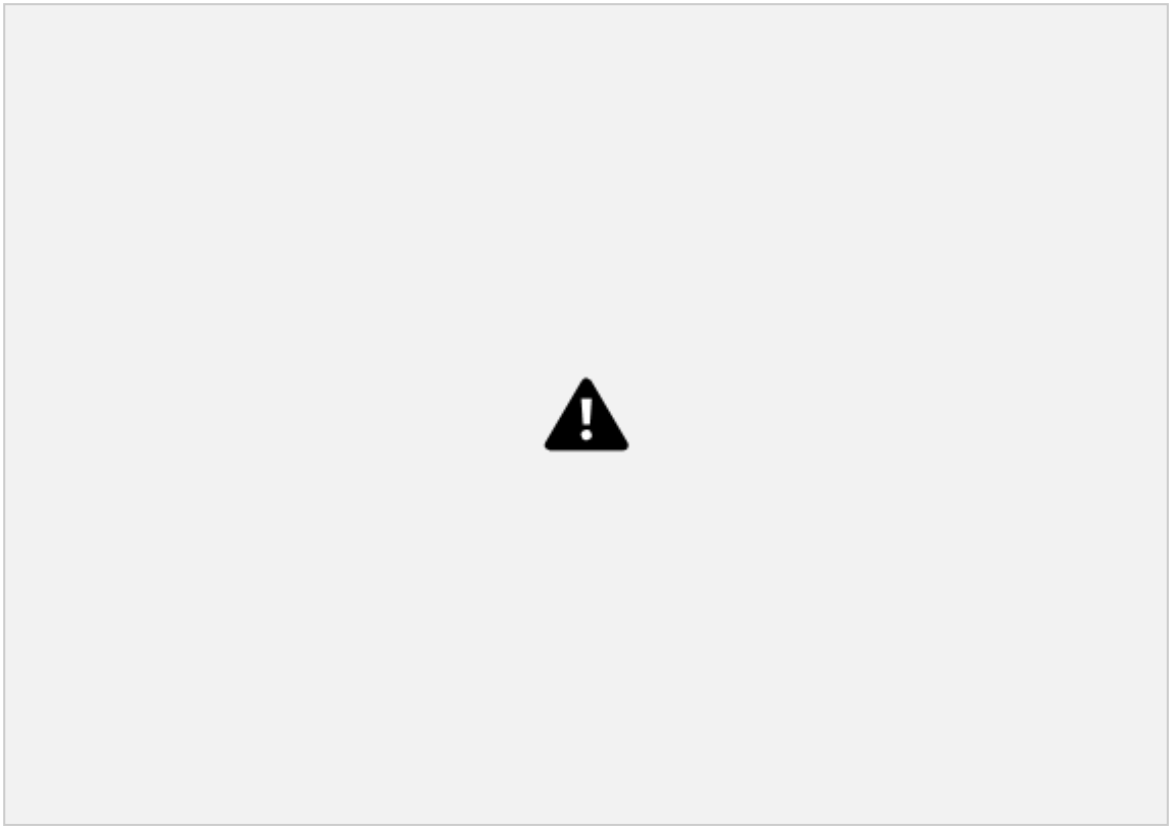
Source: Artini et al., (2013, p. 36)

In Phase 1, Metadata Mapping Definition, file specialists analyze the metadata they provide to determine how that information can structurally and semantically map the metadata schema. Structural and semantic mapping rules are delivered to infrastructure administrators, who encode them in the form of D-NET scripts. In Phase 2, Metadata Transformation and Cleaning, the metadata is collected via OAI-PMH or FTP protocols to be processed by the mapping scripts produced in Phase 1 and generate the corresponding EFG records. The resulting records are not immediately available for access, but are stored in a "pre-production" information space. In Phase 3, Metadata Quality Control and Enrichment, the records can be validated and inspected to identify mapping errors, typos and duplicates. Specifically, the Content Checker Tool can be used to verify that the structural mapping has been performed correctly. The Vocabulary Checker Tool notifies data providers about EFG records that do not yet conform to common vocabularies and the Authority File Manager (AFM), a tool that curators can use to merge duplicate records and disambiguate information. The Metadata Editor Tool allows trustees to edit EFG records, while AFM can trigger record merge actions and effectively remove duplicates. In Phase 4, Metadata Publishing, records approved in Phase 3 are visible on the EFG portal and can also be exported to external vendors, such as Europeana.

Open Cultural Digital Content Infrastructure

Stathopoulou et al. (2014) present the development of an infrastructure to aggregate centrally produced metadata records and digital files and automatically validate their compliance with interoperability and quality specifications. Figure 09 presents the architecture of the infrastructure and service components of Open Cultural Digital Content.

Figure 09. Architecture of Open Cultural Digital Content



Source: Stathopoulou et al. (2014, p. 286)

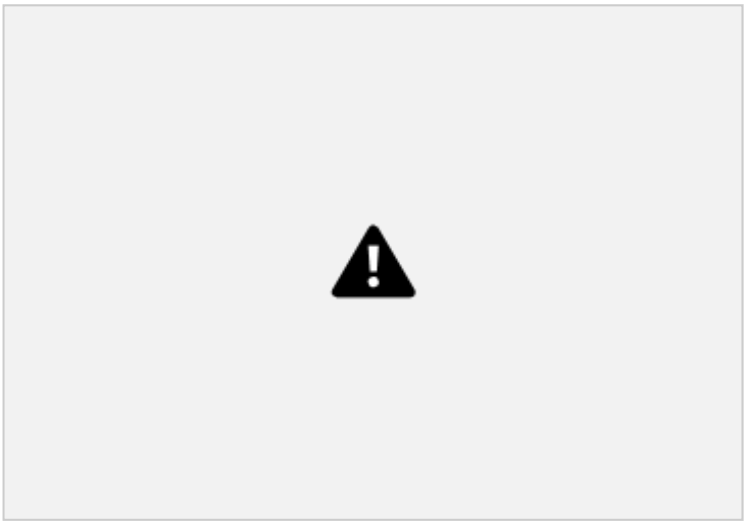
Although it presents a detailed architecture in Figure 09, the article does not explain all the items, emphasizing the validation layers: Validation Back End and Front-End Validator. In general, they inform that the system is built with an autonomous data collection layer, which collects and stores metadata and digital files. The REST API has been implemented to allow external software components, such as the Validator or Aggregator, to trigger and manage the collections. In relation to the Validation Back End layer, it allows to specify complex validation rules at any level (repository, metadata, digital files), perform validations and record detailed results, while the Front-End Validator provides a user interface, which will be able to perform validation procedures, aggregate reports and connect with the repository providers and administrative procedures for validations.

Meta-Share

Piperidis et al. (2014) present META-SHARE, an open language resource infrastructure, formed by a network of repositories that store language resources (data, tools and processing services) documented with high quality metadata, aggregated in central

inventories, allowing uniform search and access. Figure 10 presents its architecture.

Figure 10. Architecture of META-SHARE



Source: Piperidis et al. (2014, p. 1533)

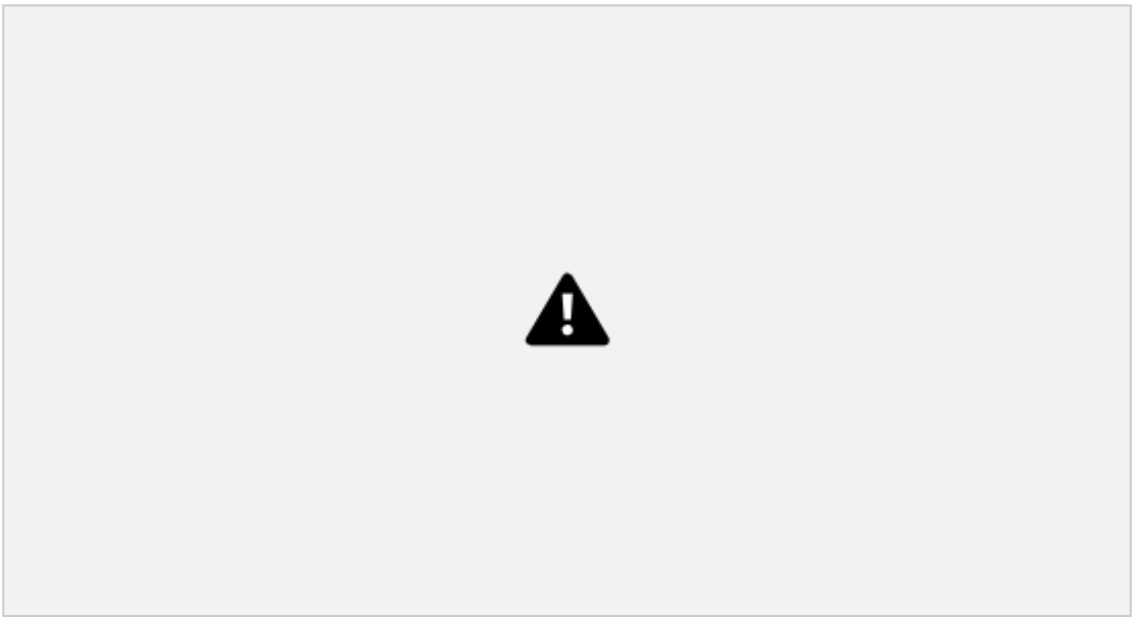
The authors do not detail all the steps, but generally explain that the metadata records are collected and stored on the central META-SHARE servers, which share metadata and create, host and maintain a central inventory, including metadata based descriptions of all available resources on the distributed network. Resource providers can create, store and edit resource descriptions using an editor implemented by META-SHARE, as well as upload actual resources directly or provide a link to an existing storage. They can also obtain statistics on the number of views and downloads of their resources, as well as the origin of viewers. Data centers or other organizations with existing resource catalogs can obtain support for mapping their metadata into the infrastructure model schema. Resource consumers can register, create a user profile, and log into the network so they can browse and search the central inventory, using multifaceted search capabilities, and access and obtain information on specific resources.

Project EAGLE

The EAGLE project aims to provide Europeana with a comprehensive collection of historical sources from the Mediterranean region and to provide a user-friendly portal to

access the same collection, built with material from about 15 different epigraphic files. The EAGLE data aggregation infrastructure is powered by the D-NET software toolkit. Figure 11 presents the EAGLE workflow.

Figure 11. D-NET adjusted to EAGLE



Source: Mannocci et al. (2014, p. 293)

Similarity to the EFG Project, the data congestion process, based on D-NET, consists of four phases. The Metadata mapping definition, which in cooperation with domain experts provide records, structural and semantic rules to map the records in the common EAGLE metadata schema are encoded in the form of D-NET scripts; Metadata transformation and cleaning, in which the metadata records are collected and processed to generate the "EAGLE objects", thus creating the pre-production space; Metadata quality control, in which the records are inspected and validated to identify mapping errors and others, such as typing; Metadata provisioning, in which the records that pass Phase 3 are moved to the Production Information Space, where they are indexed and become available for ingestion to Europeana or for consultation and navigation in the EAGLE portal.