# Recuperação de informação: descoberta e análise de *workflows* para agregação de dados do patrimônio cultural

**Joyce Siqueira**
Doutoranda em Ciência da Informação pela Universidade de Brasília (UnB) - Brasília, DF – Brasil. Mestre em Ciência da Computação pela Universidade Federal de Goiás (UFG) – GO - Brasil. Professora da Universidade Católica de Brasília (UCB) - Brasília, DF – Brasil.
http://lattes.cnpq.br/7006325340635761
E-mail: joycitta@gmail.com

**Dalton Lopes Martins**
Pós-Doutorado pela Universidade de São Paulo (USP) – SP - Brasil. Doutor em Ciência da Informação pela Universidade de São Paulo (USP) - São Paulo, SP - Brasil. Professor da Universidade de Brasília (UnB) - Brasília, DF - Brasil.
http://lattes.cnpq.br/3774617443225038
E-mail: dmartins@gmail.com

Submetido em: 28/10/2020. Aprovado em: 25/11/2020. Publicado em: 28/07/2021.

**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 objetos digitais e facilite a recuperação de dados para os usuários leigos. Contudo, integrar dados culturais não é uma tarefa trivial, pois estes são diversos e singulares, necessitando de uma variedade de etapas entre a coleta e a apresentação. Com objetivo de identificar estas etapas, esta pesquisa visa localizar *workflows* de agregação de dados e discuti-los. Para tal, realizou-se pesquisa descritiva e bibliográfica, de natureza qualitativa, em bases de dados acadêmicas e na literatura cinzenta. Como resultado, apresentam-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*.

## *Information retrieval: discovery and analysis of workflows for aggregating cultural heritage data*

***ABSTRACT***

*In recent years, different cultural institutions have been making efforts to spread culture through the construction of a unique search interface, which integrates digital objects and facilitates data retrieval for lay users. However, integrating cultural data is not a trivial task, as these are diverse and unique, requiring a variety of steps between collection and presentation. In order to identify these steps, this research aims to locate data aggregation workflows and discuss them. To this end, descriptive and bibliographic research, of a qualitative nature, was carried out in academic databases and in gray literature. As a result, eight projects are presented: American Art Collaborative, DigitalNZ, D-NET Software, Europeana, Mexicana, Parthenos Aggregator, TROVE and UNLV's Linked Data Project. 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. Publish, 7. Expose and 8. Enable new applications. In addition, the need for more detailed stages is also visible, so that it is possible to replicate the workflow, and enjoy its benefits in other institutions..*

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

## *Recuperación de información: descubrimiento y análisis de flujos de trabajo para agregar datos del patrimonio cultural*

***RESUMEN***

*En los últimos años, diferentes instituciones culturales se han esforzado por difundir la cultura mediante la construcción de una interfaz de búsqueda única, que integra objetos digitales y facilita la recuperación de datos para usuarios legos. Sin embargo, la integración de datos culturales no es una tarea trivial, ya que son diversos y únicos, y requieren una variedad de pasos entre la recopilación y la presentación. Para identificar estos pasos, esta investigación tiene como objetivo localizar los flujos de trabajo de agregación de datos y discutirlos. Para ello, se realizó una investigación descriptiva y bibliográfica, de carácter cualitativo, en bases de datos académicas y en literatura gris. Como resultado, se presentan ocho proyectos: American Art Collaborative, DigitalNZ, D-NET Software, Europeana, Mexicana, Parthenos Aggregator, TROVE y Linked Data Project de UNLV. El análisis del conjunto de flujos de trabajo resultó en ocho pasos diferentes a realizar: 1. Extraer, 2. Estructurar, 3. Transformar, 4. Reconciliar, 5. Almacenar, 6. Publicar, 7. Exponer y 8. Habilitar nuevas aplicaciones. Además, también es visible la necesidad de etapas más detalladas, para 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, reinventando-se e inovando suas formas de interagir com público, com destaque, a disponibilização de objetos digitais e seus metadados em sites e/ou repositórios institucionais, como um meio para exercer sua prática comunicacional e 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 a 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, nos anos 2000, algumas bibliotecas adotaram a pesquisa federada, que realiza a busca simultânea em diversas fontes, apresentando os resultados em uma lista única. No entanto, 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. 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.

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, 2020, *on-line*). 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.

Assim, o objetivo deste estudo é localizar e discutir *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.

Para melhor compreensão do objetivo da pesquisa, *workflow* pode ser definido como:

> uma coleção de atividades organizadas para realizar um processo, quase sempre de negócio. Essas atividades podem ser executadas por um ou mais sistemas de computador, por um ou mais agentes humanos ou de software, ou então por uma combinação destes. Do que consistem, a ordem de execução e as pré-condições das atividades estão definidas no workflow, sendo que o mesmo é capaz ainda de representar a sincronização das atividades e o fluxo de informações entre elas (PEREIRA e CASANOVA, 2003, p. 1).

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, 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.

### DESCRIÇÃO DOS WORKFLOWS

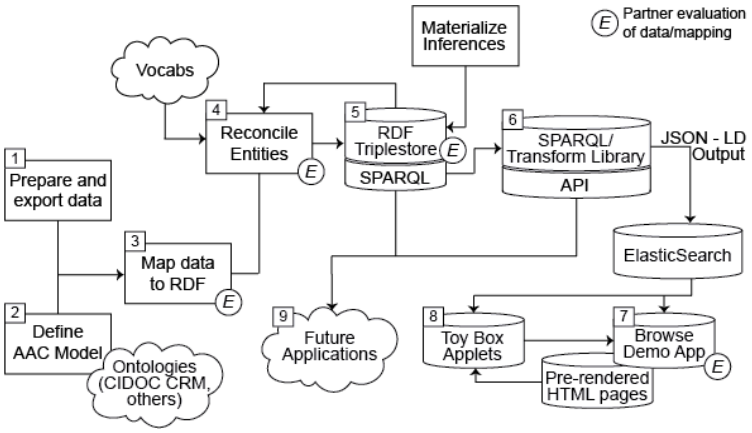
As pesquisas resultaram em oito projetos de agregação, realizados por sete instituições, listadas no quadro 1, cujos *workflows* são apresentados e detalhados nesta seção. Em alguns projetos, não foram encontrados os *workflows*, mas sim sua arquitetura, que, por mostrar etapas, foram considerados neste estudo.

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 | Itália | D-NET Software |
| 06 | Instituto de Ciência e Tecnologia da Informação | Itália | Parthenos Aggregator |
| 07 | Secretaria de Cultura | México | Repositório Mexicana |
| 08 | Universidade de Nevada | EUA | UNLV's Linked Data Project |

Fonte: Elaborado pelos autores (2020).

Figura 1 − Workflow de agregação da AAC




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

## AMERICAN ART COLLABORATIVE - AAC PIPELINE

A *American Art Collaborative* (AAC) é um consórcio de 14 instituições de arte que visam a investigar e a 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.

O *workflow* prevê nove etapas. A Etapa 1."*Prepare and export data*", em tradução livre, "Preparar e exportar os dados", visa a 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", trata-se 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 a mapear os dados das instituições parceiras para um modelo de destino e a 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 a 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 a armazenar e a 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 a 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.

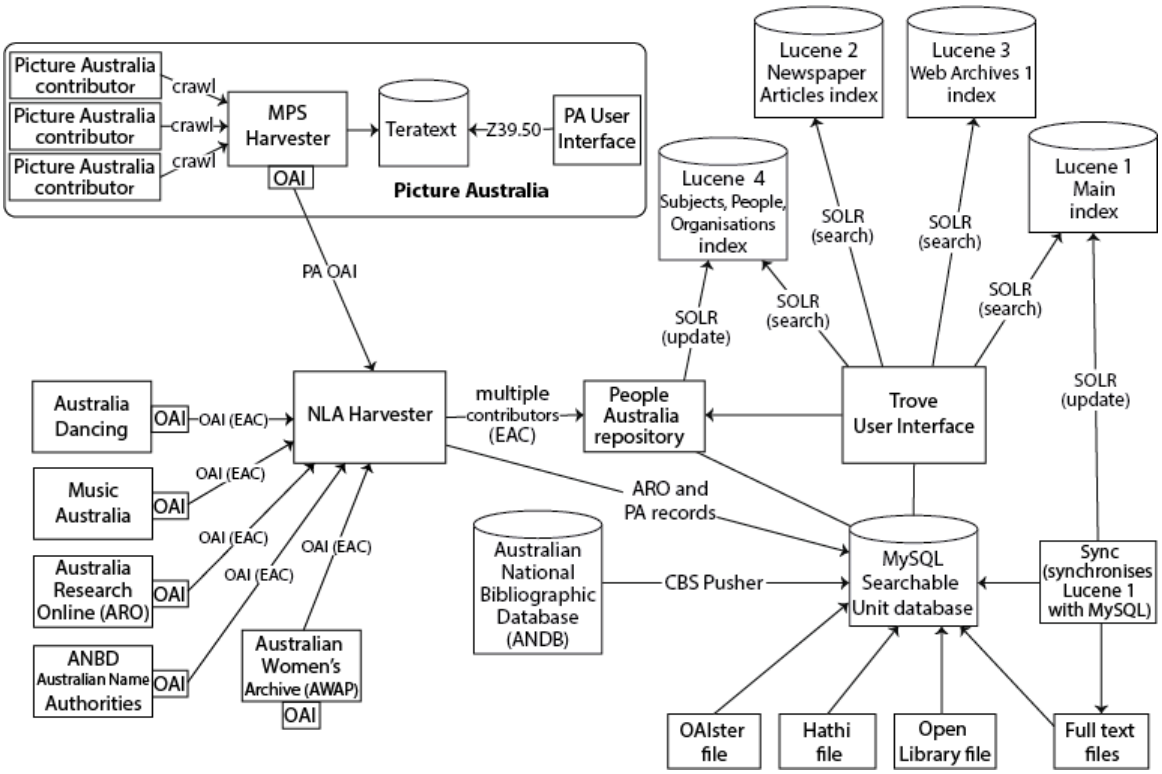
A etapa 7. "*Browse Demo App*" ou "Aplicações via browsers", visa a apresentar uma ilustração inicial suficientemente rica de dados vinculados para os usuários. A etapa 8. "*Toy Box Applets*" ou "Ampliação das aplicações via browser", visa a ampliar as possibilidades e ideias do AAC, por meio de um aplicativo de navegação. A etapa 9. "*Future Applications*" ou "Produzir aplicações futuras" visa a 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.

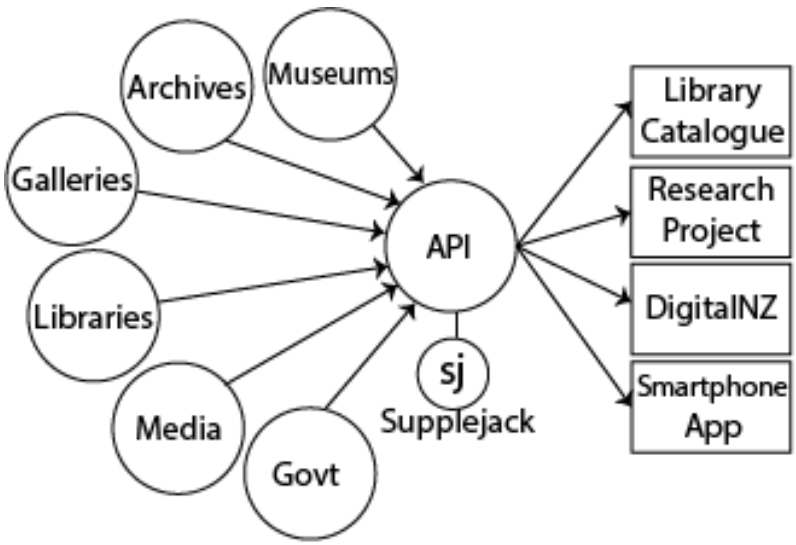
O Trove mantém um site com informações relevantes do projeto, no entanto, ainda que apresente o *workflow* de agregação, não traz, nas fontes pesquisadas, documentação que explique cada etapa do *workflow*. Contudo, na análise realizada pela pesquisa, percebe-se um grande uso do protocolo OAI-PMH para 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.

Figura 2 – Workflow de agregação da Trove




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

Figura 3 – Workflow de agregação proposto pela DigitalNZ



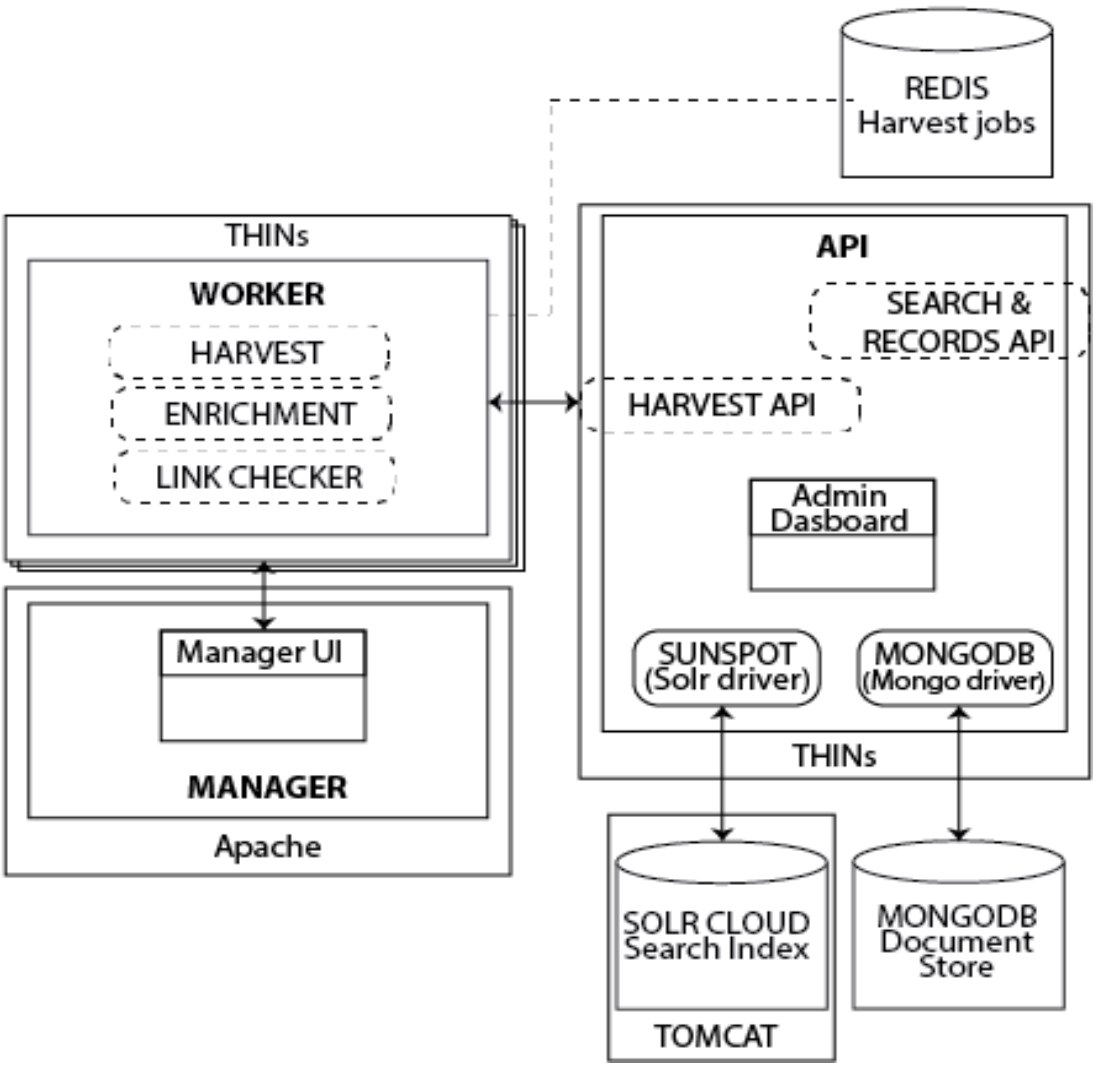

Fonte: Digital New Zealand (2018). Adaptada.

**BIBLIOTECA NACIONAL DA NOVA ZELÂNDIA – DIGITALNZ**

A Biblioteca Nacional da Nova Zelândia junto à 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.

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

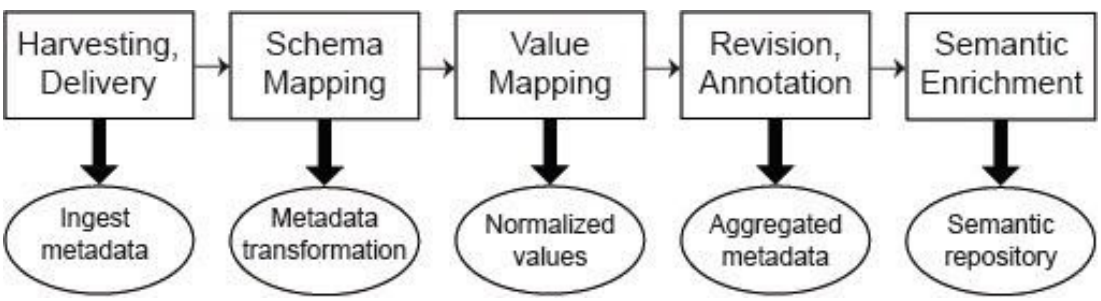
A Fundação Europeana desenvolveu a Europeana, que reuniu mais de 55 milhões de objetos digitais das coleções on-line de mais de 3.500 galerias, bibliotecas, museus, coleções audiovisuais e arquivos de toda a Europa (SCHOLZ, 2019). A figura 5 apresenta seu workflow de agregação.

A primeira etapa, "*Harvesting, Delivery*" ou "Colheita, Entrega", refere-se à coleta de metadados de provedores de conteúdo, por meio de protocolos de entrega, como o OAI-PMH, o HTTP e o 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 na 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).

Figura 5 − Workflow de agregação de dados proposto pela Europeana




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

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 a adaptação de soluções em outros contextos.

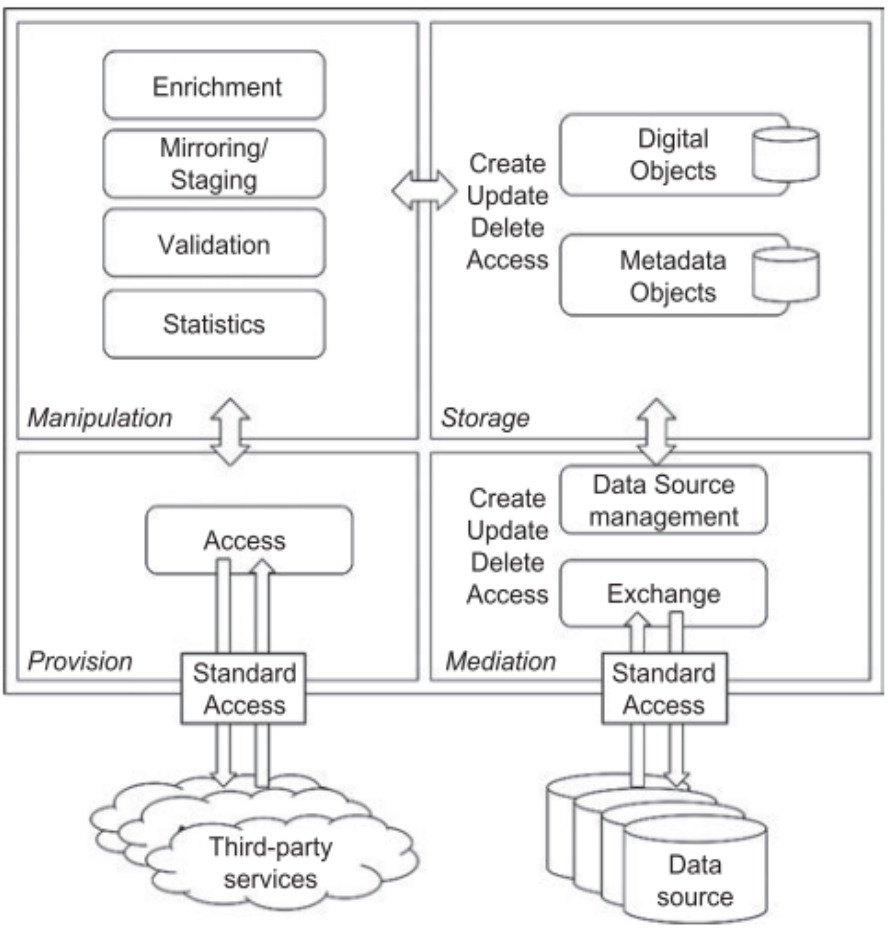
## INSTITUTO DE CIÊNCIA E TECNOLOGIA DA INFORMAÇÃO - D-NET

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 têm 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).

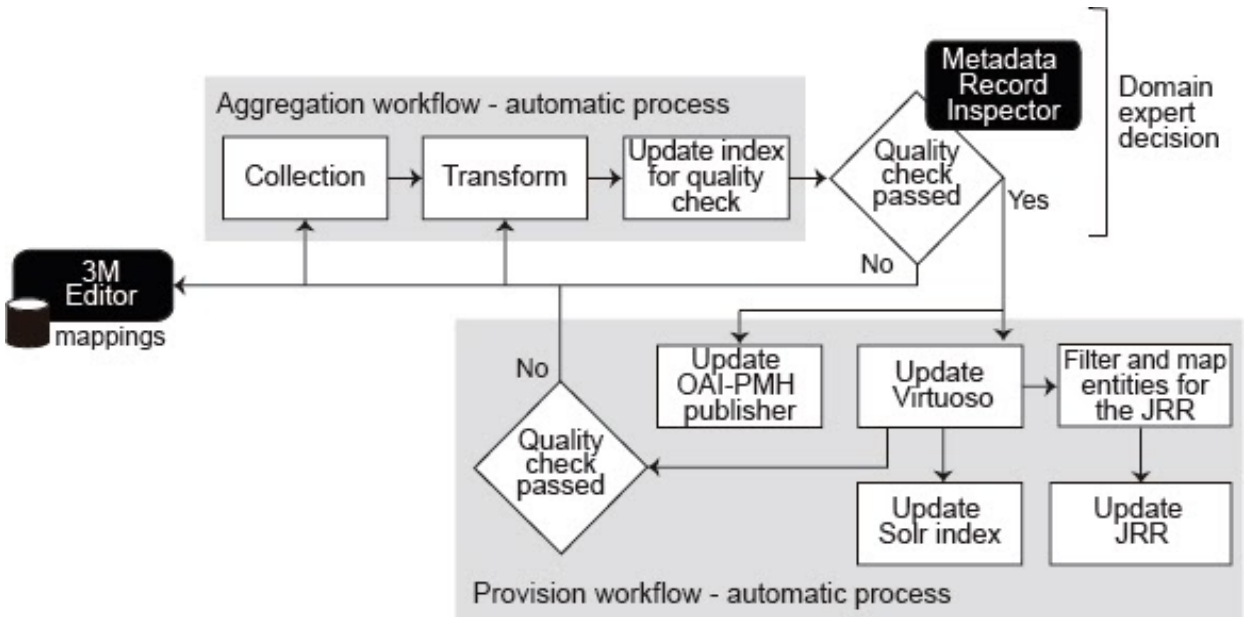
O D-NET subdivide sua arquitetura em 4 camadas principais: *Manipulation, Storage, Provision* e *Mediation*. Os serviços de manipulação, *Manipulation*, foram projetados para executar o enriquecimento, a validação, o espelhamento e a geração de estatísticas. Os serviços de armazenamento, *Storage*, como 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 a 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; 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, a Comissão Europeia lançou o projeto *Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies* (PARTHENOS) (FROSINI *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.

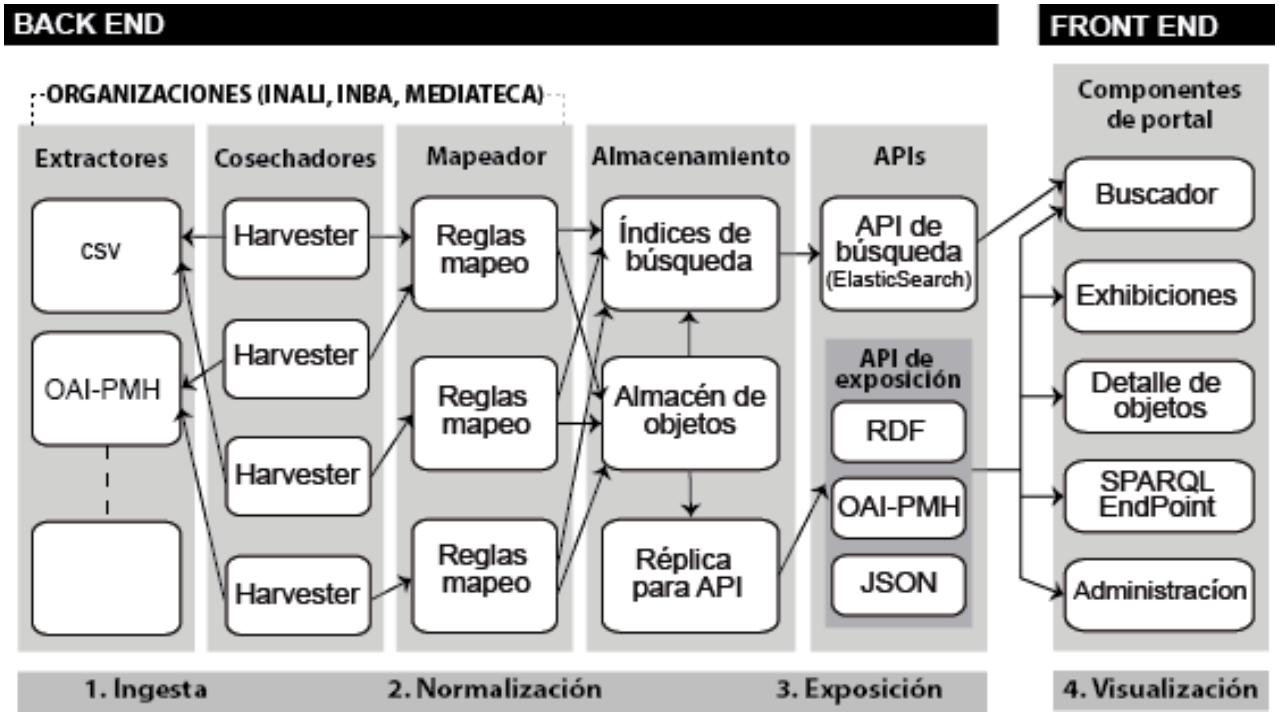
O *workflow* se divide em dois fluxos de trabalho: agregação e provisionamento. No fluxo de agregação, a etapa "*Collection*" ou "Coleta" visa a lidar com a coleta de metadados por meio de diferentes protocolos de acesso: OAI-PMH, FTP(S), SFTP, HTTP(S), RESTful. A etapa "*Transformation*" ou "Transformação" visa a mapear os metadados a uma ontologia única. A "*Metadata Cleaner*", ou "Limpeza de metadado", trata-se 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 integrada 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 transformar 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 (MÉXICO, 2018). Seu *workflow* está apresentado na figura 8.

Figura 8 – Workflow de agregação Mexicana




Fonte: México (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", trata-se dos componentes de código responsáveis pela cópia de dados locais ou remotos, realizando uma reestruturação mínima dos dados. Nesta etapa, considera-se 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 a 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 inseridos 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 a configuração e execução das coletas para fontes de dados.

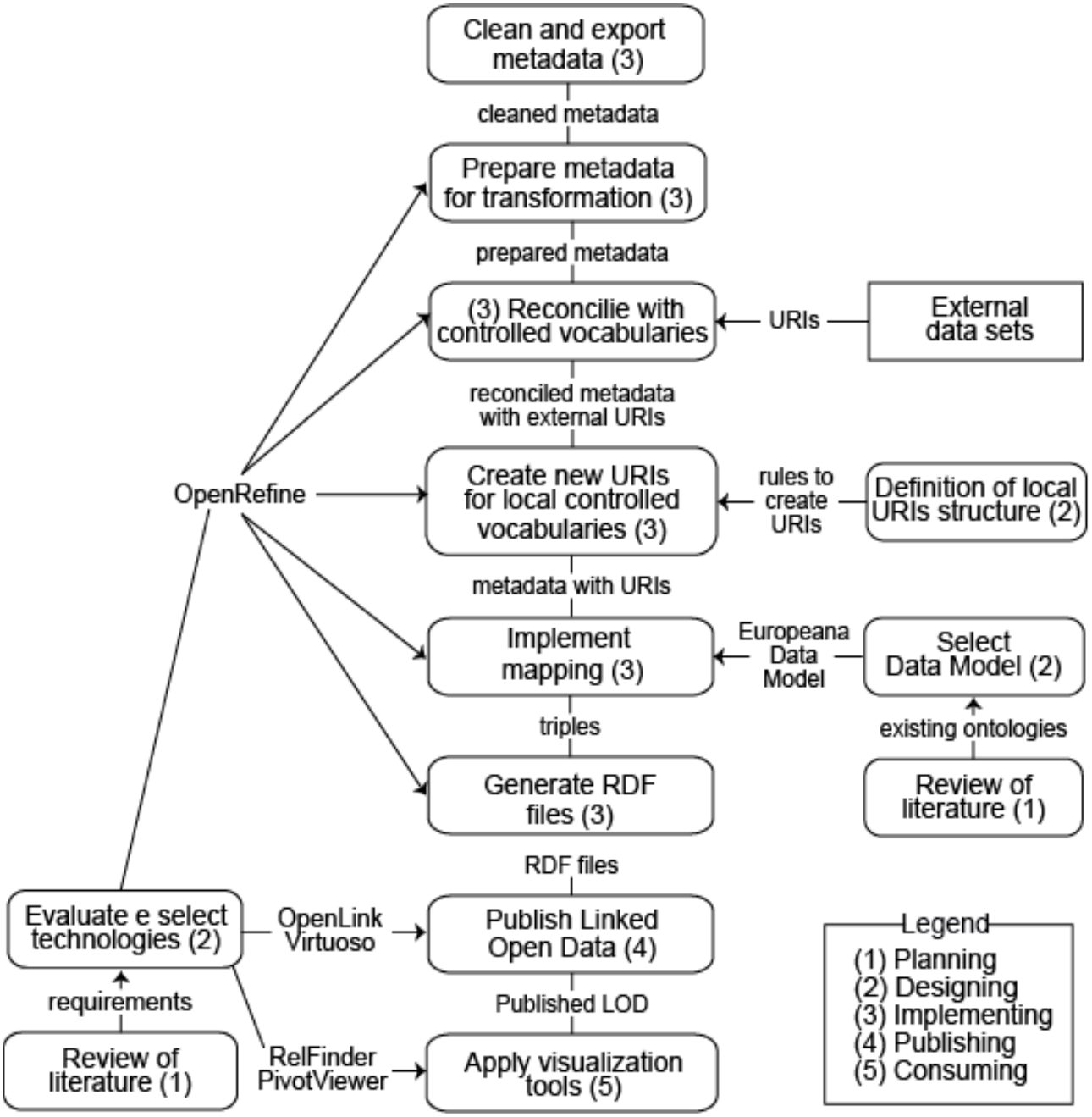
## UNIVERSIDADE DE NEVADA - UNLV'S LINKED DATA PROJECT

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 o 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*.

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.

Figura 9 – Workflow de agregação de dados proposto pela Universidade de Nevada




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

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, utilizaram-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 a 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.

## DISCUSSÃO DOS RESULTADOS

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 (2020).

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 | - | - |
| DigitalNZ | X | - | - | - | X | X | - | - |
| Europeana [2] | X | X | - | X | X[2] | X[2] | - | - |
| D-NET Software | X | X | X | X | X | X | X | X |
| Parthenos Aggregator | X | X | X | X | X | X | X | X |
| Repositório Mexicana | X | X | - | X | X | X | X | - |
| UNLV's Linked Data Project | X | X | X | X | X | X | X | X |
| [1] Dados observados somente a partir da visualização dos workflows [2] Itens não explícitos no workflow, mas identificados na documentação | | | | | | | | |

Fonte: Elaborado pelos autores (2020).

O quadro 4 apresenta a nomenclatura original das etapas na literatura revisada, classificando-as dentro das etapas elencadas neste estudo. 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 software utilizados 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 há 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 clara 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 são 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 à 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 para 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 AAC Model* | - | *Reconcilie Entities* | *RDF Triple Store - SPARQL* | *Browse Demo APP e Toy Box Applets* | *SPARQL/ Transform Library API* | *Future Applications* |
| Trove [1] | *NLA Harvest* | - | - | - | *MySQL Seachable Unit Database* | *Trove User Interface* | - | - |
| DigitalNZ | *Manager e Common* | - | - | - | *API* | *API* | - | - |
| Europeana | *Harvesting Delivery* | *Schema Mapping* | - | *Value Mapping* | *X* [2] | *X* [2] | - | - |
| D-NET Software | *Mediation* | *Manipulation* | *Manipulation* | *Manipulation* | *Storage* | *Provision* | *Provision* | *Provision* |
| Parthenos Aggregator | *Collection* | *Transformation* | *Metadata Record Inspector* | *Metadata Cleaner* | *Index Service* | *X* [2] | *OAI-PMH publisher service* | *OAI-PMH publisher service* |
| Repositório Mexicana | *Extractores e Cosechadores* | *Mapeador* | - | - | *Almacienamento* | *Buscador e Exhibiciones* | *API de Búsqueda* | - |
| UNLV's Linked Data Project | *Clean and export* | *Implement mapping* | *Prepare matadata for transformation* | *Reconcilie with controlled vocabularies e Create new URIs for local controlled vocabularies* | *Publish LOD* | *Publish LOD* | *Apply visualization tools* | *Publish LOD* |
| [1] Dados observados somente a partir da visualização do workflow | | | | | | | | |
| [2] Item não explicito no workflow | | | | | | | | |

Fonte: Elaborado pelos autores (2020).

## REFERÊNCIAS


BARDI, A.; MANGHI, P.; ZOPPI, F. Aggregative Data Infrastructures for the Cultural Heritage. *In*: DODERO, J. M.; PALOMO-DUARTE, M.; KARAMPIPERIS, P. (org.). *Metadata and Semantics Research*. Communications in Computer and Information Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. v. 343, p. 239–251. DOI 10.1007/978-3-642-35233-1_24. Disponível em: http://link.springer.com/10.1007/978-3-642-35233-1_24. Acesso em: 11 mar. 2021.

BRIGHAM, T. J.; FARRELL, A. M.; OSTERHAUS TRZASKO, L. C.; ATTWOOD, C. A.; WENTZ, M. W.; ARP, K. A. Web-Scale Discovery Service: Is It Right for Your Library? Mayo Clinic Libraries Experience. *Journal of Hospital Librarianship*, v. 16, n. 1, p. 25–39, 2 jan. 2016. DOI 10.1080/15323269.2016.1118280. Disponível em: http://www.tandfonline.com/doi/full/10.1080/15323269.2016.1118280. Acesso em: 11 mar. 2021.

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

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

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

FINK, E. 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, L.; BARDI, A.; MANGHI, P.; PAGANO, P. An Aggregation Framework for Digital Humanities Infrastructures: The PARTHENOS Experience. *SCIentific RESearch and Information Technology*, v. 8, n. 1, 11 jul. 2018. DOI 10.2423/i22394303v8n1p33. Disponível em: https://doi.org/10.2423/i22394303v8n1p33. Acesso em: 11 mar. 2021.

KOLLIA, I.; TZOUVARAS, V.; DROSOPOULOS, N.; STAMOU, G. A systemic approach for effective semantic access to cultural content. *Semantic Web*, v. 3, n. 1, p. 65–83, 2012. DOI 10.3233/SW-2012-0051. Disponível em: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-2012-0051. Acesso em: 11 mar. 2021.

MANGHI, P.; ARTINI, M.; ATZORI, C.; BARDI, A.; MANNOCCI, A.; LA BRUZZO, S.; CANDELA, L.; CASTELLI, D.; PAGANO, P. The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures. *Program*, v. 48, n. 4, p. 322–354, 27 ago. 2014. DOI 10.1108/PROG-08-2013-0045. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/PROG-08-2013-0045/full/html. Acesso em: 11 mar. 2021.

MÉXICO. SECRETARÍA DE CULTURA. *Mexicana Repositorio del Patrimonio Cultural de México*. Ciudad de México: Secretaría de Cultura, 2018. Disponível em: https://mexicana.cultura.gob.mx/work/models/repositorio/Resource/126/2/images/documentacion.pdf. Acesso em: 18 abr. 2020.

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, T. *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, C. M. G.; CAREGNATO, S. E. Serviços de descoberta em rede: a experiência do modelo Google para os usuários de bibliotecas universitárias. *Em Questão,* v. 21, n. 3, p. 130, 2015. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/58410/36046. Acesso em: 27 abr. 2020.

PEREIRA, L. A. M.; CASANOVA, M. A. *Sistemas de gerência de workflows: características, distribuição e exceções.* Rio de Janeiro: PUC-Rio, 2003. Disponível em: ftp://ftp.inf.puc-rio.br/pub/docs/techreports/03_11_pereira.pdf. Acesso em: 27 ago. 2020.

SCHOLZ, H. *A guide to the metadata and content requirements for data partners publishing material in Europeana Collections.* [*S. l.*]: Europeana Foundation, 2019(Europeana Publishing Guide, v1.8). Disponível em: https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Publishing%20Guide%20v1.8.pdf. Acesso em: 18 abr. 2020.

SOUTHWICK, S. 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, 2 jan. 2015. DOI 10.1080/19386389.2015.1007009. Disponível em: http://www.tandfonline.com/doi/abs/10.1080/19386389.2015.1007009. Acesso em: 11 mar. 2021.

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

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.


## AGRADECIMENTOS

Agradecimento ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPQ que financiou a pesquisa por meio da bolsa de doutorado.