WORKFLOW DE AGREGAÇÃO DE DADOS: PROCESSOS PARA CRIAÇÃO DE UMA INTERFACE DE BUSCA INTEGRADA DO PATRIMÔNIO CULTURAL

DATA AGGREGATION WORKFLOW: PROCESSES FOR CREATING A CULTURAL HERITAGE INTEGRATED SEARCH INTERFACE

Joyce Siqueira[1], Dalton Lopes Martins[2]
(1) Universidade de Brasília, Campus Darcy Ribeiro, Asa Norte, Brasília – DF, joycitta@gmail.com
(2) Universidade de Brasília, Campus Darcy Ribeiro, Asa Norte, Brasília – DF, daltonmartins@unb.br

Resumo:
Nos últimos anos diferentes instituições culturais vêm envidando esforços para difundir a cultura por meio da construção de uma interface única de busca integrada para seus objetos digitais. Estes esforços resultaram em diferentes propostas para agregação de dados, para as quais foram construídos workflows, que apresentam as etapas necessárias a esse intento. Considerando a possibilidade de diferentes workflows, com etapas distintas, esta pesquisa objetiva destaca-los e discuti-los. Para tal, foi realizada pesquisa descritiva e bibliográfica, de natureza qualitativa, em bases de dados acadêmicas e na literatura cinzenta. Como resultado, apresenta-se sete workflows de agregação propostos pela: American Art Collaborative, Biblioteca Nacional da Nova Zelândia, Fundação Europeana, Instituto de Ciência e Tecnologia da Informação, Secretaria de Cultura do México, Biblioteca Nacional da Austrália e Universidade de Nevada. A análise do conjunto de workflows resultou em oito diferentes etapas a serem executadas: 1. extrair, 2. usar ontologias, 3. transformar, 4. reconciliar, 5. armazenenar, 6. expor, 7. publicar 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 benefício em outras instituições.
Palavras-chave: workflow de agregação; busca integrada; patrimônio cultural; instituições culturais

Abstract:
In recent years, different cultural institutions have been making efforts to spread culture through the construction of a single, integrated search interface for their digital objects. These efforts resulted in different proposals for data aggregation, for which workflows were built, which present the necessary steps for this purpose. Considering the possibility of different workflows, with different steps, this research aims to highlight and discuss them. To this end, a descriptive and bibliographical research of qualitative nature was conducted in academic databases and in gray literature. As a result, we present seven aggregation workflows proposed by: American Art Collaborative, New Zealand National Library, Europeana Foundation, Institute of Information Science and Technology, Secretariat of Culture of Mexico, National Library of Australia, and University of Nevada. Analyzing the set of workflows resulted in eight different steps to perform: 1. extract, 2. use ontologies, 3. transform, 4. reconcile, 5. store, 6. expose, 7. publish, and 8. enable new applications. In addition, it is also visible the need for more detailed steps, so that it is possible to replicate the workflow, and enjoy its benefits in other institutions.
Keywords: aggregation workflow; integrated search; cultural heritage; cultural institutions

1. Introdução

Difundir a cultura por meio da oferta de uma interface de busca integrada, que possibilite aos usuários uma navegação eficiente pelos diversos objetos digitais que compõe o patrimônio cultural é um objetivo fortemente almejado, de tal forma que, nos últimos anos, grandes instituições culturais envidam esforços para construí-la.

Parte integrante deste procedimento se configura na construção de workflows que apresentam as etapas necessárias a agregação de dados, de forma que o mesmo contemple todos os processos e tecnologias, da extração a publicização dos dados agregados.

Neste ponto, é importante ter claro que um workflow se trata de uma maneira de organizar etapas em uma sequência produtiva e eficiente, sendo estas planejadas, modeladas e automatizadas de forma a atingir propósitos bem definidos (SANTOS, 2013).

O objetivo deste estudo é localizar diferentes workflows de agregação de dados, desenvolvidos por instituições culturais, para realizar uma análise qualitativa das etapas escolhidas por cada instituição.

Dessa forma, o primeiro passo dado foi encontrar quais instituições propuseram soluções para a agregação de dados. Procurou-se, inicialmente, por consolidadas instituições reconhecidas por realizar este trabalho, tais como a Europeana e a Mexicana, e na sequência, por meio de palavras-chaves, novas propostas foram selecionadas.

Ao final, foram localizados sete workflows de Instituições culturais que são brevemente apresentadas na seção de Resultados.

Este artigo está assim dividido: seção 2, Objetivos, seção 3, Procedimentos Metodológicos, seção 4, Resultados, e por último, na seção 5, as Considerações Finais.

2. Objetivos

Esta pesquisa tem o propósito de analisar workflows de agregação de dados, desenvolvidos por instituições culturais. Dessa forma os objetivos específicos são: 1. localizar as instituições culturais que propuseram workflows de agregação; 2. apresentar os workflows e 3. identificar as etapas propostas.

3. Procedimentos Metodológicos

Pesquisa de caráter descritivo e bibliográfico, de natureza qualitativa, realizada em bases de dados acadêmicas e na literatura cinzenta, como intuito de encontrar o workflow de agregação de dados de reconhecidas instituições, além novas iniciativas.

As buscas foram realizadas no Google, Google Acadêmico, EBSCOhost e BRAPCI, utilizando os termos: *"pipeline", "workflow", "architecture", "aggregation", "metadata ingest", "metadata aggregation", "europeana", "mexicana", "dpla", "digital public library of america", "trove", "digitalnz"*, "aggregative data infrastructures".

Optou-se pelo Google para localizar workflows na literatura cinzenta e as demais bases por serem agregadoras de outras bases, tornado a pesquisa mais ampla. No caso da BRAPCI, também é uma base específica da área de Ciência da Informação.

4. Resultados

Foram localizados sete workflows de agregação, apresentados nas Figura 01 a 08, dispostas no Apêndice A, que são tratados nessa seção. Inicialmente, apresenta-se um breve resumo de cada Instituição propositora dos workflows.

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

Para Fink (2018), LOD que se trata de um método para publicar dados estruturados na web de forma que as informações sejam interconectadas e, assim, tornadas amplamente úteis.

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).

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 LOD culminando no desenvolvimento do *UNLV's Linked Data Project* (SOUTHWICK, 2015).

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, 2018).

O Istituto di Scienza e Tecnologie dell'Informazione desenvolveu o D-NET, um software que oferece um kit de serviços para a construção de Infraestruturas de dados (BARDI, MANGHI E ZOPPI, 2012).

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 Biblioteca Nacional da Austrália desenvolveu o Trove, que tem o objetivo de fornecer recursos relacionados à Austrália. Além de um mecanismo de busca, reúne conteúdo de bibliotecas, museus, arquivos e outras organizações de pesquisa e fornece um conjunto de serviços (TROVE HELP CENTRE, 2019).

Considerando a análise dos workflows, foram encontradas oito fases para agregação, sendo elas: extração, uso de ontologias, 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. Utilizar ontologias: selecionar vocabulários controlados pré-existentes 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' novas aplicações podem ser criadas.

O Quadro 01 mostra um panorama do uso de cada fase.

Quadro 01. Etapas dos Workflow de Agregação

Projeto/ Etapas	Extrair	Utilizar ontologias	Transformar	Reconciliar	Armazenar	Publicar	Expor	Possibilitar novas aplicações
AAC	X	-	X	X	X	X	X	X
DigitalNZ	X	-	-	X	X	X	-	-
D-NET Software	X	-	X	X	-	X	X	-
Europeana	X	X	X	X	X	-	-	-
Mexicana	X	-	X	X	X	X	X	X
TROVE	X	-	-	-	X	X	-	-
UNLV's Linked Data Project	X	X	X	X	X	X	X	-

Fonte: elaborado pelos autores

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.

Além das etapas, as publicações apresentação algumas ferramentas de softwares utilização para execução do workflow, e contata-se que não há escalabilidade, ou seja, à medida que o fluxo de dados cresce, o workflow torna-se impraticável.

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.

5. Considerações Finais

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.

Como trabalho futuro, pretende-se realizar pesquisas direcionadas a cada etapa do workflow, além de propor um novo workflow, ainda mais completo.

Referências

BARDI, Alessia; MANGHI, Paolo; ZOPPI, Franco. **Aggregative data infrastructures for the cultural heritage.** In: Research Conference on Metadata and Semantic Research. Springer, Berlin, Heidelberg, 2012. p. 239-251.

DIGITAL NEW ZEALAND. **This is Digital New Zealand**. YouTube, 20 dez 2018. Disponível em: https://www.youtube.com/watch?v=UWbIDwsaA4o. Acesso em 14 set 2019.

DIGITAL NEW ZEALAND. **Our History**. Disponível em: https://digitalnz.org/about/our-history. 2019. Acesso em 14 set 2019.

FINK, Eleanor E. **American Art Collaborative (AAC) Linked Open Data (LOD) Initiative: overview and recommendations for good practices**. *Am. Art Collab.*, 2018.

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.

NATIONAL LIBRARY OF AUSTRÁLIA. Trove Help Center. **Trove System Architecture Diagram**. 2010. Disponível em: https://www.nla.gov.au/trove/marketing/Trove%20architecture%20diagram.pdf. Acesso em 14 set 2019.

SANTOS, Daniel Soares. **Automatização de Processos de Negócios Utilizando BPM/BPMS.** Monografia (Ciência da Computação) - Universidade Estadual do Sudoeste da Bahia. Vitória da Conquista – Bahia, p. 109. 2013.

SCHOLZ, Authors Henning; FANTONE, Federica. Europeana Publishing. n. September, p. 1–31, 2018.

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 Mexico. 2018.

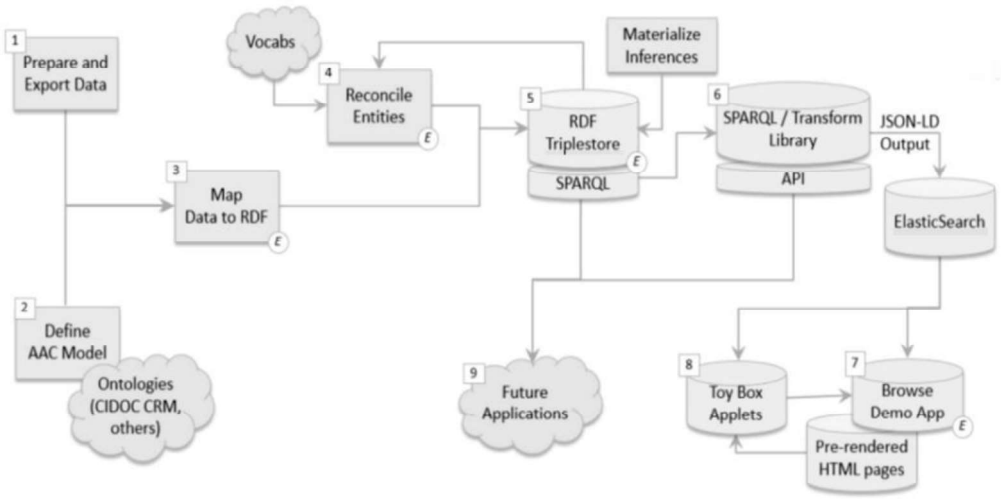
SOUTHWICK, Silvia B. **A guide for transforming digital collections metadata into linked data using open source technologies**. Journal of Library Metadata, vol. 15, no. 1, pp. 1–35, 2015.

TROVE HELP CENTRE. **About Trove.** 2019. Disponível em: https://help.nla.gov.au/trove/using-trove/getting-to-know-us. Acesso em 22 set 2019

SUPPLEJACK. **Architecture.** Documentation (Version 0.1). 2019. Disponível em: http://digitalnz.github.io/supplejack/architecture.html. Acesso em 22 set 2019.

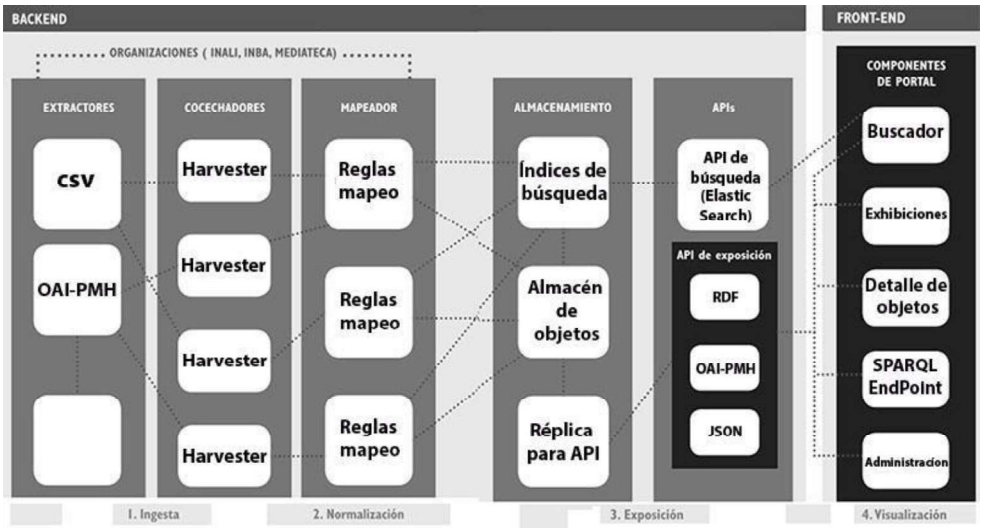
Apêndice A – Workflow de Agregação

Figura 01. Workflow de agregação proposto pela *American Art Collaborative*



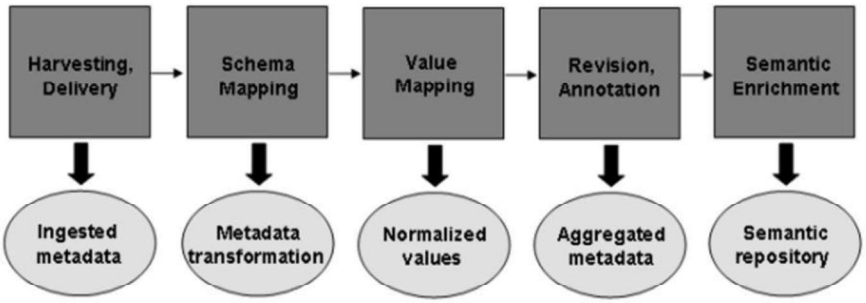
Fonte: Fink (2018), adaptada.

Figura 02. Workflow de agregação proposto pela Mexicana



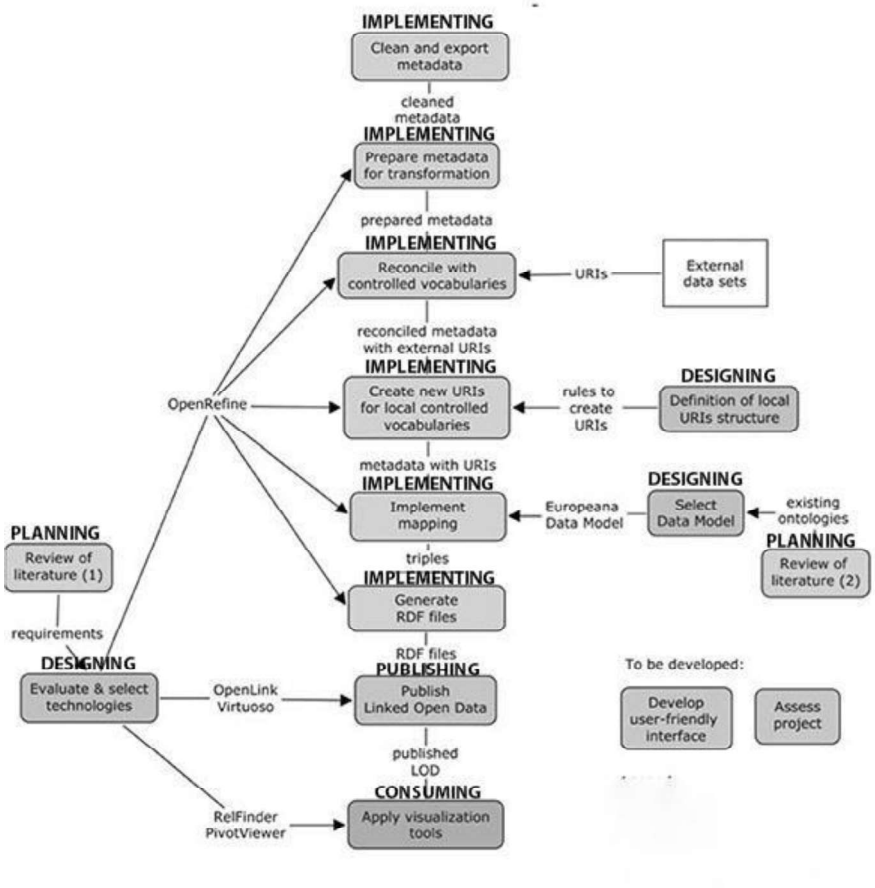
Fonte: Secretaría de Cultura (2018), adaptada

Figura 03. Workflow de agregação proposto pela Europeana



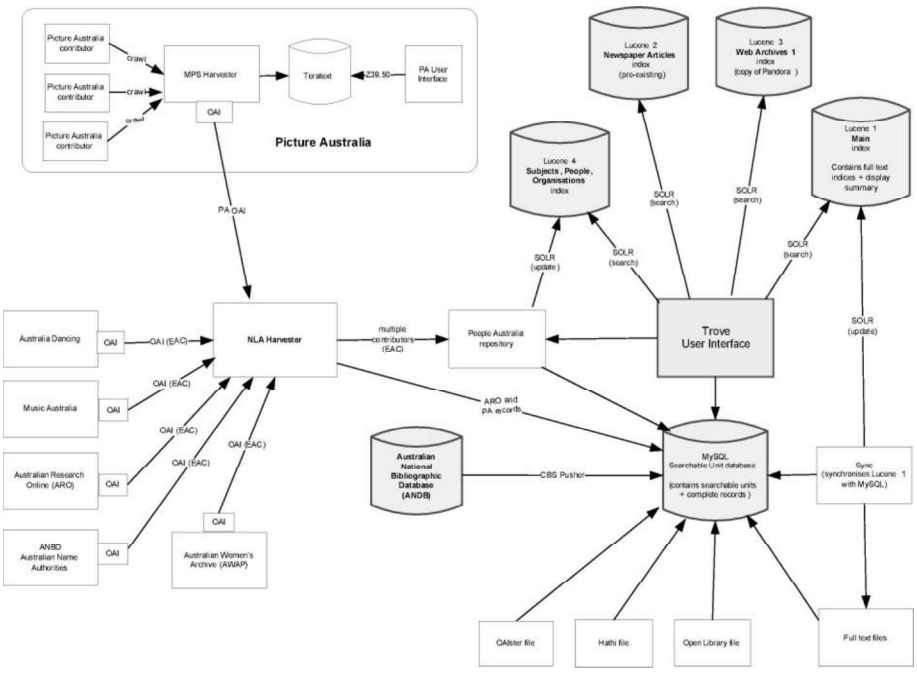
Fonte: Kollia et al (2012), adaptada

Figura 04. Workflow de agregação proposto pela University of Nevada



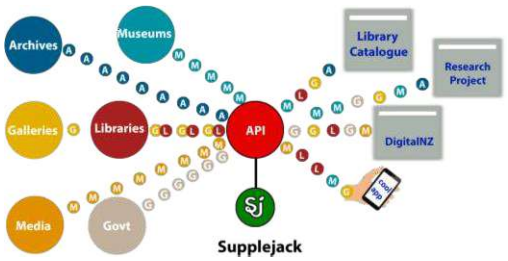
Fonte: Southwick (2015)

Figura 05. Workflow de agregação proposto pela TROVE



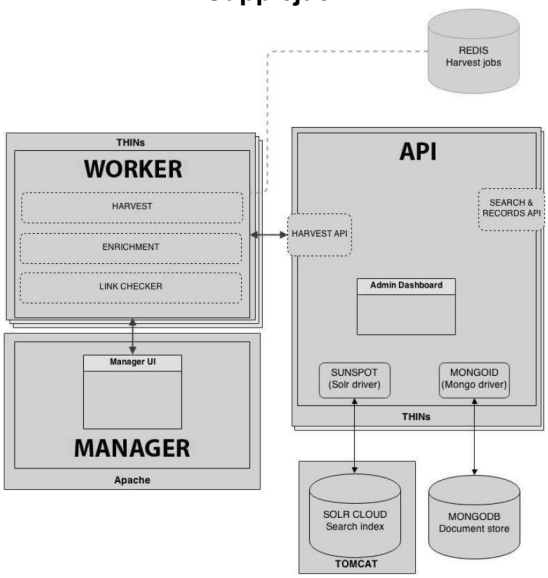
Fonte: National Library of Austrália (2010)

Figura 06. Workshop de agregação proposto pela DigitalNZ



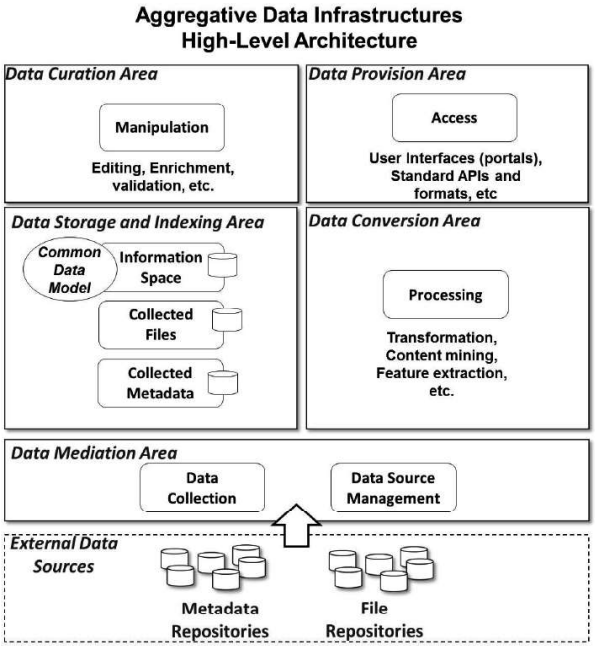
Fonte: Digital New Zealand (2018), adaptada

Figura 07. Arquitetura da plataforma Supplejack



Fonte: Supplejack (2019)

Figura 08. D-NET Software Toolkit, proposto pelo Istituto di Scienza e Tecnologie dell'Informazione



Fonte: Bardi, Manghi e Zoppi (2012), adaptada