Solução para integração entre plataformas com tradução de modelo

Harrysson Gilgamesh de Medeiros Nóbrega[a]

1. Introdução

No ecossistema da Educação, Ciência, Tecnologia e Inovação – ECTI, o desafio de permitir o uso ideal de dados e métodos de pesquisa é complexo, com várias partes interessadas: pesquisadores que desejam compartilhar seus dados e interpretações; editores de dados profissionais que oferecem os seus serviços, software e construtores de ferramentas que fornecem serviços de análise e processamento de dados; agências de financiamento (privadas e públicas) cada vez mais preocupadas com a qualidade dos trabalhos financiados; e uma comunidade de ciência de dados, integrando e analisando a saída para avançar na descoberta.

A análise computacional para descobrir padrões significativos em conjuntos de dados massivos e interligados está rapidamente se tornando uma atividade de pesquisa de rotina. Fornecer dados legíveis por máquina como o substrato principal para a descoberta do conhecimento e para que esses processos funcionem sem problemas e de forma sustentável é um dos

a Mestrando em Ciência da Informação pela Universidade Estadual Paulista – UNESP. Gerente de Projetos na Rede Nacional de Pesquisa – RNP. E-mail: harrysson.nobrega@unesp.br. ORCID: https://orcid.org/0000-0002-0378-4782.

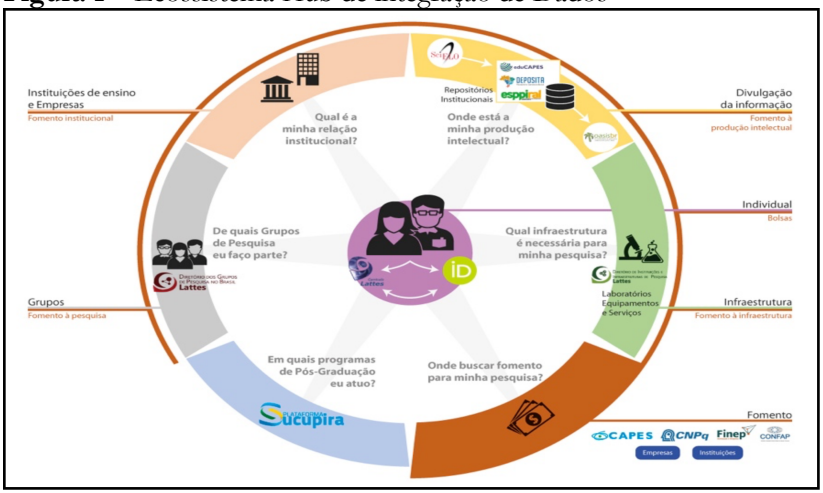
grandes desafios das equipes de ciência da informação dentro das agências do governo brasileiro.

Vivenciamos um cenário de mudança na forma como utilizamos os recursos públicos destinados a educação brasileira. Com os constates cortes no orçamento das agências de governo e a necessidade de dispormos de mecanismos tecnológicos eficientes objetivando mitigar o retrabalho dos pesquisadores e coordenadores de programas no contexto do preenchimento dos dados acadêmicos produzidos para fins de avaliação pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES); é um exemplo dos problemas enfrentados atualmente. Em um cenário heterogêneo, em termos de qualidade de dados disponíveis nos sistemas acadêmicos espalhados em diversas instituições de ensino, identificamos a oportunidade de construirmos um conjunto de ferramentas de apoio a integração dos sistemas acadêmicos existentes nas Pró-reitorias de ensino superior e as plataformas de gestão da informação do governo brasileiro.

Nesse artigo trataremos especificamente de duas soluções: FAIS – Expositor e FAIS – Integrador. Ambos foram frutos de um trabalho em conjunto entre o time de desenvolvimento do *Hub* de integração de Dados (Figura 1) e o time de pesquisa e desenvolvimento da UNESP. O grupo de trabalho teve início em março de 2020 com a apresentação da ideia e constituição do grupo. Durante o ano de 2020 fizemos vários experimentos dentro do ambiente virtual da UNESP, identificando as principais

dificuldades e necessidades de construção para a criação de um conjunto de ferramentas que posteriormente foram chamadas de FAIS – Ferramentas de Apoio a Integração de Sistemas.

Figura 1 – Ecossistema Hub de integração de Dados



Fonte: ConectiBrasil. Disponível em: https://www.conectibrasil.org/. Acesso em: 06 nov. 2021.

O *Hub* de Integração de Dados se configura como um consórcio multi-institucional que prevê a criação de uma plataforma nacional de integração de dados relacionados a educação, ciência, tecnologia e inovação (ECTI) com a finalidade de: (i) Promover ferramentas integradoras de informações entre as instituições brasileiras partícipes do Consórcio, para aumento da racionalidade do sistema nacional de informações de ECTI; (ii) Gerar e disponibilizar informações da ECTI brasileira seguindo

padrões nacionais e internacionais; (iii) Promover programas de geração e disponibilização de dados abertos; (iv) Viabilizar acordos internacionais em consórcio (ex. assinatura e adoção do ORCID como identificador persistente de pesquisadores).

O *Hub* de Integração de Dados prevê criar um ecossistema de informações em pesquisa com foco no pesquisador, integrando os sistemas federais e estaduais e otimizando a gestão do conhecimento.

2. Procedimentos metodológicos

O tema dados abertos de pesquisa vêm sendo estudado desde janeiro de 2004, quando os ministros de Ciência e Tecnologia dos países-membros da OECD se reuniram em Paris, para discutir a necessidade de um guia internacional voltado para o acesso aos dados de pesquisa. Esta reunião culminou na "*Declaration on access to research data from public funding"*, que estabelece objetivos e princípios relacionados com as ações de abertura dos dados, tais como: transparência, conformidade legal, responsabilidade formal, profissionalismo, proteção da propriedade intelectual, interoperabilidade, qualidade e segurança, além da eficiência e prestação de contas para a sociedade. Em 2007 esta mesma instituição lança uma nova versão dessa declaração, incluindo quatro novos princípios: qualidade, segurança, flexibilidade e sustentabilidade (ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2007).

Silva et al. (2017) propõem um modelo de comunicação científica que prevê os elementos da Ciência Aberta desde o início da pesquisa, por meio das novas práticas de compartilhamento. No processo tradicional, o compartilhamento/publicação ocorria principalmente por intermédio de periódicos, teses e dissertações; agora, pode ser feito desde o delineamento da pesquisa, por meio do registro do projeto, da gestão de dados científicos, da publicação em *pre-print* e da avaliação por pares, todos estes de forma aberta, cujo reconhecimento é baseado em práticas abertas e em sua reutilização pela educação aberta.

O *National Institute of Health* (NIH), instituto de pesquisa em saúde dos Estados Unidos, pode ser considerado um exemplo que demonstra a preocupação com os dados abertos já há alguns anos. Esta instituição foi a pioneira na criação de uma política de compartilhamento de dados de pesquisa. O documento gerado em outubro de 2003 determinou que: "Os dados devem estar disponíveis livremente o mais rápido possível, salvaguardando a privacidade dos participantes e protegendo dados confidenciais e proprietários" (NATIONAL INSTITUTES OF HEALTH, 2003 *apud* COSTA; LEITE, 2017, p. 92).

As primeiras manifestações referentes aos princípios FAIR surgiram no início de 2014, quando o nó holandês do grupo Elixir, em cooperação comas instituições: Netherlandse Science Center, o *Dutch Techcentre for Life Sciences* (DTL) e o Centro Lorentz, organizaram uma conferência intitulada *Jointly designing a*

data FAIRPORT. Nesta conferência foram discutidos os obstáculos relativos à utilização e reutilização de dados e as propostas para a solução destes problemas, o que seria feito por intermédio da criação de uma infraestrutura global para dados no contexto do eScience. Estiveram presentes especialistas e pesquisadores, membros de institutos de pesquisa, editores, especialistas em web semântica, inovadores e cientistas da computação (DATA FAIRPORT, 2014).

O resultado da reunião foi um consenso quanto à necessidade de criar uma infraestrutura global que suporte a abertura dos dados de pesquisa voltados para a publicação, o compartilhamento e a reutilização de dados. Além destes aspectos, houve também um conjunto de discussões sobre os requisitos para gerenciamento dos dados de forma aberta que foi promovido pelas agências de fomento à pesquisa. Outro resultado de destaque foi a proposta de construção de um *back bone* para possibilitar a interoperabilidade global dos dados, viabilizando que os computadores possam interagir e descobrir de forma automática todos os conjuntos de dados disponíveis para uma pesquisa específica (DATA FAIRPORT, 2014).

Estas propostas suscitaram a criação de princípios e práticas que pudessem orientar a descoberta, o acesso, a integração e a reutilização da vasta quantidade de dados e informação gerada pela ciência. Esses princípios são regulados por intermédio de protocolos que definem a interoperabilidade semântica básica dos conjuntos de dados, quando possível, usando-se padrões e

protocolos comunitários já aprovados ou emergentes (WILKINSON et al., 2016).

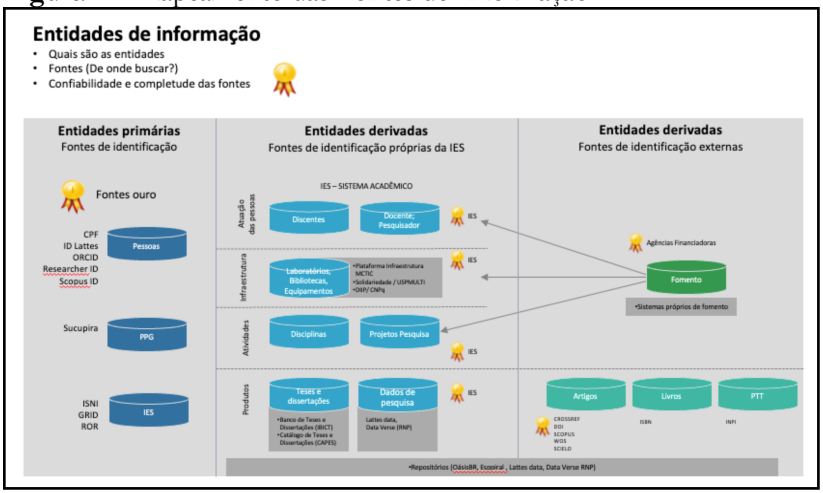
No segundo semestre de 2019 realizamos uma pesquisa de campo exploratória onde entrevistamos representantes de cada participe do projeto, objetivando identificarmos as principais dificuldades e oportunidades de cada um deles. A pesquisa durou 6 meses, resultando na identificação dos modelos dedados comuns entre cada partícipe e as principais necessidades comuns.

No primeiro semestre de 2020 iniciamos a pesquisa junto as universidades de ensino superior. Na ocasião foram entrevistadas instituições públicas e privadas, estaduais e federais. A metodologia foi a mesma utilizada com os partícipes e os resultados foram semelhantes. A pesquisa durou 12 meses, devido o maior número de universidades visitadas, resultando na identificação dos modelos de dados, tecnologias existentes e principais necessidades e expectativas. Após o processo de pesquisa de campo, pudemos mapear as fontes de informação existentes em cada parceiro, sua origem e a possibilidade de troca. Também identificamos as principais oportunidades (Figura 2).

Considerando a diversidade de modelos de dados e padrões existentes em todo o ecossistema da Educação, Ciência, Tecnologia e Inovação – ECTI, pretendemos apresentar uma solução tecnológica para interoperar e interagir com um escopo diversificado de sistemas utilizados por diversas instituições de ensino. Desenhamos um produto customizável, sendo capaz de se adaptar e ser utilizado em qualquer ecossistema independente

das diferentes faixas de dados e maturidade tecnológica. Os testes preliminares mostraram uma solução robusta e satisfatória, sendo inserido pela CAPES como um serviço dentro do *Hub* de Integração de Dados Nacional.

Figura 2 – Mapeamento das Fontes de Informação



Fonte: Autor.

O *Hub* de Integração de Dados deverá oferecer um balcão de serviços para os demais sistemas que compõem o ecossistema da Educação, Ciência, Tecnologia e Inovação – ECTI e os que forem descobertos na fase de definição de escopo do projeto do novo ecossistema (Figura 3). Tendo em vista o volume de dados que serão trocados entre diversas instituições que compõem o ecossistema de informação, e a necessidade de ser utilizado padrões pré-definidos para correta comunicação entre os

partícipes do *Hub* de Integração de Dados; os conceitos FAIR significa (*Findable Accessible, Interoperable, Reusable*) que visa implementar um conjunto de metadados definidos tanto para uso por mecanismos computacionais automatizados, quanto para uso por pessoas torna-se peça fundamental para o sucesso da integração dos diversos sistemas que compõem o ecossistema. Estes, se forem devidamente adotados viabilizam a interoperabilidade entre diferentes ambientes de dados.

Figura 3 – Balcão de Serviços



Fonte: Diretoria de Avaliação – DAV/CAPES.

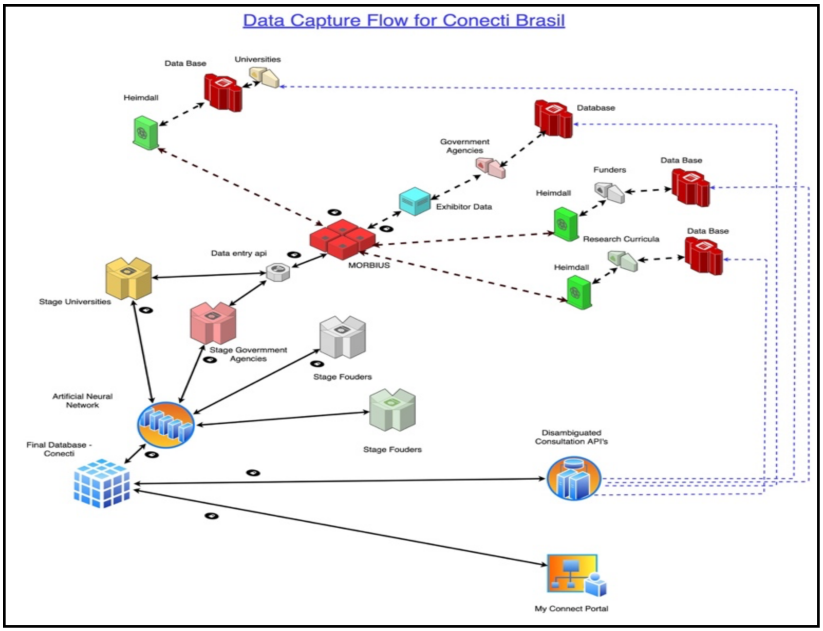
A primeira solução desenvolvida pela equipa foi a criação de um *data display*, denominado FAIS – Expositor, que está a ser implementado nos parques tecnológicos de cada participante e a segunda solução desenvolvida foi uma plataforma de consumo de API, denominada FAIS – Integrador, que consumirá os dados

compartilhados através do Expositor ou através de outra solução proprietária e alimentados pelo ambiente do Hub Nacional de Integração de Dados.

No *Integrator*, é possível traduzir os dados recebidos para o padrão semântico adotado na base de dados do Conecti, seguindo todas as regras dos sistemas CRIS. Essa tradução é configurada apenas uma vez, pelas equipes de TI de ambos os envolvidos. Essas soluções foram desenvolvidas de forma dinâmica o suficiente para serem customizadas de acordo com as necessidades de cada instituição, gerando a documentação das API's geradas, que podem ser escaláveis, aderindo a qualquer protocolo de segurança utilizado no mundo; eles são independentes e podem ou não ser implementados em conjunto (Figura 4).

A solução foi construída para usar poucos recursos computacionais, e também pode ser encapsulada em *docker*, *kubernetes* ou *openshift*, construído em Java 11 – LTS, usando o *framework Spring Boot* 2.2.1.RELEASE, *Spring cloud* 2.2.1.RELEASE, *Microservices with* validação FK, AOP (programação orientada a aspectos), *Service Mesh*, *Hateoas*, *GraphQL*, *Query Search*, *Rancher* v2.3.2, *MongoDB*, *PostgreSQL*, *Oracle*, *Kong*, *Jfrog*, *Gitlab Pipeline*, *Harbor*, *Graylog* 2, *Jwt and Jwe*, *Grafana* and *Prometheus*.

Figura 4 – Arquitetura das Ferramentas de Apoio a Integração de Sistemas – FAIS



Fonte: Autor.

3. Resultados e discussões

Inicialmente testamos as soluções nas bases de dados da Capes, Biblioteca Nacional, Google Scholar, Biblioteca Aberta, BD ISBN, Scopus, ORCID, Google Livros, Lattes e uma Universidade parceira para obter dados do sistema acadêmico.

Preparamos a Solução para suportar 60.000 requisições simultâneas por segundo, com a possibilidade de dobrar a capacidade escalando mais computadores no *cluster*. Ambas as

soluções foram projetadas para se adequar a qualquer fonte de dados existente nos participantes, sejam elas estruturadas ou não. Ambas estão preparadas para receber dados via API, via *upload* de arquivo, entre outros.

Toda a configuração para receber e exibir dados é simplificada objetivando atender a qualquer cenário existente dentro do ecossistema da Educação, Ciência, Tecnologia e Inovação – ECTI.

4. Considerações finais

A solução idealizada visa abstrair a necessidade de se definir modelos de dados a serem adotados pelos diversos atores do ecossistema da Educação, Ciência, Tecnologia e Inovação – ECTI, traduzindo os modelos existentes para o modelo adotado pelo *Hub* Nacional de Integração de Dados, como também para o modelo existente na organização a qual as soluções estiverem instaladas. Entretanto, a necessidade dos atores definirem a semântica de dados por trás de cada metadado compartilhado no ecossistema será o ponto diferencial para uma integração entre as plataformas.

De modo que a não utilização de técnicas eficientes capazes de promover a colaboração entre organizações que desejam trocar informações, mantendo diferentes estruturas internas e processos de negócios variados acarretará a inviabilização do projeto de integração proposto para a utilização dessas soluções como também na construção do *Hub* Nacional de Integração de Dados.

Referências

BRASIL. Presidência da República. **Decreto nº8.777**, de 11 de maio de 2016. Institui a Política de Dados Abertos do Poder Executivo Federal. Diário Oficial da União, 2016. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/D8777.htm. Acesso em: 12 maio 2021.

BRASIL. Ministério da Transparência e Controladoria-Geral da União. **4º Plano de Ação Nacional em Governo Aberto**. Brasília: Open Government Partnership, 2018. Disponível em: http://governoaberto.cgu.gov.br/no-brasil/copy_of_planos-de-acao/4o-plano-de-acao-brasileiro/4o-plano-de-acao-nacional_portugues.pdf. Acesso em: 6 maio 2021.

PAULK, M. C. Extreme Programming from a CMM perspective, **IEEE Software**, v. 18, n. 6, 2001, pp 19-26.

ELIXIR. **Webinar**: BYOD and FAIR data stewardship. Elixir, Netherlands, 2015. Disponível em: http://www.elixir-europe.org/documents/elixir-webinar-byod-and-fair-data-stewardship-august-2015. Acesso em: 25 abr. 2021.

ABU-ELKHEIR, M.; HAYAJNEH, M.; ALI, N. **Data Management for the Internet of Things: Design Primitives and Solution. Sensors**, v.13, n. 11, p.15582-15612, 2013.

DECHAMP, J. F. 2006-2015: celebrating a decade of Open Access policy in the European Commission. In: 4th National

Open Access Workshop. **Anais…** 2015, Ankara, Turkey. Disponível em: http://www.acikerisim.org/wp-content/uploads/2016/01/jean_francois_dechamp.pdf. Acesso em: 02 maio 2021.

FORCE11. Guiding principles for Findable, Accessible, Interoperable and Re-usable data publishing version b1.0. Disponível em: http://www.force11.org/node/6062. Acesso em: 02 maio 2021.

KAPUR, D.; NARENDRAN, P. Matching, unification and complexity. **ACM SIGSAM Bulletin**, v.21, n.4, p.6-9, 1987.

WASSERMAN, A. The extension of data abstraction to database management. **ACM SIGPLAN Notices**, v.16, n.1, pp.198-200, 1981.

WILKINSON, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. **Sci. Data**, v. 3, 2016.

TORINO, E.; TREVISAN, G. L.; CONEGLIAN, C. S.; BOTEGA, L. C.; SANTAREM SEGUNDO, Jose Eduardo; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Enriquecimento semântico para a disponibilização de dados abertos: teoria e prática. **Encontros Bibli**, v. 25, p. 01-19, 2020.

MARIN-ARRAIZA, P.; GONCALVEZ, P. R. V. A.; VIDOTTI, S. A. B. G. Recomendações para a integração de publicações ampliadas em repositórios digitais confiáveis. **Encontros Bibli**, v. 24, p. 1-23, 2019.

TORINO, E.; TREVISAN, G. L.; VIDOTTI, S. A. B. G. Dados abertos CAPES: um olhar à luz dos desafios para publicação de dados na web. **Ciência da Informação**, v. 48, p. 38-46, 2019.