XXII Encontro Nacional de Pesquisa em Ciência da Informação – XXII ENANCIB

ISSN 2177-3688

GT-8 – Informação e Tecnologia

INTEGRAÇÃO DE DADOS DE ECTI ENTRE DIFERENTES SISTEMAS DE INFORMAÇÃO: PROPOSTA DE SOLUÇÃO

INTEGRATION OF ECTI DATA BETWEEN DIFFERENT INFORMATIONS SYSTEMS: SOLUTION PROPOSAL

Harrysson Gilgamesh Medeiros Nobrega. UNESP.

Emanuelle Torino. UTFPR/UNESP.

Silvana Aparecida Borsetti Gregorio Vidotti. UNESP.

Jaqueline Pereira Carvalho Halicki. UNESP.

Talita Moreira de Oliveira. CAPES.

Modalidade: Resumo Expandido

Resumo: Considerando a necessidade de integrar os diferentes sistemas de informação das instituições que compõem o ecossistema brasileiro da Educação, Ciência, Tecnologia e Inovação, foram desenvolvidas soluções tecnológicas para interoperar e interagir com um amplo escopo de sistemas de instituições de ensino, pesquisa e fomento. Com isso, foram desenvolvidas ferramentas tecnológicas customizáveis, capazes de se adaptarem e serem utilizadas em qualquer ecossistema independente das diferentes faixas de dados e maturidade tecnológica. Os testes preliminares mostraram uma solução robusta e satisfatória que pode ser utilizada amplamente, tanto por instituições que gerenciam esses dados em arquivos quanto por instituições que dispõem de sistemas de informações robustos.

Palavras-Chave: Integração entre sistemas. Integração de dados. Ecossistema de ECTI.

Abstract: Considering the need to integrate the different information systems of the institutions that make up the Brazilian ecosystem of Education, Science, Technology and Innovation, technological solutions were developed to interoperate and interact with a scope of systems from educational, research and funding institutions. As a result, customizable technological tools were developed, capable of adapting and being used in any ecosystem, regardless of different data ranges and technological maturity. Preliminary tests showed a robust and satisfactory solution, which can be widely used, both by institutions that manage these data in archives and by institutions that have robust information systems.

Keywords: Integration between systems. Data Integration. ECTI ecosystem.

1 INTRODUÇÃO

No ecossistema brasileiro de Educação, Ciência, Tecnologia e Inovação (ECTI) a coleta de dados armazenados em diferentes sistemas de informação é complexa em função das formas de armazenamento e estrutura dos dados. Esse contexto possui vários atores, como: educadores, pesquisadores, coordenadores de cursos de graduação e pós-graduação, gestores de instituições de ensino, órgãos governamentais e agências de fomento[1].

Vivenciamos um cenário de mudança na forma como são destinados e utilizados os recursos públicos na ECTI brasileira, sejam eles financeiros, físicos, tecnológicos ou humanos. Com isso, há a necessidade de dispor de ferramentas tecnológicas eficientes para mitigar o retrabalho dos atores supramencionados na alimentação de diferentes sistemas de informação, com finalidades específicas e que necessitam de dados armazenados em outros sistemas para fins institucionais, atualização de currículo, avaliação pelo Ministério da Educação (MEC) ou pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), entre outros. Ainda é perceptível um cenário heterogêneo em termos de quantidade, qualidade e confiabilidade de dados disponíveis nos sistemas de informação das diversas instituições que compõem o ecossistema de ECTI.

Dietrich et. al. (2022) afirmam que tecnologias emergentes podem ser utilizadas para fornecer dados aos cidadãos de forma automática, por meio dos preceitos de Dados Abertos, uma vez que grande parte deles são gerados por órgãos públicos. Por outro lado, apontam que tais dados, em geral, não estão disponíveis de modo a favorecer o uso.

Diante do exposto, este trabalho tem como problema de pesquisa: como integrar os sistemas de informação que armazenam dados nas instituições, nas fontes de publicação e disponibilização de resultados de pesquisa que compõem o ecossistema brasileiro de ECTI?

Para tanto, objetivou-se estudar e desenvolver uma solução tecnológica para integrar os dados armazenados nos sistemas de informação utilizados pelas instituições que compõem o ecossistema brasileiro de ECTI, bem como das fontes de publicação e disseminação de resultados de pesquisa, possibilitando integrá-los às plataformas do governo brasileiro. Por fim, este estudo abordou a solução intitulada Ferramentas de Apoio à Integração de Sistemas (FAIS), composta pelo FAIS – Expositor, FAIS – Integrador e Rede Neural.

[1] Quando nos referimos às Agências de Fomento, incluímos as Fundações de Amparo à Pesquisa (FAP).

2 METODOLOGIA

Na busca de um método para orientar a descoberta, o acesso, a integração e a reutilização da vasta quantidade de dados gerados e armazenados no contexto da ECTI foram utilizados os preceitos estabelecidos pelo movimento de dados abertos (TORINO; TREVISAN; VIDOTTI, 2019; TORINO; VIDOTTI, 2021; DIETRICH et. al., 2022; OPEN KNWLEDGE FOUNDATION, 2022), pela transparência pública (BRASIL, 2011) e pelos princípios FAIR (FORCE11, 2020; GOFAIR, 2020).

O estudo de natureza empírica investigou um fenômeno em um contexto com pouca literatura disponível, o que reforça sua relevância para o cenário do ecossistema brasileiro de ECTI e para a Ciência da Informação.

Os antecedentes que marcaram a necessidade de realização desta pesquisa são identificados na própria CAPES que, ao longo dos anos, realizou diferentes iniciativas visando auxiliar a coleta de dados dos Programas de Pós-Graduação (PPGs). Foram utilizados diferentes recursos tecnológicos, dentre eles o INFOCAPES, o DATACAPES, o Coleta e, a partir de 2012, a Plataforma Sucupira.

Em 2019, com o objetivo de identificar pontos passíveis de melhorias na Plataforma Sucupira para torná-la uma plataforma de gestão da informação das Instituições de Ensino Superior Brasileiras (IES), foram selecionadas universidades com base em requisitos tecnológicos e corpo técnico disponível. Isso visa a cocriação de soluções desejáveis, econômicas e tecnologicamente viáveis para desenvolvimento e disponibilização.

A coleta das informações incluiu uma pesquisa realizada no Seminário de Meio Termo de 2019 e, posteriormente, no segundo semestre de 2019 foram realizadas visitas às IES selecionadas. As visitas objetivaram conhecer as formas de coleta, tratamento e armazenamento dos dados, bem como as tecnologias computacionais existentes nas instituições. Para tanto, os representantes das IES participantes foram entrevistados e as técnicas de coleta de dados utilizadas incluem: entrevista em profundidade; Canvas de proposta de valor; teste de usabilidade da Plataforma Sucupira; e facilitação de *design thinking*.

No primeiro semestre de 2020 a pesquisa foi ampliada para outras IES privadas, públicas federais e estaduais, além de agências de fomento [1], utilizando a mesma metodologia

empregada anteriormente, com resultados semelhantes obtidos. Ao todo foram visitadas 26 agências de fomento, 8 instituições de ensino superior, 2 agências de fomento federais, 1 empresa privada, 2 órgãos de governo e 2 empresas sem fins lucrativos.

A pesquisa de campo exploratória visou identificar o nível de maturidade no tratamento e armazenamento dos dados, os padrões tecnológicos, bem como as dificuldades que as instituições encontram nesse processo. Ainda possibilitou identificar os modelos de dados adotados e mapear as similaridades, necessidades e expectativas.

Dentre as dificuldades encontradas destacam-se a necessidade de obtenção de dados certificados e o retrabalho nas prestações de contas para as agências de fomento.

O mapeamento subsidiou o desenvolvimento das Ferramentas de Apoio à Integração de Sistemas (FAIS) realizado pela equipe do Hub Nacional de Integração de Dados da CAPES em parceria com uma equipe da Unesp que foi designada para atuar no projeto. Esse grupo de trabalho iniciou as atividades em março de 2020, quando começaram a ser realizados experimentos em caráter de testes no ambiente digital da Unesp. O objetivo era identificar as principais necessidades para a criação de um conjunto de ferramentas de integração de dados com capacidade de uso por qualquer instituição brasileira.

3 RESULTADOS E DISCUSSÕES

A pesquisa de campo exploratória possibilitou identificar, nas IES participantes, diversos cenários tecnológicos. Algumas IES possuem setores de Tecnologia da Informação (TI) bem definidos e estruturados, com sistemas acadêmicos implantados e equipe suficiente para colaborar com novos projetos; enquanto outras IES não possuem equipe e setor de TI definido, orçamento para investimento em TI e sistema de gerenciamento implantado.

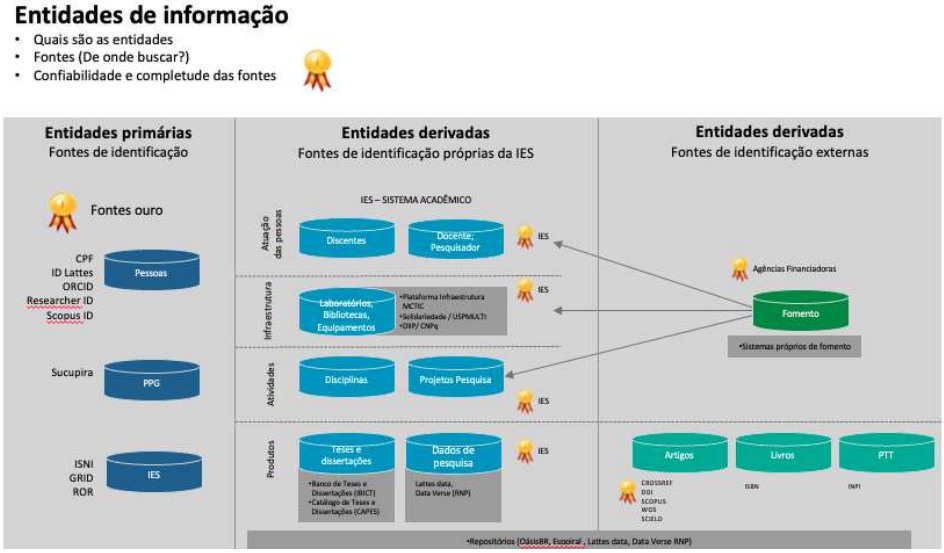
Poucas agências de fomento tinham estrutura necessária para iniciar a colaboração em um projeto de integração de dados. Das 26 agências de fomento visitadas, apenas 4 possuíam maturidade tecnológica e pessoal disponível para a participação projeto.

Não foram identificados padrões de dados e ontologias nas IES e tampouco conceitos de semântica e ontologias nas agências de fomento. Em muitos casos, as dificuldades se expandem para os conceitos, como, por exemplo, chegar a um consenso sobre o que é bolsa de pesquisa e auxílio.

Os dados coletados na pesquisa de campo possibilitaram mapear as fontes de armazenamento de dados existentes em cada instituição, sua origem e a possibilidade de integração, conforme apresentado na Figura 1.

Considerando a diversidade de sistemas de informação, modelos de dados e tecnologias utilizadas pelas instituições brasileiras que compõem o ecossistema da ECTI, a solução Ferramentas de Apoio à Integração de Sistemas (FAIS) foi desenvolvida como um produto customizável, capaz de se adaptar e ser utilizado em qualquer instituição independente das tecnologias disponíveis, bem como das diferentes faixas de dados e maturidade no tratamento e armazenamento de dados.

Figura 1 – Mapeamento das Fontes de Armazenamento de Dados



Fonte: Autoria própria.

Os testes preliminares mostraram uma solução robusta e satisfatória, o que culminou na inclusão das FAIS como um serviço do Hub Nacional de Integração de Dados da CAPES. Com isso, o Hub Nacional de Integração de Dados oferecerá serviços para as demais instituições que compõem o ecossistema brasileiro de ECTI, conforme explicitado na Figura 2.

Figura 2 – Hub Nacional de Integração de Dados

Fonte: Diretoria de Avaliação (DAV) da CAPES.

Os princípios FAIR tornam-se fundamentais para o sucesso da integração dos sistemas de informação que compõem esse ecossistema, tendo em vista a diversidade e o volume de dados que podem ser interoperáveis entre as diversas instituições do ecossistema brasileiro de ECTI e a necessidade de serem utilizados padrões pré-definidos para a comunicação de dados entre as instituições participantes do Hub de Integração de Dados. Assim, a adoção dos princípios FAIR (FORCE11, 2020; GOFAIR, 2020) viabilizam a interoperabilidade entre as diferentes fontes de dados, objetivo principal deste trabalho.

Segundo Torino, Coneglian e Vidotti (2020, p. 16),

> Os princípios FAIR possibilitam identificar pontos de convergência no tratamento dos dados, metadados e infraestruturas visando maximizar sua localização, acesso e uso. Dessa forma, é necessário que todos os envolvidos – desde o planejamento da pesquisa, a coleta, o tratamento e o armazenamento dos dados – tenham em mente a adoção dos princípios, considerando as especificidades dos dados e do domínio.

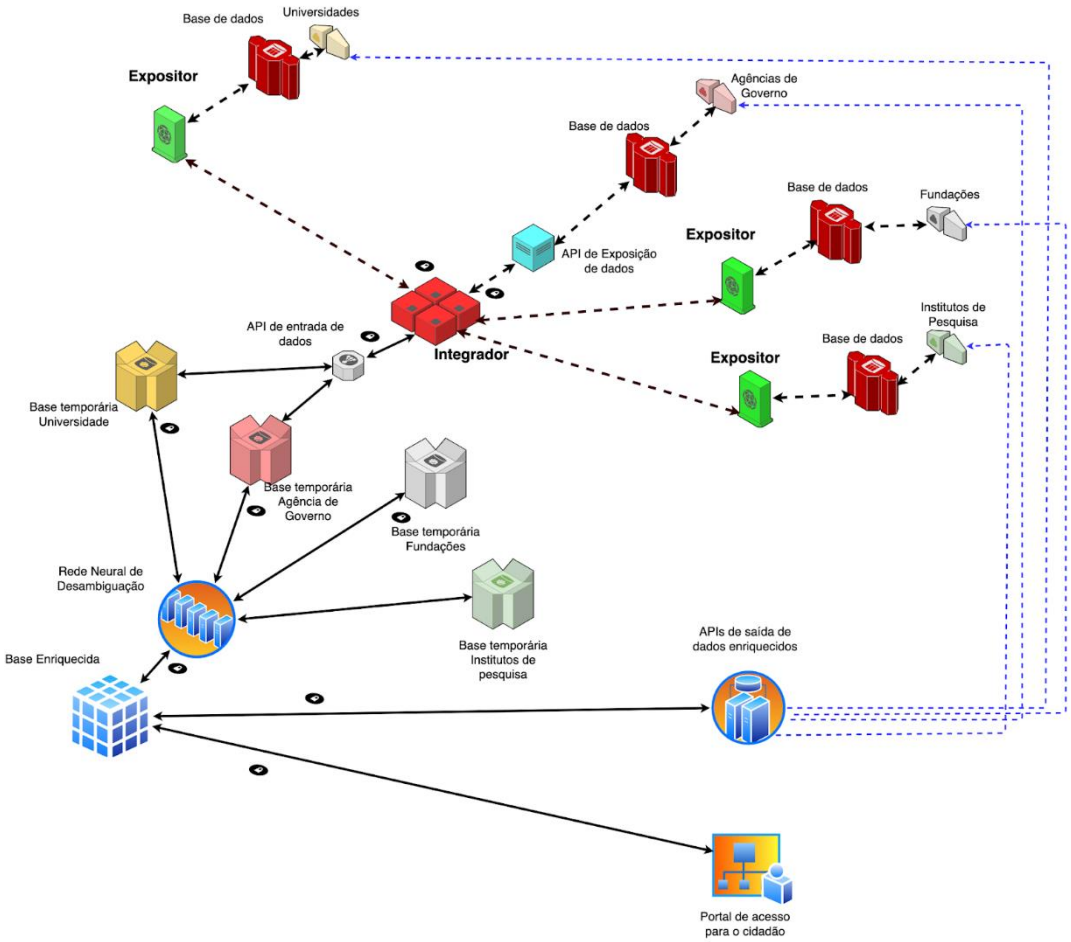
A arquitetura das Ferramentas de Apoio à Integração dos Sistemas (FAIS) foi desenvolvida após análise dos dados coletados na pesquisa de campo exploratória. A primeira solução desenvolvida foi um *data display*, denominado **FAIS - Expositor**, que pode ser implementado nos parques tecnológicos de cada instituição participante e a segunda solução desenvolvida foi uma plataforma de consumo de *Application Programming Interface* (API), denominada **FAIS - Integrador**, que consumirá os dados compartilhados pelo FAIS - Expositor ou por outra ferramenta tecnológica - a exemplo de protocolos e *web services* -, e alimentará o ambiente informacional digital do Hub Nacional de Integração de Dados (Figura 3).

Além disso, no FAIS - Integrator é possível traduzir os dados recebidos para o padrão semântico adotado na base de dados de destino, seguindo todas as regras adotadas pela equipe responsável pela coleta dos dados. Essa tradução é configurada apenas uma vez pelas equipes de tecnologia da informação (TI) de ambos os envolvidos.

Foi idealizada uma solução capaz de se conectar em uma fonte de dados com modelos relacionais para atender a heterogeneidade dos cenários encontrados e, rapidamente, ser capaz de expor as tabelas de dados dessa fonte em forma de API Rest. Essa solução designada como FAIS - Expositor é alinhada aos modernos padrões existentes e aderente aos padrões de segurança atuais.

De acordo com Dietrich et. al. (2022), no contexto dos dados abertos, os dados podem ser disponibilizados por API, que permitem estabelecer quais dados serão disponibilizados a partir de um banco de dados atualizado em tempo real.

Figura 3 – Arquitetura das Ferramentas de Apoio à Integração de Sistemas (FAIS)



Fonte: Autoria própria.

Havia, ainda, a necessidade de capturar dados de instituições que não possuem fonte de dados estruturados, mas que possuem dados relevantes a serem capturados e que também não possuem estrutura computacional que permita uma integração. Também foi necessário encontrar uma forma de capturar dados em páginas web, cuja fonte de dados está apenas em portais e sem forma de exposição, como, por exemplo, a base da Biblioteca Nacional Brasileira.

Desse modo, foi desenvolvida a solução FAIS - Integrator para coleta de dados por meio da importação de planilhas e de arquivos de texto. Essa solução é capaz de capturar informações de páginas web, o que possibilita a realização de buscas e coletas de forma automatizada e periódica. Além disso, a solução precisava ser capaz de traduzir os modelos de dados semânticos coletados para o modelo de dados adotados no receptor.

Cada conjunto de metadados coletado utilizando o FAIS - Integrador é gravado em um banco de dados relacional temporário, que apenas o fornecedor dos dados tem acesso por meio de uma chave de acesso única para alterar os dados armazenados. Com isso, tem-se bancos de dados separados para cada parceiro, que se configura como uma fonte de dados.

Ao observar uma amostra desses dados coletados percebeu-se que seria necessário uma higienização e o cruzamento de dados entre essas bases, objetivando atender às políticas públicas vigentes, como, por exemplo, a necessidade de identificar pessoas de forma única entre as bases coletadas.

Para solucionar esse problema foi desenvolvida uma solução de inteligência artificial composta de 13 algoritmos rodando em pilha, capaz de identificar o tipo de metadado que está sendo trafegado e, de forma autônoma, decidir o melhor algoritmo a ser utilizado em cada conjunto de metadados. O conjunto de dados resultantes dessa desambiguação é salvo em uma base de dados enriquecida e dessa base são disponibilizados API's de consulta para que os parceiros possam obter dados mais limpos e completos para seus sistemas.

Também nessa base final foi planejada uma forma única para que os usuários finais tenham acesso aos dados pessoais trafegados entre essas instituições. Desse modo, eles passam a ter controle sobre a disponibilização ou não dos próprios dados nesse banco, em cumprimento à Lei Geral de Proteção de Dados Pessoais (BRASIL, 2019).

Em suma, foi construída uma solução capaz de expor dados de instituições sem tecnologia para tal; uma outra solução para captura de dados em qualquer tipo de base ou

arquivo estruturado; um conversor de modelo de dados e semântica; uma rede neural com aprendizado supervisionado destinada a higienização e desambiguação de dados; e uma forma de visualização centralizada para o cidadão averiguar e autorizar o uso de dados pessoais entre as instituições.

Essas soluções foram desenvolvidas de forma dinâmica o suficiente para serem customizadas de acordo com as necessidades de cada instituição. É possível gerar a documentação das API's que podem ser escaláveis para aderir a qualquer protocolo de segurança utilizado. Destaca-se, ainda, que as soluções são independentes e podem ou não ser implementadas em conjunto.

As soluções foram construídas para usar poucos recursos computacionais, e também podem ser encapsuladas 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 validation 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*.

Inicialmente as soluções FAIS - Expositor e FAIS - Integrador foram testadas para coletar dados das 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. As soluções foram preparadas para suportar 60 mil requisições simultâneas por segundo, com a possibilidade de dobrar a capacidade escalando mais computadores no cluster.

As soluções FAIS - Expositor e FAIS - Integrador foram projetadas para se adequar a qualquer fonte de dados existente nos participantes, sejam os dados estruturados ou não estruturados, que podem ser recebidos, por exemplo, via API ou *upload* de arquivos. Toda a configuração para receber e exibir dados é simplificada objetivando atender a qualquer cenário existente no ecossistema brasileiro de Educação, Ciência, Tecnologia e Inovação (ECTI).

4 CONSIDERAÇÕES FINAIS

As Ferramentas de Apoio a Interoperabilidade de Sistemas (FAIS) idealizadas visam abstrair a necessidade de se definir modelos de dados a serem adotados pelos diversos atores

do ecossistema brasileiro de Educação, Ciência, Tecnologia e Inovação (ECTI), possibilitando a conversão dos modelos existentes para o modelo de dados adotado pelo Hub Nacional de Integração de Dados, bem como para o modelo de dados existente na organização na qual as soluções estiverem instaladas. Entretanto, a necessidade de os atores definirem a semântica de dados no ecossistema será o ponto diferencial para a integração entre as plataformas.

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 e a construção do Hub Nacional de Integração de Dados.

O objetivo do serviço ou do desenvolvimento destas tecnologias é otimizar a obtenção dos dados, gerando mais agilidade e eficiência para a instituição e para os próprios envolvidos, além de simplificar o trabalho entre os envolvidos: instituição, governo ou cidadão.

Financiamento: Coordenadoria de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERÊNCIAS

BRASIL. Lei nº 12.527, de 18 de novembro de 2011. **Diário Oficial da União**, Brasília, DF, 18 nov. 2011. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 1 maio 2018.

BRASIL. **Lei nº 13.709**, de 14 de agosto de 2018. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 1 maio 2022.

DIETRICH, Daniel; GRAY, Jonathan; McNAMARA, Tim; POIKOLA, Antti; POLLOCK, Rufus; TAIT, Julian; ZIJLSTRA, Ton**. Guia de dados abertos**. Disponível em: http://opendatahandbook.org/guide/pt_BR/. Acesso em: 20 jul. 2022.

FORCE21. **Guiding principles for findable, accessible, interoperable and re-usable data publishing version b1.0**. Disponível em: https://www.force11.org/fairprinciples#Annex1-1. Acesso em: 07 abr. 2020.

GOFAIR. **FAIR principles**. Disponível em: https://www.go-fair.org/fair-principles/. Acesso em: 07 abr. 2020.

OPEN KNOWLEDGE FOUNDATION. **A fair, free and open future**. Disponível em: https://www.go-fair.org/fair-principles/. Acesso em: 07 abr. 2022.

TORINO, Emanuelle; CONEGLIAN, Caio Saraiva; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Estruturas de representação para reuso de dados no contexto da ecologia de pesquisa: CRIS Institucional. **Informação & Informação**, Londrina, v. 25, n. 3, p. p. 1-27, jul./set. 2020. DOI:

http://dx.doi.org/10.5433/1981-8920.2020v25n3p1. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/41946. Acesso em: 24 nov. 2020.

TORINO, Emanuelle; TREVISAN, Gustavo Lunardelli; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Dados abertos CAPES: um olhar à luz dos desafios para publicação de dados na web. **Ciência da Informação**, Brasília, v. 48, n. 3, p. 38-46, 2019. Disponível em: https://revista.ibict.br/ciinf/article/view/4866. Acesso em: 10 jun. 2022.

TORINO, Emanuelle; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Boas práticas para dados na web: análise do portal Dados Abertos Capes. **Informação & Sociedade: Estudos**, João Pessoa, v. 31, n. 1, p. 1-25, 2021. DOI: 10.22478/ufpb.1809-4783.2021v31n1.50790. Disponível em: https://periodicos.ufpb.br/index.php/ies/article/view/50790. Acesso em: 10 jun. 2022.