ISSN 1678-765X doi 10.20396/rdbci.v21i00.8673171

Artigo

RDBCI
Revista Digital de Biblioteconomia e Ciência da Infor
Digital Journal of Library and Information Science

BrCris: desenvolvimento de ferramentas no tratamento, análise e disseminação da informação em apoio à Ciência Aberta no Brasil

Vivian dos Santos Silva [1] Jesiel Viana da Silva [2] Thiago Magela Rodrigues Dias[3] Renê Faustino Gabriel Junior[4] Washington Luís Ribeiro de Carvalho Segundo[5]

RESUMO

Introdução: Os sistemas CRIS constituem-se em sistemas de informação abrangentes sobre todo o ecossistema do processo científico. O BrCris tem como propósito integrar e organizar informações referentes a atividades de pesquisa, projetos, publicações, pesquisadores, instituições, financiamentos e outros dados relevantes no contexto científico brasileiro. Objetivo: Este estudo visa discutir o processamento dos dados na Plataforma BrCris e analisar as ferramentas computacionais empregadas para essa finalidade, explorando três abordagens principais: integração e consistência dos dados, visualização e validação, além da certificação dos dados. Metodologia: O estudo se configura como descritivo, apresentando em detalhes as etapas de tratamento de dados do BrCris, discutindo os desafios encontrados no manuseio de grandes volumes de informações. Além disso, descreve o ferramental computacional utilizado para o tratamento das informações científicas e tecnológicas. Resultados: O estudo revela os procedimentos para o tratamento de dados e as ferramentas computacionais desenvolvidas para os sistemas informacionais, bem como a integração e análise dos dados obtidos. São apresentados os resultados do tratamento e modelagem das informações baseadas no VIVO, com dados e painéis gráficos, e as oportunidades de reutilização dos dados gerados. Também é detalhada a integração dos dados em um repositório autodeclarado (Lattes) e no repositório agregador de teses e dissertações (Oasisbr), culminando na emissão de um selo de certificação. Conclusão: Os resultados evidenciam que a adoção dessas ferramentas computacionais proporciona um acesso facilitado e ágil a um extenso conjunto de informações consolidadas, previamente dispersas em várias fontes, especialmente devido à diversidade de repositórios e limitações de acesso individualizados. Assim, este estudo apresenta um conjunto de ferramentas computacionais cujas funcionalidades estão alinhadas com as diretrizes da Ciência Aberta no Brasil.

PALAVRAS-CHAVE

Ciência aberta. Repositórios de dados. Tratamento da informação. BrCris. Dados científicos.

Correspondência dos autores

[1] Universität Passau, Passau - Alemanha
vivian.ss@gmail.com

[2] Instituto Federal do Piauí
Teresina, PI - Brasil
jesiel.analista@gmail.com

[3] Centro Federal de Educação Tecnológica de Minas Gerais,
Belo Horizonte, MG - Brasil
thiagomagela@gmail.com

[4] Universidade Federal do Rio Grande do Sul
Porto Alegre, RS - Brasil
renefgj@gmail.com

[5] Instituto Brasileiro de Informação em Ciência e Tecnologia
Brasília, DF - Brasil
washingtonsegundo@ibict.br

BrCris: tools for treatment, analysis, and dissemination of scientific information in support of Open Science in Brazil

ABSTRACT

Introduction: CRIS systems constitute comprehensive information systems over the entire ecosystem of the scientific process. BrCris aims to integrate and organize information regarding research activities, projects, publications, researchers, institutions, financing and other relevant data in the Brazilian scientific context. Objective: This study aims to discuss data processing on the BrCris Platform and analyze the computational tools used for this purpose,

exploring three main approaches: data integration and consistency, visualization and validation, in addition to data certification. Methodology: The study is descriptive, presenting in detail the BrCris data processing steps, discussing the challenges encountered in handling large volumes of information. Furthermore, it describes the computational tools used to process scientific and technological information. Results: The study reveals the procedures for data processing and the computational tools developed for information systems, as well as the integration and analysis of the data obtained. The results of the processing and modeling of information based on VIVO, with data and graphic panels, and the opportunities for reusing the generated data are presented. The integration of data into a self-declared repository (Lattes) and the theses and dissertations aggregator repository (Oasisbr) is also detailed, culminating in the issuance of a certification seal. Conclusion: The results show that the adoption of these computational tools provides easy and agile access to an extensive set of consolidated information, previously dispersed across various sources, especially due to the diversity of repositories and individual access limitations. Thus, this study presents a set of computational tools whose functionalities are aligned with the guidelines of Open Science in Brazil.

KEYWORDS

Open science. Data repositories. Information processing. BrCris. Scientific data.

CRediT

- Reconhecimentos: Não aplicável.
- Financiamento: Este estudo foi parcialmente financiado pelas agências brasileiras Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF) - processo 00193-00000788/2021-31; Conselho Nacional de Desenvolvimento Científico e Tecnológico(CNPq) - processo 400038/2023-4; Financiadora de Estudos e Projetos (FINEP) - convênio 01.16.0051-00.
- Conflitos de interesse: Os autores certificam que não têm interesse comercial ou associativo que represente um conflito de interesses em relação ao manuscrito.
- Aprovação ética: Não aplicável.
- Disponibilidade de dados e material: Os conjuntos de dados gerados e/ou analisados durante o presente estudo estão disponíveis no Repositório de Dados científicos Zenodo.
- Contribuições dos autores: Conceitualização, Análise formal, Investigação, Metodologia, Software, Visualização; Escrita – rascunho original: SILVA, V. S.; Curadoria de dados, Análise formal, Investigação, Software, Visualização; Escrita – rascunho original: VIANA, J.; Conceitualização, Curadoria de dados, Análise formal, Investigação, Metodologia, Administração do projeto, Supervisão, Validação; Escrita – rascunho original; Escrita – revisão & edição: DIAS, T. M. R.; Curadoria de dados, Análise formal, Investigação, Metodologia, Software, Validação; Escrita – rascunho original; Escrita – revisão & edição: GABRIEL-JUNIOR, R. F.; Conceitualização, Análise formal, Aquisição de financiamento, Investigação, Metodologia, Administração do projeto, Recursos; Escrita – rascunho original: CARVALHO-SEGUNDO, W. L. R.

JITA: IN. Open science.
ODS: 9. Indústria, Inovação e Infraestrutura

Artigo submetido ao sistema de similaridade

Submetido em: 21/03/2023 – Aceito em: 23/11/2023 – Publicado em: 17/12/2023

Editor: Gildenir Carolino Santos

1 INTRODUÇÃO

A humanidade demonstra ser capaz de enfrentar os mais importantes desafios, em diferentes contextos, que impactam o modo de vida e impedem seu desenvolvimento. Contemporaneamente, assistimos ao surgimento de tratamentos, drogas e vacinas para as mais diversas enfermidades, novos processos de produção de alimentos que garantem a perpetuação da espécie e evoluções tecnológicas que transformam a vida cotidiana das pessoas. Estes avanços possuem uma origem comum, ou seja, são resultados da aplicação prática de alguma descoberta científica teórica que, por sua vez, derivam do esforço de pesquisa que é empreendido por pesquisadores nas mais diversas áreas do conhecimento (TANG *et al*., 2008).

A produção do conhecimento científico é um processo que leva tempo e é incremental (Huang; Glänzel; Zhang, 2021). Pesquisadores buscam as bases para suas pesquisas no desenvolvimento observado da sua área do conhecimento (estado da arte). A identificação do estado da arte de uma determinada área pode ser feita principalmente pela análise de publicações que concentram as informações a respeito daquele tópico do conhecimento.

Pesquisadores que fazem seu trabalho de forma colaborativa, comumente, publicam os resultados de suas pesquisas da mesma forma. A identificação das relações de coautoria em trabalhos científicos/tecnológicos e projetos de pesquisa possibilita a estruturação de redes de pesquisadores e instituições que são interligados por suas produções acadêmicas (Abbasi; Altmann; Hossain, 2011).

Atualmente, tudo que se conhece sobre o surgimento e o desenvolvimento das disciplinas, a difusão do conhecimento e a evolução da ciência e tecnologia é resultado, predominantemente, da análise de publicações científicas (De Meis *et al*. 2003; Leta; Glänzel; Thijs, 2006), da colaboração científica (Yoshikane; Kageura, 2004) e de registros de patentes (Abbas; Zhang; Khan, 2014).

O Brasil tem uma parcela relevante na produção científica internacional principalmente em nichos específicos ao atuar em sua economia. O país é líder na produção do conhecimento no contexto de América Latina (Collazo-Reyes, 2014) e um atrator de talentos no contexto regional (Saraiva; Miranda, 2004). Destaca-se pela sua implementação de plataformas digitais de registro nacional do atuar de seus pesquisadores.

O destaque internacional é a Plataforma Lattes, em que se observa a importância estratégica de ter as informações científicas curriculares disponíveis de forma ampla (LANE, 2010). Similar à Plataforma Lattes existem outras plataformas nacionais que registram parte do atuar acadêmico e tecnológicos, como o Sucupira da CAPES, o Banco de teses e dissertações do Ibict, o Banco de propriedade industrial do INPI e Portal Transparência do Governo Federal, por exemplo. Embora todas as informações dessas fontes de dados sejam abertas e livres para consulta, elas estão restritas a suas bases de dados, não existindo uma integração de forma a possibilitar a exploração dessas informações de forma ampla a ciência e a sociedade do Brasil.

Neste contexto, é importante destacar que a informação proveniente de pesquisas em Ciência, Tecnologia e Inovação (CT&I) representa um dos principais pilares para o desenvolvimento econômico e social de um país. Essa informação constitui-se como um elemento orientador fundamental para a geração sistemática de conhecimento, tanto teórico quanto aplicado.

A informação gerada no âmbito de Ciência, Tecnologia e Inovação (CT&I) é altamente especializada, distinguindo-se de outros tipos de informação. Sua produção é fundamentada em um método específico - o método científico - e a divulgação de seus resultados segue procedimentos distintos, abrangendo avaliação, validação, publicação e acesso por meio de fontes especializadas. (Meadows, 1999).

A diversidade de fontes de informação, juntamente com a variedade de seus modelos de dados (metadados), resultante das pesquisas em Ciência, Tecnologia e de Inovação (CT&I) são frequentemente armazenados em bases de dados e repositórios que tem estruturas de dados

distintas, visando assegurar a preservação, visibilidade e recuperação eficaz das informações contidas, e em poucos casos objetivam a integração.

Complementando as fontes de informações brasileiras, existem fontes com abrangência internacional, que possibilitam a troca de dados tanto para seres humanos como para computadores. Dentre essas bases de dados pode-se citar o Wikidata, o Crossref, OpenCitations, OpenAIRE Research Graph, o Latindex e o DOAJ entre outras de fontes abertas de informação em CT&I.

Reunir e integrar os dados dessas e de outras fontes é um desafio que exige um grande esforço intelectual e um grande poder computacional. A capacidade de armazenamento necessária ultrapassa a que um sistema gerenciador de banco de dados tradicional consegue suportar, exigindo soluções tecnológicas avançadas, provenientes da chamada Web 4.0, bem como, de análises que envolvam técnicas de inteligência computacional.

A partir desse cenário, surgiram iniciativas voltadas para a criação de sistemas destinados a gerenciar a produção acadêmica, sejam eles institucionais, nacionais ou temáticos. Esses sistemas são conhecidos pela sigla CRIS (Current Research Information Systems). Seu objetivo principal é integrar informações provenientes de diversas bases de dados, fornecendo relatórios, indicadores e dados consolidados para gestores, pesquisadores e demais usuários, permitindo a análise da produção em seus respectivos países ou áreas de conhecimento.

Os sistemas CRIS são facilitadores na promoção da Ciência Aberta ao tornar visíveis e acessíveis dados relacionados a projetos, resultados, publicações, patentes, grupos de pesquisa, pesquisadores e instituições. Isso não apenas facilita a interoperabilidade entre diferentes sistemas, mas também ressoa com o cerne da Ciência Aberta, que enfatiza a colaboração e a partilha de recursos. Assim, os CRIS podem ser ajustados para fomentar a interconexão entre pesquisadores, permitindo a identificação de áreas de interesse comum, potencializado colaborações e o intercâmbio de informações (Singh *et al*., 2021).

Neste contexto o problema de pesquisa busca relatar quais foram os desafios enfrentados e as estratégias adotadas durante a construção do BrCris e no desenvolvimento das ferramentas computacionais destinadas ao tratamento, análise e divulgação da informação científica brasileira, e quais foram os resultados e impactos alcançados por essas iniciativas?

Desta forma a pesquisa tem como objetivo discutir o tratamento e processamento dos dados da Plataforma BrCris, bem como as ferramentas computacionais utilizadas para essa finalidade. De forma mais específica, o objetivo geral da pesquisa se subdivide em discutir a integração e consistência dos dados para o BrCris; validar os dados por meio de visualizados dos dados; e por fim, discutir os dados gerados e a certificação dos dados.

2 REFERENCIAL TEÓRICO

O ecossistema da pesquisa abrange a participação de múltiplos intervenientes que interagem entre si. Isso compreende desde a obtenção de financiamento por meio de um projeto de pesquisa, indo além para incluir o papel fundamental do pesquisador, que utiliza infraestrutura, como laboratórios e equipamentos físicos, para conduzir suas investigações (Lee; Bozeman, 2005). Por sus vez, os pesquisadores estão vinculados a instituições onde conduzem suas pesquisas com a geração de conhecimento, frequentemente documentado em artigos científicos e ou relatórios técnicos disponibilizados em bases de dados (Singh et al., 2021) ou repositórios de pesquisa.

O modelo CRIS define um sistema de informação sobre todo o ecossistema do processo científico. São organizadas em um só lugar todas as informações do ciclo da pesquisa Científica, desde o Fomento, passando pelos projetos, pesquisadores, instituições de pesquisa e laboratórios, até os outputs de uma pesquisa científica, tais como artigos científicos, teses, dissertações, livros, capítulos de livro, patentes e conjuntos de dados científicos (Sivertsen, 2019).

Para Joint (2008) o CRIS pode ser definido como um sistema de informação capaz de gerenciar toda a informação relevante de pesquisa, desde o estágio inicial de identificação de oportunidades de financiamento, passando pela redação e submissão de propostas, acompanhando as propostas bem-sucedidas que se transformam em projetos ativos, os quais são gerenciados até sua conclusão. Neste ponto, são produzidos resultados, incluindo várias publicações ou outros artefatos relacionados à atividade de pesquisa.

Torino, Coneglian e Vidotti (2020) ressaltam que para a constituição de um CRIS institucional é necessária a integração das estruturas de representação institucionais, conhecer os sistemas de informação que as armazenam, as formas de exibição dos dados, os protocolos de comunicação disponíveis, para definir a estrutura de conversão.

Iniciativas de mapeamento do ecossistema ocorrem na Europa com o Directory of Research Information Systems (DRIS) executado pela euroCRIS (Eurocris , 2023), No Brasil, o BrCris é o mapeamento do ecossistema de informação da pesquisa científica brasileira. Sua concretude se faz com uma plataforma agregadora de diversas fontes de informações, o que permite recuperar, certificar e visualizar dados e informações relativas aos diversos atores que atuam na pesquisa científica do contexto brasileiro. Dentre as principais fontes estão os dados curriculares de indivíduos, organizações, programas de pós-graduação, publicações, orientações acadêmicas, revistas científicas, patentes, grupos de pesquisa, softwares, outras fontes ainda serão agregadas (Dias; *et al.*, 2022).

O BrCris oferece uma interface unificada de busca de informações, a visualização de redes de colaboração e painéis de indicadores em ciência, tecnologia e inovação. E estabelece um modelo único de organização da informação científica de todo o ecossistema da pesquisa brasileira (Dias; et al., 2022). Entre os agentes deste ecossistema estão os pesquisadores, os projetos, infraestruturas, laboratórios e instituições de pesquisa, os financiadores, além dos resultados da pesquisa expressos principalmente por publicações científicas, teses, dissertações, conjuntos de dados científicos e patentes (Kong, et al., 2019).

Neste contexto, a idealização do projeto do Sistema BrCris (Pinto *et al.,* 2021), que é o CRIS no contexto da Ciência Aberta Brasileira, concebida em 2014, quando inspirado no modelo proposto por Portugal de um CRIS nacional (o PTCRIS - https://ptcris.pt), o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) iniciou uma sequência de estudos e parcerias interinstitucionais para a execução do projeto. Em 2020, houve a implementação formal do projeto de pesquisa para a construção do BrCris. O objeto do projeto era disponibilizar ferramentas tecnológicas visando munir dados consolidados científicos e tecnológicos brasileiros para toda a comunidade acadêmica.

Para essa disponibilização, o BrCris adota um esquema de representação de dados em dois níveis. O primeiro é o nível lógico, materializado como um modelo de entidades e relacionamentos baseado no CERIF (Jörg, 2010), também chamado de metamodelo. Este modelo é traduzido para um esquema relacional físico na plataforma LA Referencia, onde os dados são carregados e processados. Trata-se, portanto, de um modelo que atende às necessidades internas de armazenamento de dados, sendo responsável pela organização e integração das informações coletadas para que se tornem insumos para os objetivos do projeto (Pinto *et al*., 2021).

Para representar o domínio acadêmico e científico, pode-se utilizar a ontologia VIVO-ISF (*Integrated Semantic Framework*), utilizada pela plataforma VIVO. A ontologia VIVO-ISF é baseada na ontologia de alto nível Basic Formal Ontology (BFO), que fornece uma base conceitual bem fundamentada. A ontologia também permite extensões que incorporam características institucionais locais (Rathke; Rocha, 2019). A ontologia VIVO-ISF foi elaborada integrando várias outras ontologias e vocabulários, o que a torna uma ferramenta compreensível para diversas outras plataformas. Essa ontologia é versátil e pode ser empregada em várias aplicações, pois representa o domínio das informações acadêmicas, abrangendo aspectos como publicações, ensino, orientação e outras áreas pertinentes. Vale destacar ainda

que o uso de outras ontologias e vocabulários facilita a ligação com as bases de dados de linked data, por haver uma maior compatibilização entre os recursos com as demais bases (Lyrasis, 2016).

Sendo um modelo que descreve o domínio de pesquisa acadêmica, a ontologia VIVO-ISF é composta por classes e propriedades que representam uma rede de pesquisadores, as instituições e projetos aos quais estão vinculados, e as publicações, patentes, softwares e eventuais outros produtos de suas pesquisas. Sua principal vantagem é a reutilização de outras ontologias já bem estabelecidas, como a *Bibliographic Ontology* (BIBO), a *Event Ontology* (EO), a *Friend of a Friend* (FOAF), a *Geopolitical Ontology* (GEO), a *Software Ontology* (SWO), a *Simple Knowledge Organization System* (SKOS) e a *vCard,*entre outras. Destaca-se também a integração da *Basic Formal Ontology* (BFO), uma ontologia de fundamentação que fornece uma sólida base conceitual para as classes e propriedades do modelo.

O modelo semântico do BrCris é composto por um subconjunto da ontologia VIVO, representados pelas classes e propriedades equivalentes às entidades, atributos e relacionamentos do metamodelo lógico, acrescido de uma extensão local que cobre informações específicas do contexto brasileiro. Para cada entidade do modelo, é atribuído um identificador único, ou seja, cada entidade dentro do ecossistema de pesquisa possui um link permanente na web, garantindo a ausência de ambiguidades.

A utilização de um modelo semântico baseado em ontologia permite a representação dos dados como um grafo de conhecimento, o que permite a publicação e o consumo destes dados como *Linked Open Data* (LOD). Bauer e Kaltenböck (2011) ressaltam que, para que possamos realmente tirar proveito de dados abertos, é crucial colocar informação e dados em contexto, criando novo conhecimento que alimenta serviços e aplicações eficientes. Também acrescentam que, por ser um importante mecanismo de integração e gerenciamento de informação, a disponibilização de LOD facilita a inovação e multiplicação do conhecimento a partir dos dados interligados, o que vai ao encontro dos princípios e metas do BrCris e de plataformas CRIS de forma geral.

A publicação de dados como LOD constitui-se como uma boa prática de compartilhamento de dados na Ciência Aberta, principalmente pelo fato de permitir a rastreabilidade das informações disponibilizadas. Isso é particularmente importante no contexto do BrCris devido ao volume e diversidade das fontes de dados. Gerando dados interligados é possível, por exemplo, vincular entidades do tipo "Person" com seus perfis nas plataformas Lattes ou Orcid; "*OrgUnits*" com seus registros no *Research Organization Registry* (ROR) ou na *Global Research Identifier Database* (GRID); "*Publications*" com suas entradas na BDTD, no Oasisbr, ou em qualquer repositório onde estejam identificadas por seu DOI; e assim por diante.

2 METODOLOGIA

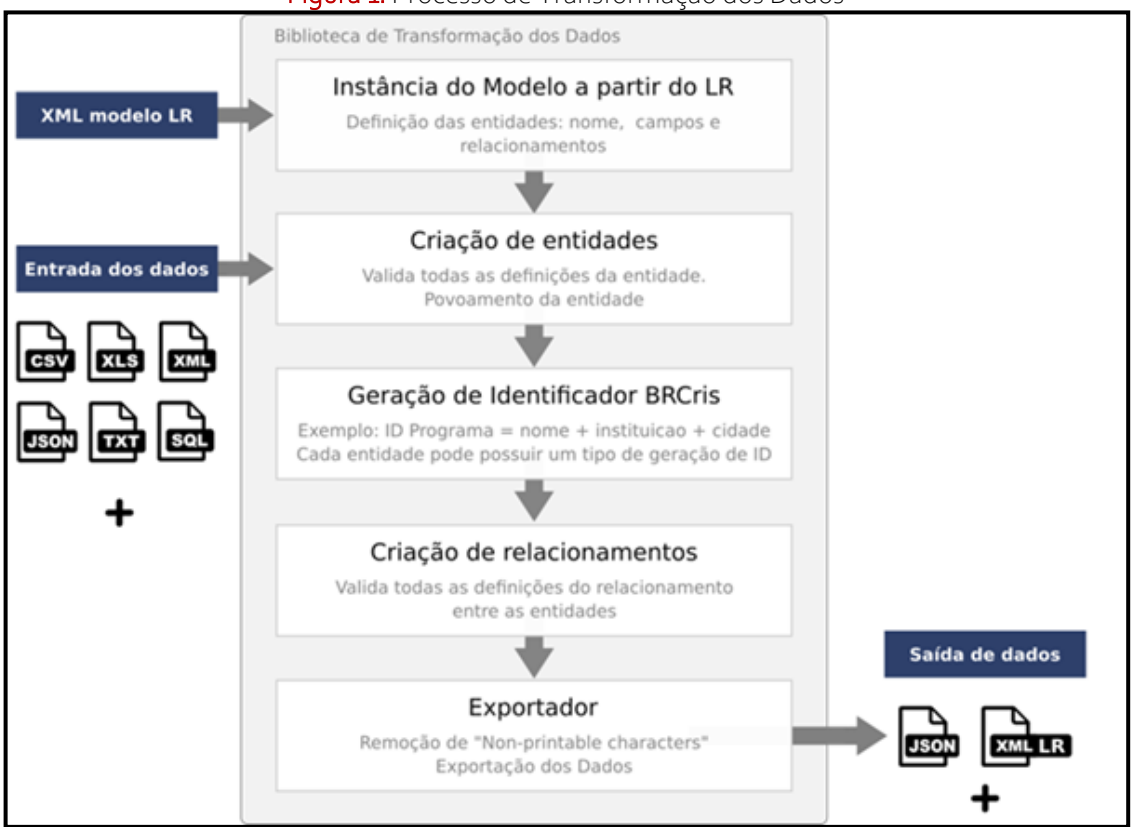
O estudo se caracteriza como descritivo ao apresentar detalhadamente as etapas de tratamento de dados do BrCris, bem como discutir os desafios enfrentados durante os processos de grandes volumes de dados. Para o desenvolvimento da Plataforma do BrCris foi necessário o desenvolvimento de ferramentas para tratamento da informação. Neste contexto, e de forma a atender o primeiro objetivo de discutir a integração e consistência dos dados para o BrCris se fez necessário o desenvolvimento e testes de uma metodologia para essa transformação dos dados estruturados e não estruturados em formato padronizado. Para essa padronização, o modelo de dados do BrCris, iniciou-se pela estruturação de nove entidades de dados, seguindo padrões amplamente utilizados na comunidade científica internacional, a saber: publicações, teses e dissertações, pessoas, revistas, patentes, grupos de pesquisa, softwares, redes de especialistas e temáticas (agrupamento de conceitos).

O processo de transformação de dados para o BrCris é representado na Figura 1. Esse processo se inicia com a coleta de dados das fontes e o mapeamento de seus metadados a partir do modelo La Referência (LR), de forma a definir os tipos de entidades no BrCris. A intenção é reutilizar ontologias já existentes, evitando a criação de novas classes e ou relacionamentos. Cada entidade é identificada e validada por meio de uma combinação de elementos descritivos, garantindo a ausência de ambiguidade. Por exemplo, no caso de um programa de Pós-Graduação, são utilizados o nome do programa, a instituição e a cidade para gerar um identificador único no BrCris.

Estando as entidades criadas, parte-se para a criação e validação dos relacionados entre elas, conforme definida na ontologia. Validada as informações são realizados tratamentos de conteúdo eliminando conteúdo "*Non-printable characters*" (caracteres não imprimíveis) como caracteres de controle, espaços não visíveis e caracteres de formatação, como negrito. Após esse tratamento, os dados estão prontos para serem exportados para a Plataforma do BrCris.

Na criação do modelo semântico foram adotadas como premissas a maximização do reuso de recursos existentes e a utilização de padrões internacionais para representação de dados na área, para que o BrCris fosse compatível com sistemas similares ao redor do mundo.

Figura 1. Processo de Transformação dos Dados



Fonte: Autores (2023)

Após todo o processo de tratamento, independentemente da fonte de dados, os dados gerados como saída, são importados em um único banco de dados, e utilizando-se dos identificadores únicos gerados ou identificados, os conjuntos de dados são vinculados e deduplicados, viabilizando dessa forma a interoperabilidade dos dados, independentemente de sua fonte e formato. Os dados de saída que podem ser caracterizados sob diversos formatos de dados, possibilitam ainda a importação dos conjuntos previamente tratados de tal forma que possam ser incorporados por outras ferramentas de análise.

Para a visualização e validação dos dados são analisados os dados gerados pelo VIVO com a consolidação dessas informações. Essa análise foi realizada pela equipe de desenvolvimento e tratamento de informação do Ibict.

A certificação dos dados do BrCris é realizada por meio da integração com a plataforma Lattes e o Oasisbr, com a descrição do processo de validação de teses e dissertações entre esses sistemas.

4 RESULTADOS

São apontados nessa seção, os resultados obtidos com todo o ferramental desenvolvido conforme descrito anteriormente. Este conjunto de ferramentas tem auxiliado no acesso e avaliação da produção científica brasileira. Podendo ainda, fornecer insumos como conjuntos de dados em formatos padronizados que facilitam a integração com outros conjuntos e ferramentas. Tais ações visam fornecer mecanismos para impulsionar o acesso aos dados científicos de forma facilitada, contribuindo de forma significativa com o avanço da Ciência Aberta no Brasil.

4.1 Integração e Consistência dos Dados

Tendo em vista todo o processo de curadoria dos dados a serem coletados, integrados e analisados no contexto deste projeto, uma estratégia de geração de identificadores se faz necessária. Tais identificadores são importantes pois todos os dados coletados são mapeados para entidades, previamente identificadas, em que se faz necessário identificá-las de forma única, tendo em vista todo o processo de tratamento a ser realizado, em especial o processo de desambiguação e deduplicação dos dados.

Uma ferramenta desenvolvida foi a *Ocean Dragon*, que permite obter dados individuais, instituições, formações, publicações, orientações, entre outras. *Ocean Dragon* possui uma estrutura de dados formada por entidades, relacionamentos, campos aninhados com a possibilidade internacionalização, fazendo um intercâmbio com a Plataforma La Referencia, utilizada pelo Ibict para coletar os dados e indexar em diversas outras Plataformas.

Para tanto, uma biblioteca computacional, utilizando a linguagem de programação Python foi proposta em que ela se torna responsável por gerar Identificadores BrCris, criados com o intuito de realizar uma pré-desambiguação dos dados, evitando entidades duplicadas no conjunto a ser analisado.

Para cada conjunto de dados a serem analisados, uma estratégia para a geração dos identificadores foi considerada. Tal estratégia objetiva utilizar o mínimo possível de informações que são extraídas, mas que possam gerar com um nível satisfatório de confiança, identificadores únicos, que serão utilizados em diversas etapas futuras.

A biblioteca desenvolvida para o tratamento dos dados também é responsável por gerar Identificadores BrCris, criados com o intuito de realizar uma pré-desambiguação dos dados, evitando entidades duplicadas na plataforma La Referencia. A biblioteca ainda permite a validação dos dados em relação ao modelo utilizado para estruturar a Plataforma La Referencia. Ou seja, verifica se os nomes de entidades, campos e relacionamentos estão em conformidade com a Plataforma, evitando problemas durante a carga.

A criação do metamodelo de entidades e relacionamento levou em consideração as características dos diversos conjuntos de dados coletados, visando a facilitar as rotinas de tratamento dos dados, em especial a deduplicação de entidades. Porém, neste formato, as informações resultantes não seriam facilmente acessadas e reutilizadas por agentes externos. Para resolver este problema, foi desenvolvido o segundo nível de representação, o nível

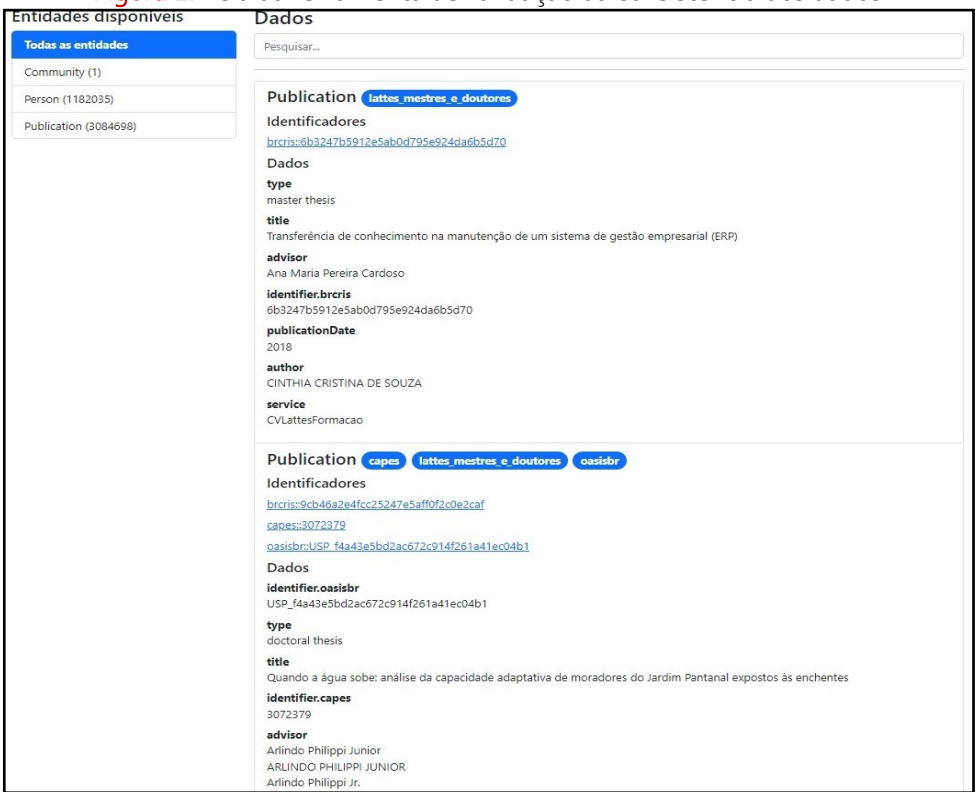
semântico, implementado como uma ontologia para permitir a visualização e navegação dos dados como um grafo de conhecimento.

Diante de todas as etapas descritas anteriormente foi possível realizar de forma consistente todo o processo de coleta, transformação e integração dos dados. Tais estratégias, desenvolvidas com a adoção de diversas técnicas validadas em diversos estudos, viabilizaram a caracterização de um único conjunto de dados devidamente validado.

Ao realizar a carga dos dados na plataforma La Referencia, é fundamental garantir que todas as informações sejam deduplicadas corretamente. Para alcançar esse objetivo, a plataforma utiliza identificadores únicos, conforme mencionados acima, para identificar entidades semelhantes e aplicar o merge dos dados. Estes identificadores são conhecidos como "*brcrisId*", e possuem uma forma de criação distinta para cada tipo de entidade.

Objetivando facilitar a visualização, verificação e análise do que foi enviado para a plataforma La Referencia e como os dados foram tratados, foi desenvolvida uma ferramenta (Figura 2) que permite ao usuário navegar entre as entidades e visualize o resultado da carga e deduplicação de forma clara e intuitiva, atuando neste caso com um curador de dados.

Figura 2. Tela da ferramenta de validação da consistência dos dados



Fonte: Autores (2023)

Ao agrupar entidades com identificadores únicos, a plataforma é capaz de reduzir o tamanho do conjunto de dados e eliminar informações redundantes, tornando-o mais fácil de ser gerenciado e analisado.

No contexto do BrCris, esse processo de deduplicação também é essencial para garantir a interoperabilidade dos dados. Ao selecionar e processar os dados para integração a partir de diferentes fontes de dados, a plataforma utiliza identificadores implícitos ou gerados durante o processo de tratamento dos dados para resolver possíveis deduplicações e garantir a consistência dos dados.

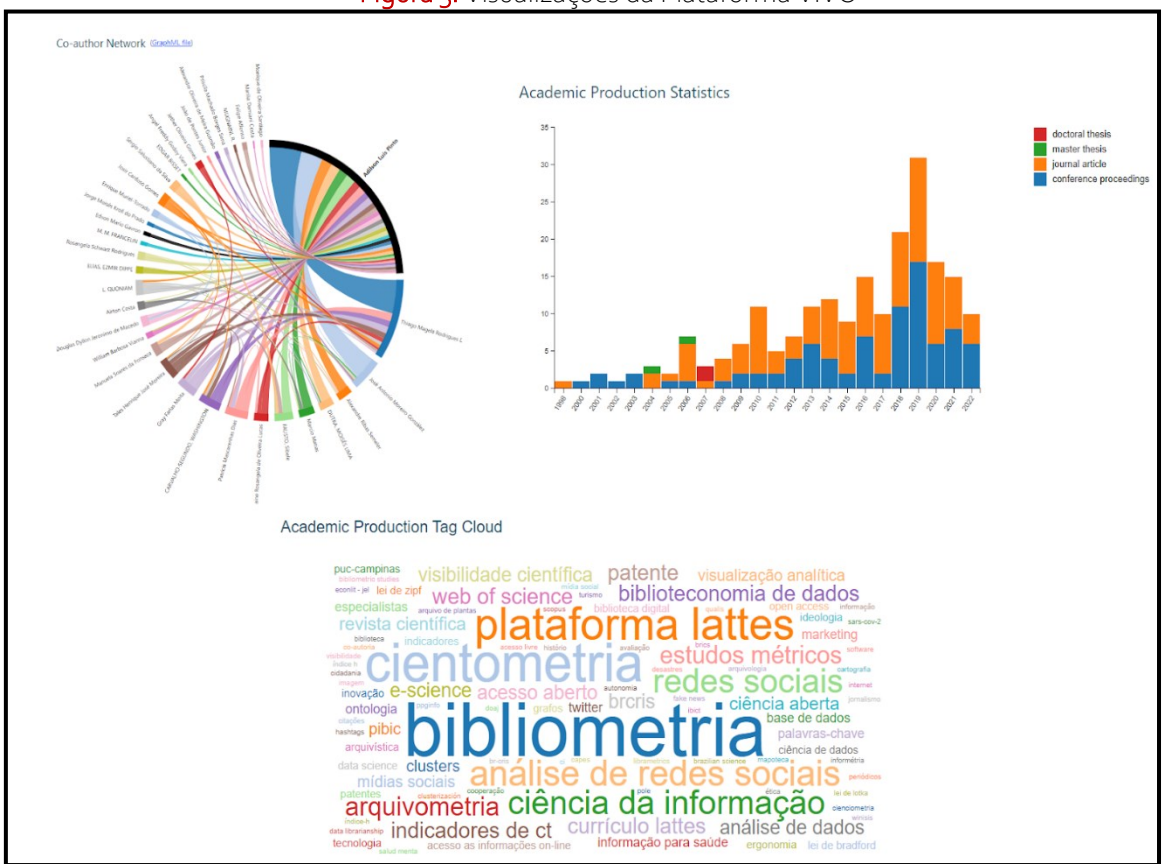
4.2 Visualização dos Dados

Com os dados agregados e deduplicados, a plataforma é capaz de oferecer diversas análises bibliométricas, permitindo que os usuários realizem uma ampla variedade de estudos e pesquisas com base nas informações disponíveis. Em suma, o processo de deduplicação é uma etapa fundamental para garantir a precisão, a confiabilidade e a interoperabilidade dos dados, tornando-os ainda mais valiosos para análises e pesquisas.

Após todas as etapas de coleta, tratamento e integração dos dados, é possível ter acesso ao conjunto de dados com auxílio de interfaces gráficas que facilitam o acesso e certificação dos conjuntos. A ontologia VIVO, em particular, possibilita que os dados sejam visualizados na Plataforma VIVO, uma ferramenta para navegação de dados do domínio acadêmico que permite que o BrCris sirva *Linked Open Data* a agentes externos, além de facilitar a exploração do grafo de conhecimento. Outro recurso importante oferecido pela plataforma VIVO são as visualizações gráficas, que fornecem um panorama mais amplo sobre um determinado indivíduo. Além das visualizações pré-definidas, também é possível implementar e incluir na interface de forma simples gráficos customizados.

A Figura 3 ilustra algumas destas visualizações: a rede de coautoria de um pesquisador, já disponível de forma nativa na plataforma, e dois gráficos customizados implementados para o BrCris: o total de publicações por tipo e a nuvem de palavras dada pelas palavras-chave das publicações de um determinado pesquisador.

Figura 3. Visualizações da Plataforma VIVO



Fonte: BrCris (Ibict, 2023)

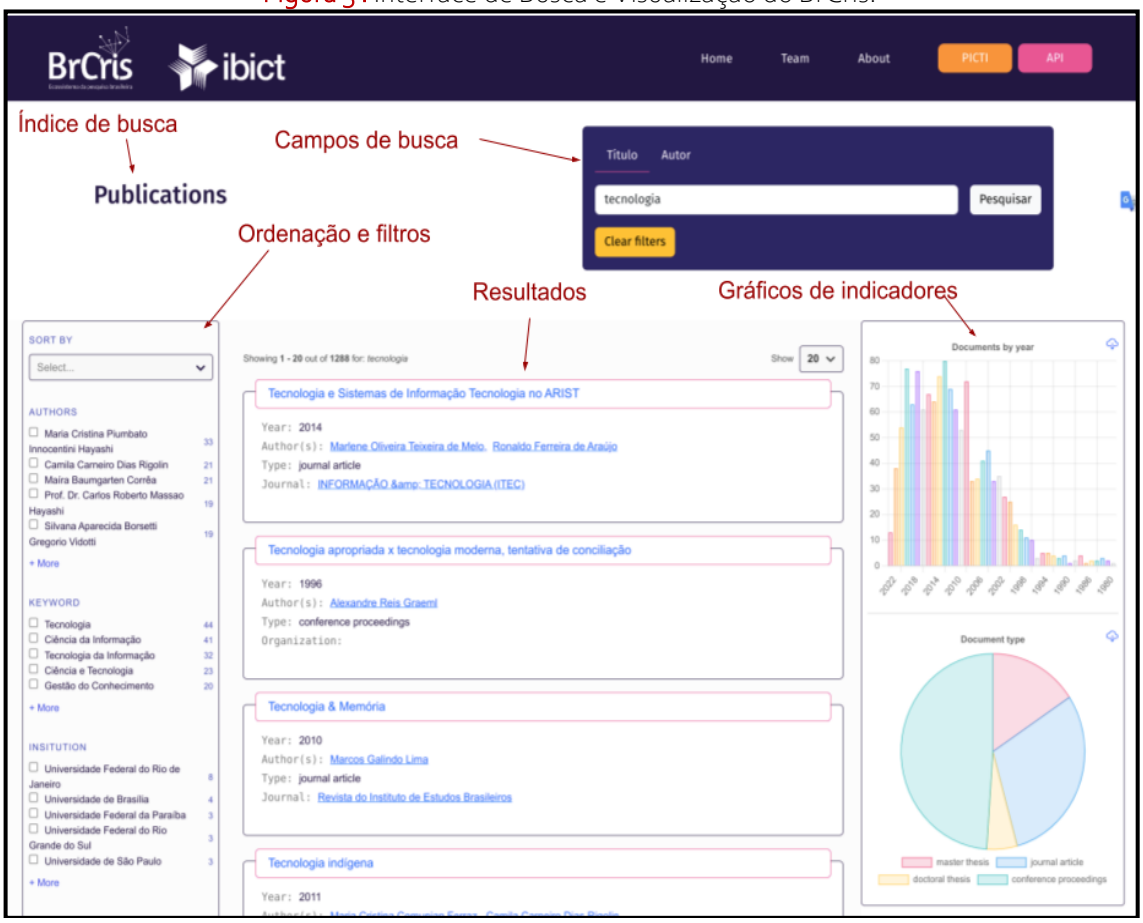
Reforçando o tema de boas práticas de publicação e compartilhamento de dados, é importante destacar que, seguindo padrões internacionais e adotando o mesmo modelo semântico de sistemas semelhantes, garantimos que os dados coletados e tratados no BrCris sejam acessíveis, reutilizáveis e interoperáveis, desta forma aderindo também aos princípios

FAIR (*Findability, Accessibility, Interoperability, Reusability*). A representação em RDF com base na ontologia VIVO possibilita tanto a descrição dos dados sem ambiguidade quanto a recuperação automática e referenciamento cruzado da informação, o que se alinha com a ênfase dos princípios FAIR na descoberta e utilização dos dados de forma automática, em adição ao seu reuso também por indivíduos (Wilkinson *et al*., 2016).

Como componente de recuperação de informação foi desenvolvido uma interface gráfica web (Figura 4) baseada na *Search-UI* da *Elastic. Search-UI* que é uma biblioteca livre de código aberto, escrita em *Typescript*, que disponibiliza uma variedade de componentes web customizáveis e compatíveis com aplicativos desktop e aplicativos móveis, que se integra com o *Elasticsearch* para fornecer uma interface de busca e visualização de informações na web.

O Search-UI [1] dispõe de um conjunto de componentes de busca configuráveis facilitando a construção de interfaces ricas de busca totalmente customizáveis com funcionalidades avançadas de filtragem que ajudam os usuários encontrarem exatamente o que eles precisam. Disponibilizando recursos de paginação, digitação preditiva, autocomplete, filtros, classificação e geração de gráficos de indicadores.

Figura 3 . Interface de Busca e Visualização do BrCris.



Fonte: BrCris (Ibict, 2023)

Com a utilização da Search-UI foi possível desenvolver uma interface para realizar buscas por texto em linguagem natural com a adição de operadores booleanos. As solicitações de buscas dos usuários são cruzadas com um índice do *Elasticsearch*[2], possibilitando que a recuperação e o ranqueamento sejam realizados rapidamente.

[1] A Search-UI é responsável por apresentar ao usuário uma interface intuitiva, amigável e eficiente para realizar buscas de informações em um determinado sistema.

[2] Motor de busca e análise de dados distribuído e de código aberto, desenvolvido pela empresa Elastic.

Realizada a busca, a interface de recuperação do BrCris permite realizar classificação dos resultados, tais como: classificação alfabética, classificação cronológica, bastante útil para pesquisadores, pois permite a seleção dos resultados mais recentes, classificação por relevância de acordo com o ranqueamento do *Elasticsearch*. Também é possível aplicar filtros de refinamento da busca por algum atributo dos documentos, como por exemplo: nome do autor, tipo de documento.

Todas estas visualizações, atualmente disponibilizadas possibilita uma fácil interação com todo o conjunto de dados que foi inicialmente, coletado, tratado e integrado. Além disso, diversos filtros podem ser aplicados e consequentemente terem os resultados de suas buscas exportadas para formatos padronizados, contribuindo de forma significativa com o avanço da Ciência Aberta, tendo em vista que proporciona análises a conjuntos de dados certificados que anteriormente não eram acessíveis.

4.3 Certificação de Dados

Com a integração dos dados em um repositório de dados padronizado e devidamente avaliado, pode-se realizar todo um processo de certificação de dados originários de outras fontes ainda não validadas, proporcionando uma visão real da produção científica e tecnológica brasileira.

Os sistemas de certificação de dados caracterizam-se como mecanismos de verificação da origem, veracidade, integridade e confiabilidade dos conjuntos de dados armazenados em diferentes sistemas (Dias *et al*., 2023).

As ações autodeclaratórias de um sistema podem ser validadas por um agente denominado "terceiro de confiança" (assinatura autodeclaratória versus assinatura certificada), conferindo segurança e veracidade às informações prestadas.

Em fontes de informação de ciência e tecnologia, no entanto, sistemas de certificação de dados ainda são raros no Brasil, tendo em vista os diversos desafios envolvidos no processo de certificação. A ausência de identificadores persistentes que as diversas entidades que compõem todo o ecossistema da investigação científica destacam-se como uma das principais limitações.

Com os dados coletados e já deduplicados, classificados e categorizados, eles podem ser posteriormente adaptados e validados, estabelecendo relações com registros de outras fontes. Um registro coletado na fonte "A" tem um atributo comum com o registro coletado na fonte "B", podendo ser estabelecida uma ligação entre ambos, com certo grau de confiabilidade. Os outros atributos de registro podem ser mesclados para resultar em um único registro enriquecido, eliminando as réplicas. Um esquema de validação pode ser criado para descartar registros malformados, redundantes, inconsistentes ou ambíguos.

No contexto deste trabalho, o primeiro modelo de certificação testado é a integração da Plataforma Lattes com o Oasis.br (https://oasisbr.ibict.br/). Por meio de desdobramentos no âmbito do Projeto BrCris, foi possível criar um mecanismo inteligente de identificação de teses e dissertações declaradas nas sessões de formação acadêmica e orientações concluídas de um determinado Currículo cadastrado na Plataforma Lattes, que também constavam no cadastro agregado pelo Oasisbr.

Desta forma, a Oasisbr torna-se o "terceiro de confiança" neste processo, sem a necessidade da pré-existência de um identificador persistente explicitamente atribuído à tese ou dissertação.

Todo o processo de certificação é baseado em estratégias computacionais testadas e validadas em diversos estudos, por meio da análise de informações autodeclaradas, comparadas com informações inseridas em repositórios, bibliotecas digitais e portais agregados pelo Oasisbr. O selo de certificação é exibido próximo aos títulos das teses ou dissertações no currículo do usuário (Figura 4).

Figura 4 . Fragmento de Um Currículo com Certificação Incluída

Fonte: Plataforma Lattes (CNPq, 2023)

Com o selo é possível obter, de forma rápida e simples, a comprovação documental do título informado e acessar o documento no Oasisbr. O certificado pode ser emitido automaticamente pela Plataforma Lattes ou solicitado manualmente pelo usuário.

Por meio de desenvolvimentos no âmbito do Projeto BrCris, foi possível criar um mecanismo inteligente para identificar teses e dissertações declaradas nas seções de treinamento e orientação concluídas de um determinado Currículo Lattes, que também foram incluídas no conjunto de registros agregados pelo Oasisbr. Desta forma, a Oasisbr torna-se o "terceiro de confiança" neste processo, sem a necessidade da pré-existência de um identificador persistente explicitamente atribuído à tese ou dissertação. Dentre os currículos da Plataforma Lattes, há aproximadamente 1,1 milhão de registros de teses e dissertações, em que 65% (aproximadamente 700 mil) já são passíveis de certificação. Outras 10 mil teses e dissertações defendidas no exterior também foram mapeadas para receber o selo de certificação.

As vantagens do processo de certificação são muitas. Por meio da certificação, é possível verificar que os trabalhos científicos, orientações, participação em bancas, entre outros elementos de fontes autodeclaradas são realmente verdadeiros, evitando informações falsas. Portanto, a certificação acaba sendo um fator que promove maior credibilidade e autoridade ao pesquisador. Desta forma o BrCris é uma importante contribuição para compreensão da ciência aberta brasileira.

5 CONSIDERAÇÕES

O BrCris se configura como um importante espaço de pesquisa e análise de dados. As informações agregadas e organizadas segundo um modelo de dados semântico, permitem a geração de serviços para diversos atores, nos contextos de gestão e pesquisa acadêmica, assim como na área de informação para a inovação.

A plataforma é uma iniciativa que coleta e enriquece dados de repositórios e bases de dados abertas sendo uma proposta, ímpar no mundo, que facilita obter um Panorama Brasileiro da Produção e Atuação de todos os seus atores acadêmicos/científicos. Entretanto requer muitos recursos computacionais e humanos no tratamento e padronização desses dados.

A partir do BrCris, os envolvidos no ecossistema da pesquisa científica brasileira, poderão ter fácil acesso a um grande agregado de dados científicos, segundo as melhores práticas dos princípios FAIR, obtendo conjuntos de dados ou informações de interesse de forma acessível.

Ressalta-se que os currículos Lattes são de natureza autodeclaratória, destacando-se, portanto, a importância dos processos de certificação aplicados a essa base. Uma tese ou dissertação só é considerada documento oficial de titulação, se a versão final, e com as respectivas correções sugeridas pelos avaliadores, for depositada em repositório oficial e de acesso público. Estratégia esta que pode ser replicada em outros conjuntos de dados visando a sua certificação.

REFERÊNCIAS

ABBAS, A.; ZHANG, L.; KHAN, S. U. A literature review on the state-of-the-art in patent analysis. **World Patent Information**, Oxford, UK, v. 37, p. 3-13, 2014.

ABBASI, A.; ALTMANN, J.; HOSSAIN, L. Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. **Journal of informetrics**, Amsterdam, v. 5, n. 4, p. 594-607, 2011.

BAUER, F.; KALTENBÖCK, M. **Linked open data**: the essentials. Vienna: Edition mono/monochrom, 2011. p. 21, v.710.

COLLAZO-REYES, F. Growth of the number of indexed journals of Latin America and the Caribbean: the effect on the impact of each country. **Scientometrics**, Budapest, v. 98, p. 197-209, 2014.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO (CNPq). Plataforma Lattes. Brasília. Disponível em: https://lattes.cnpq.br. Acesso em: 12 out. 2023.

DE MEIS, L. *et al*. The growing competition in Brazilian science: rites of passage, stress and burnout. **Brazilian journal of medical and biological research**, Ribeirão Preto, v. 36, p. 1135-1141, 2003.

DIAS, T. M. R. *et al*. Brcris: plataforma para integração, análises e visualização de dados técnicos-científicos. , p. 622-638, **Informação e Informação**, Londrina, v. 27, n. 3, 2022. DOI: https://doi.org/10.5433/1981-8920.2022v27n3p622.

DIAS, T. M. R. *et al*. *In*: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 6., 2023, Brasília, DF. **Anais**... Brasília: Ibict, 2023. DOI: https://doi.org/10.22477/vi.widat.53.

EUROCRIS. **Directory of Research Information Systems** (DRIS). Nijmegen, Nertherlands. Disponível em: https://eurocris.org/services/dris. Acesso em: 23 out. 2023.

HUANG, Y.; GLÄNZEL, W.; ZHANG, L. Tracing the development of mapping knowledge domains. **Scientometrics**, Budapest, v. 126, p. 6201-6224, 2021.

JÖRG, B. CERIF: The common European research information format model. **Data Science Journal**, London, v. 9, p. CRIS24-CRIS31, 2010.

KONG, X. et al. Academic social networks: Modeling, analysis, mining and applications. **Journal of Network and Computer Applications**, London, v. 132, p. 86-103, 2019.

LANE, J. Let's make science metrics more scientific. **Nature**, London, v. 464, n. 7288, p. 488-489, 2010.

LEE, S.; BOZEMAN, B. The impact of research collaboration on scientific productivity. **Social Studies of Science**, London, v. 35, n. 5, p.673-702, 2005. Disponível em: https://elibrary.ru/item.asp?id=11423996. Acesso em: 27 mar. 2023.

LETA, J.; GLÄNZEL, W.; THIJS, B. Science in Brazil. Part 2: Sectoral and institutional research profiles. **Scientometrics**, Budapest, v. 67, n. 1, p. 87-105, 2006.

MEADOWS, A. J. **A comunicação científica**. Trad. A. A. B. de Lemos. Brasília, DF: Briquet de Lemos, 1999.

PINTO, A. L. *et al*. The Brazilian current research information system: BrCris. *In*: SILVA, Carlos Guardado da Silva; REVEZ, Jorge; CORUJO, Luis (coord.).Organização do conhecimento no horizonte 2030: desenvolvimento sustentável e saúde. Lisboa: Universidade de Lisboa, 2021. p. 319. (Colecção CA–Ciência Aberta). ISBN 978-989-566-137-4. DOI: https://doi.org/10.51427/10451/50067.

RATHKE, S. B.; ROCHA, R. P. Sistema de informação de pesquisa: uso da ontologia de VIVO no contexto das instituições brasileiras. **Brazilian Journal of Information Science**, Marília, v.13, n.4, 2019. DOI: https://doi.org/10.36311/1981-1640.2019.v13n4.08.p132.

SINGH, V. K. The journal coverage of web of science, scopus and dimensions: A comparative analysis. **Scientometrics**, Budapest, v. 126, Jun., 2021. DOI: https://doi.org/10.1007/s11192-021-03948-5.

SIVERTSEN, G. Developing Current Research Information Systems (CRIS) as data sources for studies of research. *In*: GLÄNZEL, W. *et al*. (ed.). **Springer handbook of science and technology indicators**. [*S.l.*]: Springer, Cham. 2019. p. 667-683. DOI: https://doi.org/10.1007/978-3-030-02511-3_25.

TANG, J. *et al.* Arnetminer: extraction and mining of academic social networks. *In*: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 14th, 2008. **Proceedings of the…** Las Vegas, Nevada: ACM, 2008. p. 990-998. DOI: https://doi.org/10.1145/1401890.1402008.

TORINO, E.; CONEGLIAN, C. S.; VIDOTTI, S. A. B. G. Estruturas de representação para reuso de dados no contexto da ecologia de pesquisa: Cris institucional. Informação e Informação, Londrina, 2020, v.25, n. 3. DOI: https://doi.org/10.5433/1981-8920.2020v25n3p1.

YOSHIKANE, F.; KAGEURA, K. Comparative analysis of coauthorship networks of different domains: The growth and change of networks. **Scientometrics**, Budapest, v. 60, p. 435-446, 2004.