# Inovação e Conectividade: Uma perspectiva sobre o Projeto BrCris e suas tecnologias para tratamento de dados científicos

*Innovation and Connectivity: A Perspective on the BrCris Project and its Technologies for Processing Scientific Data*

*Innovación y Conectividad: Una Perspectiva sobre el Proyecto BrCris y sus Tecnologías de Procesamiento de Datos Científicos*

**Washington Luís Ribeiro de Carvalho Segundo**
Doutorado em Informática, Universidade de Brasília (UnB), Brasília, DF, Brasil.
Tecnologista, Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasília, DF, Brasil.
http://lattes.cnpq.br/9453481318889500
https://orcid.org/0000-0003-3635-9384

**Thiago Magela Rodrigues Dias**
Doutor em Modelagem Matemática e Computacional, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Belo Horizonte, MG, Brasil.
Professor, Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Divinópolis, MG, Brasil.
http://lattes.cnpq.br/4687858846001290
https://orcid.org/0000-0001-5057-9936

**Marcel Garcia de Souza**
Mestre em Educação em Ciências, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brasil.
Analista em Ciência e Tecnologia, Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasília, DF, Brasil.
http://lattes.cnpq.br/9517728665816047
https://orcid.org/0000-0003-2255-199X

**Fhillipe de Freitas Campos**
Graduado em Biblioteconomia, Universidade de Brasília (UnB), Brasília, DF, Brasil.
Pesquisador, Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasília, DF, Brasil.
http://lattes.cnpq.br/2076669848354453
https://orcid.org/0000-0002-7093-703X

**Denise Aparecida Freitas de Andrade**
graduada em Biblioteconomia, Universidade de Brasília (UnB), Brasília, DF, Brasil.
Pesquisadora, Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasília, DF, Brasil.
http://lattes.cnpq.br/6698900487294293
https://orcid.org/0000-0003-3988-5929

Anais do **VII Workshop de Informação, Dados e Tecnologia**, Porto Velho, 25-27 jun. 2024.
https://doi.org/10.22477/vii.widat.182

## Resumo

**Introdução**: No contexto brasileiro, a Plataforma BrCris emerge como uma iniciativa inovadora, integrando dados de todo o ecossistema de pesquisa científica nacional. Abrangendo a diversidade da produção científica, desde artigos até produções técnicas como softwares e patentes. Metodologia: Este trabalho apresenta as tecnologias envolvidas no processo de integração de dados em um repositório padronizado, possibilitando uma visão precisa da produção científica e tecnológica brasileira. Destaca-se a interoperabilidade de dados e a aderência a padrões internacionais de modelagem como estratégias fundamentais. **Resultados**: Exploram-se e detalham-se os processos de transformação e integração de dados, enfatizando a qualidade, consistência e padronização das informações disponibilizadas. As interfaces de visualização amigáveis proporcionam uma compreensão aprofundada do ecossistema de pesquisa científica no Brasil. **Conclusão**: Destaca-se o papel crucial da Plataforma BrCris na promoção da integração e acessibilidade no ecossistema de pesquisa nacional. Ao consolidar dados de diversas fontes, a plataforma fortalece a difusão das informações, contribuindo para o avanço científico em todas as áreas do conhecimento.

**Palavras-chave:** BrCris; informação científica; ecossistema da pesquisa, certificação; análise de dados.

## *Abstract*

***Introduction**: In the Brazilian context, the BrCris Platform emerges as an innovative initiative, integrating data from the entire national scientific research ecosystem. Covering the diversity of scientific production, from articles to technical productions such as software and patents. **Methodology**: This work presents the technologies involved in the data integration process into a standardized repository, enabling an accurate view of Brazilian scientific and technological production. Data interoperability and adherence to international modeling standards stand out as fundamental strategies. **Results**: Data transformation and integration processes are explored and detailed, emphasizing the quality, consistency and standardization of the information made available. User-friendly visualization interfaces provide an in-depth understanding of the scientific research ecosystem in Brazil. **Conclusion**: The crucial role of the BrCris Platform in promoting integration and accessibility in the national research ecosystem stands out. By consolidating data from different sources, the platform strengthens the dissemination of information, contributing to scientific advancement in all areas of knowledge.*

***Keywords:** BrCris; scientific information; research ecosystem, certification; data analysis.*

## *Resumen*

***Introducción**: En el contexto brasileño, la Plataforma BrCris surge como una iniciativa innovadora, integrando datos de todo el ecosistema de investigación científica nacional. Abarcando la diversidad de la producción científica, desde artículos hasta producciones técnicas como software y patentes. **Metodología**: Este trabajo presenta las tecnologías involucradas en el proceso de integración de datos en un repositorio estandarizado, permitiendo una visión precisa de la producción científica y tecnológica brasileña. La interoperabilidad de datos y el cumplimiento de los estándares internacionales de modelización se destacan como estrategias fundamentales. **Resultados**: Se exploran y detallan los procesos de transformación e integración de datos, enfatizando la calidad, consistencia y estandarización de la información puesta a disposición. Las interfaces de visualización fáciles de usar brindan una comprensión profunda del ecosistema de investigación científica en Brasil. **Conclusión**: Destaca el papel crucial de la Plataforma BrCris en la promoción de la integración y accesibilidad en el ecosistema de investigación nacional. Al consolidar datos de*

*diferentes fuentes, la plataforma fortalece la difusión de información, contribuyendo al avance científico en todas las áreas del conocimiento.*

***Palabras clave:*** *BrCris; información científica; ecosistema de investigación, certificación; análisis de datos.*

## 1 INTRODUÇÃO

A dinâmica e expansão da produção científica brasileira, marcada por uma diversidade de campos disciplinares e uma profusão de resultados que vão desde artigos acadêmicos até inovações tecnológicas, revelam a necessidade premente de uma plataforma abrangente para integração e acesso a este vasto conjunto de dados. Nesse contexto, a Plataforma BrCris emerge como uma iniciativa fundamental, englobando todo o ecossistema da pesquisa científica no Brasil.

A produção do conhecimento científico, um processo gradual e incremental, demanda uma compreensão detalhada do estado da arte em diferentes áreas. Tradicionalmente, essa compreensão é extraída da análise criteriosa de publicações especializadas, colaborações científicas, bem como registros de patentes, dentre outros. A Plataforma BrCris surge como resposta a essa demanda, buscando integrar dados de pesquisadores, projetos, instituições e resultados de pesquisa em um sistema abrangente, conhecido como *Current Research Information System* (CRIS).

CRIS define um sistema de informação sobre todo o ecossistema do processo científico. São organizadas em um só lugar todas 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).

Neste trabalho, apresentamos uma visão detalhada da Plataforma BrCris, com foco especial nas tecnologias envolvidas no tratamento de dados. Exploramos o panorama das tecnologias que viabilizam a integração de informações

provenientes de diversas fontes, proporcionando uma visão consolidada da produção científica e tecnológica brasileira. A metodologia adotada destaca a importância da interoperabilidade de dados e a conformidade com padrões internacionais de modelagem, garantindo a qualidade e consistência das informações.

Os resultados expõem os processos de transformação e integração de dados, enfatizando a padronização das informações e a criação de interfaces de visualização amigáveis. Essas interfaces não apenas simplificam a compreensão do ecossistema de pesquisa científica no Brasil, mas também facilitam a identificação de tendências e padrões valiosos.

Este trabalho representa, portanto, uma contribuição essencial para a simplificação do acesso a conjuntos de dados extensos e o estímulo a novas pesquisas no contexto científico brasileiro

## 2 PROCEDIMENTOS METODOLÓGICOS

A metodologia proposta visa oferecer uma compreensão abrangente do Projeto BrCris e suas tecnologias, permitindo uma apresentação sólida e informada no contexto do tratamento de dados científicos.

O BrCris agrega um extenso ecossistema de dados provenientes de diversas fontes, como dados curriculares individuais, informações sobre organizações, programas de pós-graduação, publicações, patentes, revistas científicas, entre outros. O tratamento desses dados demanda um esforço significativo. Diante da diversidade de fontes que contribuem para o BrCris, é essencial realizar a transformação dos dados para um formato padronizado.

No que diz respeito ao modelo de dados adotado pelo BrCris, a abordagem começou com a incorporação de nove entidades de dados, alinhadas aos padrões amplamente reconhecidos na comunidade científica internacional. Essa escolha visa garantir consistência e interoperabilidade, facilitando a integração de

informações de diversas origens. O processo de transformação, portanto, representa uma etapa crucial para harmonizar dados heterogêneos, promovendo uma base sólida para a construção e utilização eficaz do BrCris.

O BrCris adota um esquema de representação de dados em dois níveis. O primeiro consiste no nível lógico, concretizado como um modelo de entidades e relacionamentos fundamentado no CERIF (Jörg, 2010), também denominado de metamodelo. Esse modelo é subsequentemente traduzido para um esquema relacional físico na plataforma LA Referencia, onde os dados são carregados e processados. Este modelo, conhecido como metamodelo, atende às demandas internas de armazenamento de dados, desempenhando um papel crucial na organização e integração das informações coletadas, transformando-as em insumos para os objetivos delineados no projeto (Pinto et al., 2021).

Para representar o domínio acadêmico e científico, utilizou-se 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 e 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.

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.

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

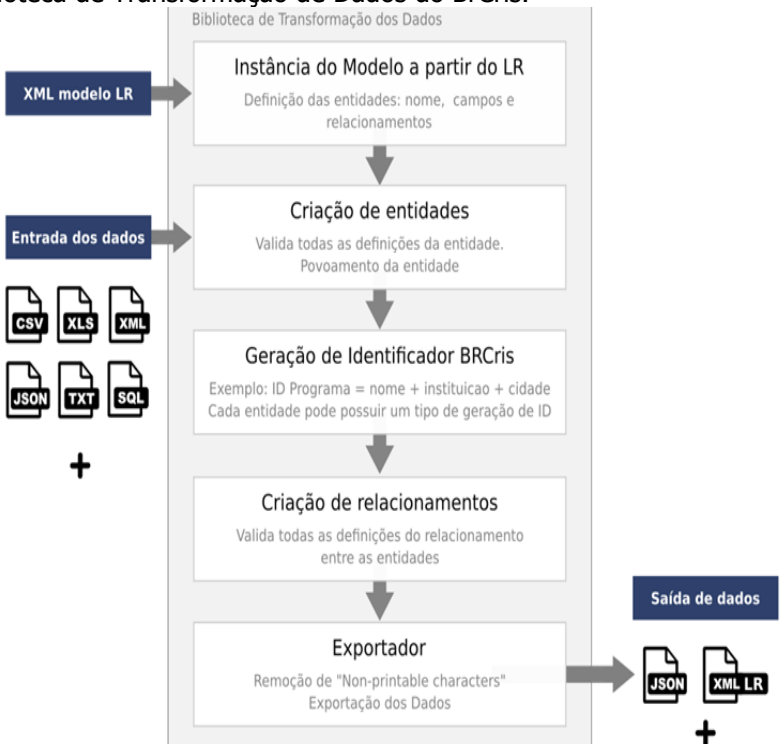
Oasisbr, ou em qualquer repositório onde estejam identificadas por seu DOI; e assim por diante.

O modelo de dados é delineado por um conjunto de entidades e relações, cada uma dotada de identificadores e atributos predefinidos. A adoção de um descritivo tem como objetivo simplificar a identificação dos atributos associados a cada entidade e suas inter-relações. Essa abordagem possibilita a integração fluida de todas as modificações efetuadas diretamente no modelo.

A estratégia adotada visa facilitar de maneira significativa a incorporação de novos atributos e relações, eliminando a necessidade de alterações diretas no modelo de dados. Esse enfoque proporciona uma flexibilidade essencial para a evolução do sistema, permitindo que novos elementos sejam adicionados de forma eficiente e sem impactos substanciais na estrutura preexistente do modelo.

Para efetuar o tratamento dos dados, concebeu-se uma biblioteca computacional que incorpora uma estrutura de dados elaborada para otimizar o processamento de informações provenientes de diversas fontes, adequando-as ao formato requerido pela plataforma LA Referencia (Figura 1).

Figura 1 – Biblioteca de Transformação de Dados do BrCris.




Fonte: Dados da pesquisa (2024).

Como pode ser observado, independentemente do formato de dados que seja importado pela biblioteca, todo um conjunto de tratamento nos dados são aplicados, com rotinas computacionais amplamente otimizadas, possibilitando com auxílio do descritivo de dados, identificar as entidades, bem como os seus relacionamentos e gerar um arquivo de saída, para que este possa ser importado pela plataforma La referencia, ou outra ferramenta de armazenamento.

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

Nesse contexto, a biblioteca desempenha um papel integral na condução da transformação e exportação dos dados. Isso implica na validação das entidades,

campos e relações estipulados pelo modelo, garantindo a conformidade e consistência dos dados processados.

## 3 RESULTADOS

Os resultados alcançados até o momento do projeto abrangem a concepção da arquitetura do BrCris, a identificação e mapeamento das diversas fontes de dados a serem integradas pelo sistema, a efetiva implementação de testes para comprovar a capacidade de agregação dos recursos previamente mapeados.

A implementação de várias visualizações já está concluída, proporcionando uma visão inovadora e abrangente da produção científica nacional. Essas visualizações permitem a aplicação de filtros e outros métodos de personalização, possibilitando análises detalhadas tanto em termos temporais quanto por áreas específicas, como ilustrado na Figura 2.

Figura 2 – Painel de Visualização de Patentes.


Fonte: Dados da pesquisa (2024).

Como parte do sistema de recuperação de informações, foi desenvolvida uma interface gráfica web (Figura 3) fundamentada na Search-UI da Elastic. A Search-

UI, uma biblioteca de código aberto escrita em Typescript, oferece uma variedade de componentes web personalizáveis compatíveis tanto com aplicativos desktop quanto móveis. Esta biblioteca integra-se ao Elasticsearch, proporcionando uma interface de busca e visualização de informações na web.

A Search-UI apresenta um conjunto de componentes de busca configuráveis, simplificando a criação de interfaces de busca altamente personalizáveis com funcionalidades avançadas de filtragem. Essas funcionalidades incluem recursos como paginação, digitação preditiva, autocompletar, filtros, classificação e geração de gráficos de indicadores. Essa abordagem visa facilitar aos usuários a localização precisa do que necessitam, promovendo uma experiência de busca eficiente e aprimorada.

Figura 3 – Interface de Busca de Publicações Científicas.

Fonte: Dados da pesquisa (2024).

Esses avanços representam uma contribuição significativa para a compreensão do cenário da pesquisa científica no Brasil, oferecendo ferramentas flexíveis e personalizáveis para os usuários explorarem dados de maneira eficaz e inédita. Este progresso destaca o potencial do BrCris em fornecer resultados e promover o avanço contínuo no ecossistema de pesquisa do país.

## 4 Considerações Finais

O presente trabalho proporcionou uma visão sobre o Projeto BrCris e suas tecnologias para o tratamento de dados científicos, destacando seu papel no cenário brasileiro de pesquisa. Ao integrar dados de todo o ecossistema científico nacional, desde artigos e patentes até softwares e teses, a plataforma emerge como uma ferramenta abrangente para a promoção da conectividade e acessibilidade no ambiente científico e acadêmico.

A metodologia de integração, transformação e visualização de dados adotada revelou-se robusta, destacando a interoperabilidade e a aderência a padrões internacionais como fundamentais para o sucesso do projeto. Ao consolidar dados de forma padronizada, certificada e acessível, a plataforma não apenas reflete a diversidade e o potencial criativo dos pesquisadores, mas também pavimenta o caminho para futuros estudos e descobertas.

## Referências

BAUER, Florian; KALTENBÖCK, Martin. Linked open data: The essentials. Edition mono/monochrom, Vienna, v. 710, p. 21, 2011.

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

PINTO, Adilson Luiz et al. The brazilian current research information system: BrCris. Colecção CA–Ciência Aberta, p. 319, 2021.

RATHKE, Sandra Beatriz; ROCHA, Rafael Port da. Sistema de informação de pesquisa: Uso da ontologia de VIVO no contexto das instituições brasileiras. Brazilian Journal of Information Science: research trends. Marília, SP: UNESP, Faculdade de Filosofia e Ciências. Vol. 13, n. 4 (dez. 2019), p. 132-151, 2019.

SIVERTSEN, Gunnar. Developing Current Research Information Systems (CRIS) as data sources for studies of research. Springer handbook of science and technology indicators, p. 667-683, 2019.