BrCris: Uma Plataforma para Integração, Interoperabilidade e Análise da Produção Científica Brasileira

Renan Carneiro[1], Thiago M. R. Dias[1], Washington L. R. C[1]. Segundo, Marcel G. de Souza[1], Fabio L. Canto[1]

[1]Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) – Brasília – DF – Brazil

renancarbat@gmail.com, thiagomagela@gmail.com, washingtonsegundo@ibict.br, marcelsouza@ibict.br, fabiocanto@ibict.br,

Abstract. *Brazilian scientific production is fragmented across multiple databases and repositories, hindering integrated and interoperable analyses. The BrCris Platform constitutes a computational ecosystem that gathers, integrates, transforms, and makes available scientific and technical data from open sources. This article presents the platform's integration architecture, its semantic model based on international ontologies, and the computational resources for exporting and visualizing data in standardized formats. It also highlights how these functionalities enhance bibliometric and scientometric studies, enabling complex analyses, interoperability with other tools, and the promotion of Open Science in Brazil.*

Resumo. *A produção científica brasileira se encontra fragmentada entre múltiplas bases de dados e repositórios, dificultando a realização de análises integradas e interoperáveis. A Plataforma BrCris, constitui um ecossistema computacional que reúne, integra, transforma e disponibiliza dados científicos e técnicos a partir de fontes abertas. Este artigo apresenta a arquitetura de integração da plataforma, seu modelo semântico baseado em ontologias internacionais, e os recursos computacionais para exportação e visualização dos dados em formatos padronizados. Destaca-se ainda como essas funcionalidades potencializam estudos bibliométricos e cientométricos, permitindo análises complexas, interoperabilidade com outras ferramentas e a promoção da Ciência Aberta no Brasil.*

1. Introdução

A produção do conhecimento científico é um processo que leva tempo e é incremental (HUANG et al., 2021). A identificação do estado da arte de uma determinada área de conhecimento é amplamente conduzida por meio da análise criteriosa de publicações especializadas que concentram informações relevantes sobre o tópico em questão.

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 et al., 2006), da análise da colaboração científica (YOSHIKANE e KAGEURA, 2004) e da análise de registros de patentes (ABBAS et al., 2014).

A produção e disseminação do conhecimento científico são atividades centrais para o desenvolvimento social, econômico e tecnológico de qualquer nação. No contexto brasileiro, a riqueza da produção acadêmica está dispersa em uma pluralidade de fontes institucionais, repositórios digitais, bases de dados temáticas, currículos de pesquisadores e outras instâncias de registro e disseminação da atividade científica. Apesar dessa pluralidade, a ausência de uma infraestrutura computacional capaz de integrar essas informações de forma interoperável tem dificultado análises sistêmicas, a gestão eficiente da informação científica e a formulação de políticas públicas embasadas em evidências.

A partir desse cenário, começaram a surgir iniciativas voltadas para a criação de sistemas que gerem a produção acadêmica de uma instituição, país ou área do conhecimento. Tais sistemas são conhecidos pela sigla CRIS (*Current Research Information Systems*) e têm como objetivo agregar informações de diversas bases de dados a fim de fornecer relatórios e dados consolidados para que os pesquisadores da área possam analisar como ocorre a produção em seus países ou áreas de atuação.

CRIS define um sistema de informação sobre todo o ecossistema de produção científica. Todas as informações sobre o ciclo de pesquisa científica estão organizadas em um só lugar, desde a divulgação, passando pelos Projetos, Pesquisadores, Instituições de Pesquisa e Laboratórios, até os outputs de uma pesquisa científica, como artigos científicos, teses, dissertações, livros, capítulos de livros, patentes e conjuntos de dados científicos (SIVERTSEN, 2019).

A Plataforma BrCris, coordenada pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), surge como uma resposta a essa lacuna, oferecendo um ecossistema de dados estruturados e integrados sobre a produção científica, técnica e tecnológica brasileira. A plataforma adota os princípios da Ciência Aberta e segue padrões internacionais de interoperabilidade como o CERIF (*Common European Research Information Format*) e a ontologia VIVO, promovendo não apenas a integração de dados, mas também a certificação, a deduplicação e a disponibilização em formatos reutilizáveis.

Dentre os desafios enfrentados na ciência da computação aplicados a esse contexto estão: (1) o tratamento de grandes volumes de dados heterogêneos; (2) a identificação e desambiguação de entidades como autores, publicações e instituições; (3) a manutenção de relacionamentos complexos entre objetos de dados; (4) a aderência a vocabulários e ontologias estabelecidos internacionalmente; (5) a exportação dos dados em formatos compatíveis com ferramentas analíticas amplamente utilizadas.

A Plataforma BrCris apresenta vantagens frente a outras soluções de gestão e análise da informação científica e tecnológica. Sua principal característica está na capacidade de integrar, em um ambiente unificado e padronizado, dados provenientes de múltiplas fontes confiáveis como currículos da Plataforma Lattes, bases de dados de publicações e registros institucionais, permitindo uma visão abrangente, atualizada e contextualizada da produção científica nacional. Além disso, a BrCris adota padrões internacionais de interoperabilidade (como o modelo CERIF), o que facilita a exportação, reutilização e comparação de dados em nível internacional.

Neste artigo, apresenta-se em detalhes a estrutura, os modelos, as funcionalidades e as possibilidades de uso da Plataforma BrCris. A partir de uma perspectiva técnica e aplicada, discutimos como a plataforma contribui para transformar

dados brutos dispersos em uma infraestrutura aberta e reutilizável, apta a impulsionar estudos bibliométricos, cientométricos, altimétricos e de redes. A proposta se insere no escopo da área de "Bases e Fontes de Dados", com forte interface com ciência de dados, engenharia de dados e sistemas de informação.

2. Metodologia

A arquitetura metodológica da Plataforma BrCris combina estratégias de coleta, integração, transformação semântica, modelagem ontológica, visualização e exportação de dados. O projeto parte da premissa de que os dados científicos são gerados por múltiplos atores e depositados em múltiplos repositórios, com diferentes padrões e formatos. É necessário, portanto, um pipeline capaz de tratar essa heterogeneidade com rigor técnico e aderência a padrões.

Inicialmente, foram mapeadas mais de 15 fontes de dados de interesse para o BrCris, incluindo: Plataforma Lattes, OpenAlex, DOAJ, Oasisbr, Espacenet, CAPES, Wikidata, NDLTD, INPI, entre outras. A coleta dos dados foi realizada via APIs REST sempre que disponíveis, ou por meio de *harvesting* OAI-PMH e *crawlers* especializados, nos casos em que APIs não estavam disponíveis. Em todos os casos, optou-se por dados de acesso aberto, em conformidade com os princípios da Ciência Aberta.

Os dados são transformados segundo dois modelos complementares: um modelo lógico baseado em Entidade-Relacionamento (E-R), com vistas ao armazenamento relacional e à visualização via dashboards; e um modelo semântico baseado na ontologia VIVO, com vistas à interoperabilidade semântica, ao fornecimento de *Linked Open Data* e à navegação baseada em grafos.

O processo de integração semântica inclui a desambiguação de autores, a deduplicação de registros e o enriquecimento com metadados de fontes externas. Além disso, foi implementado um mecanismo para certificação de dados, baseado na confiabilidade da fonte, consistência interna e presença de identificadores persistentes (ex: ORCID, DOI, ROR).

Finalmente, os dados são disponibilizados via três interfaces: (1) interface de busca baseada em *Search-UI* e *Elasticsearch*; (2) *dashboards* interativos com aplicação de filtros em tempo real; e (3) visualização semântica via VIVO, com exportação em CSV, GraphML e RDF.

3. Resultados

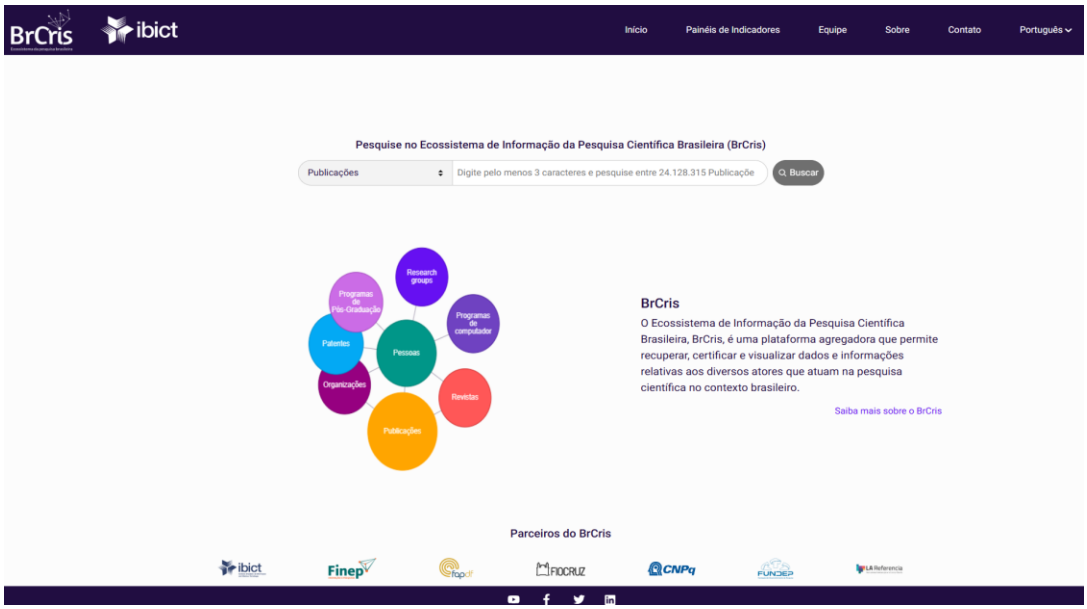
Os resultados do BrCris podem ser observados em três frentes principais: (1) integração e enriquecimento de dados; (2) visualização e exploração; (3) exportação e reuso.

No que tange à integração, atualmente a plataforma agrega mais de 16 milhões de registros distintos, relacionados a publicações, autores, grupos de pesquisa, programas de pós-graduação, patentes, softwares e instituições. Cada entidade está ligada a outras por relacionamentos como autoria, orientação, vinculação institucional, colaboração em projeto, entre outros.

Todas as interfaces são especialmente projetadas para simplificar o processo de acesso e certificação dos conjuntos de dados disponíveis, fornecendo uma experiência

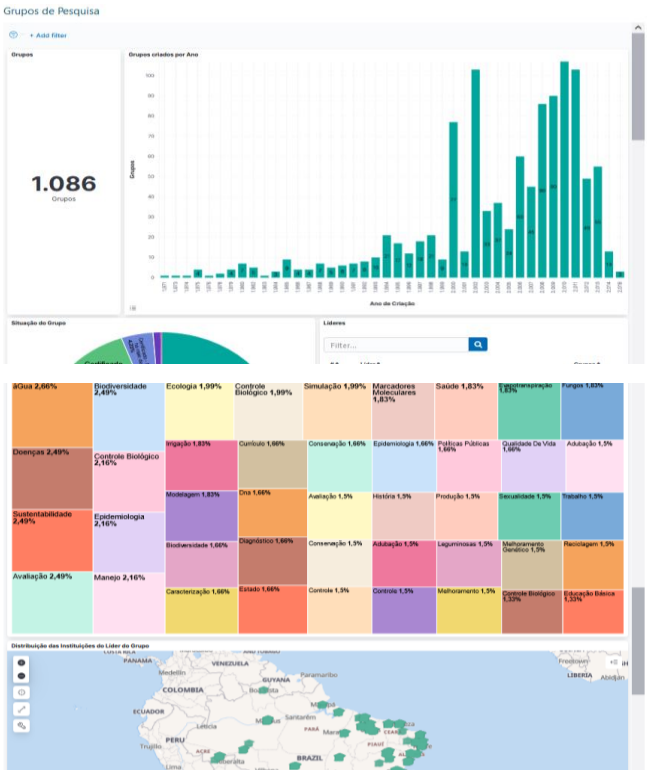
amigável e visualmente atrativa, permitindo uma interação mais intuitiva com as informações contidas nos conjuntos (Figura 1).

Figura 1. Página principal do BrCris.



Ao acessar a opção "Painéis e Indicadores", é possível realizar análises de várias entidades e explorar várias métricas associadas a elas. Essas visualizações oferecem uma experiência interativa, permitindo aplicar filtros diversos e gerar visualizações personalizadas sobre os conjuntos de dados (Figura 2).

Figura 2. Painel de Indicadores da Entidade Grupos de Pesquisa.

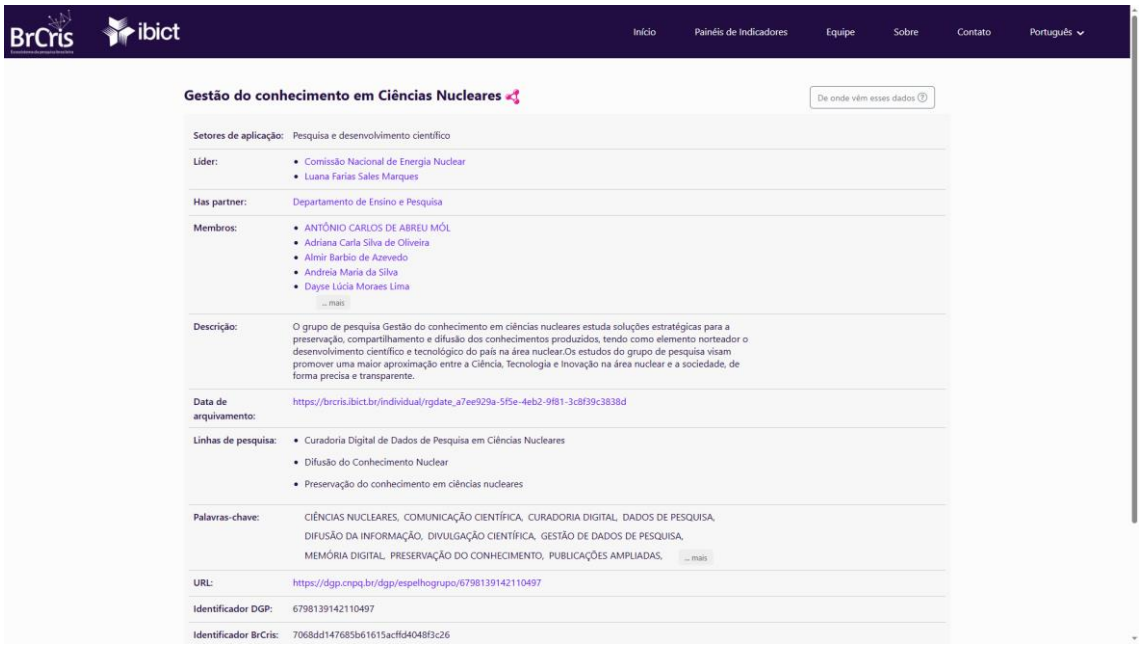


Nas visualizações disponíveis, é possível aplicar diversos filtros em praticamente todos os campos que compõem uma entidade. Isso permite a criação de visualizações específicas, levando em consideração apenas os filtros aplicados. Esses subconjuntos de dados podem ser analisados sob diferentes perspectivas e visualizações.

Um grande diferencial da plataforma é a possibilidade de exportar os subconjuntos de dados em formatos padronizados, como arquivos .csv. Esses arquivos podem ser facilmente importados por diversas outras ferramentas de análise e visualização de dados, proporcionando maior flexibilidade e permitindo uma análise mais aprofundada.

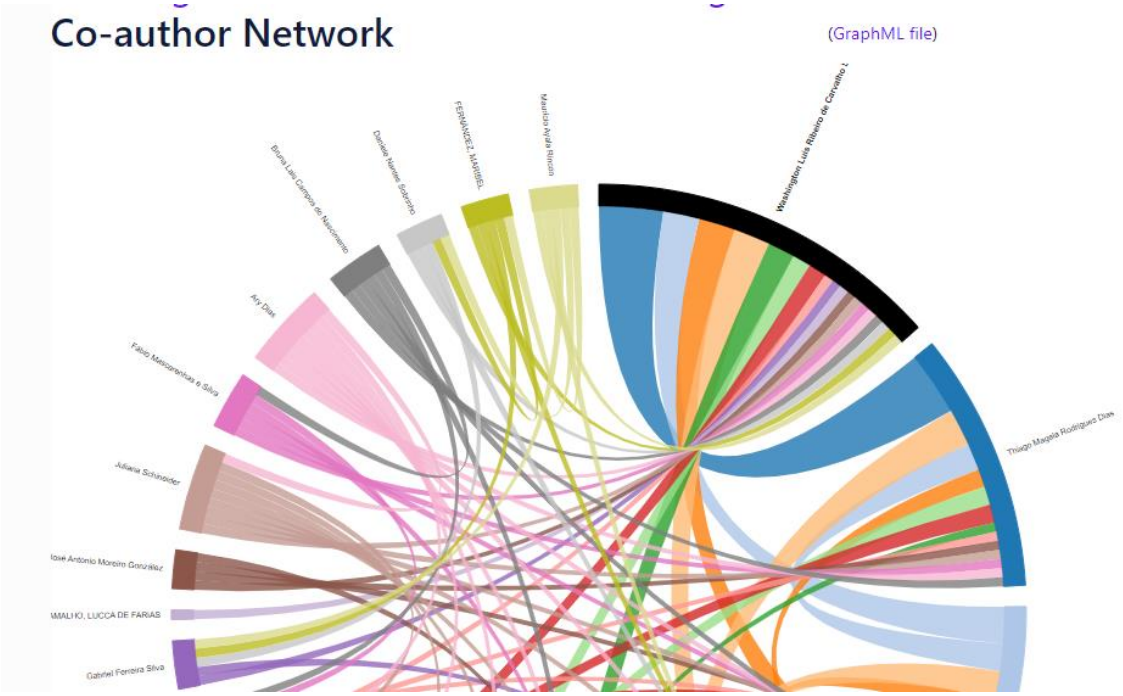
A ontologia VIVO, em particular, possibilita que os dados sejam visualizados na Plataforma VIVO, uma ferramenta para navegação de dados do domínio acadêmico que permite que o BrCris sirva *Linked Open Data* a agentes externos, além de facilitar a exploração do grafo de conhecimento. Outro recurso importante 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.

Figura 3. Visualização de uma Entidade dos Grupos de Pesquisa no ambiente VIVO .



Quanto à exportação, os dados podem ser exportados em CSV, para uso em planilhas ou ferramentas bibliométricas como VOSviewer; em GraphML, para uso em Gephi e ferramentas de análise de redes; e em RDF, para uso em SPARQL endpoints e sistemas semânticos. Também está em desenvolvimento uma API REST para consulta programável. Exemplo do grafo de colaboração de uma entidade pesquisador e sua possibilidade de exportação pode ser visualizado na Figura 4.

Figura 4. Grafo de colaboração de um Pesquisador.



Conforme evidenciado, além da representação visual da rede de coautoria do pesquisador escolhido, há a facilidade de exportar o arquivo no formato de grafo para análises mais específicas em ferramentas dedicadas ao estudo de grafos. Além disso, dados como as publicações que contribuíram para a formação da rede e as informações sobre os colaboradores identificados podem ser exportados em arquivos .csv. Essa capacidade de exportação diversificada fortalece a utilidade da plataforma VIVO como uma ferramenta flexível, proporcionando aos usuários a liberdade de explorar os dados em contextos analíticos mais especializados, além da exportação dos dados.

Diante do exposto, é evidente a ampla variedade de ferramentas disponíveis na Plataforma BrCris. Essas ferramentas têm como objetivo fornecer uma contribuição significativa para a comunidade acadêmica, oferecendo mecanismos de fácil utilização e dados confiáveis sobre o ecossistema da pesquisa científica brasileira. Isso viabiliza análises que buscam compreender de forma mais aprofundada a ciência no Brasil e promover a realização de novos estudos.

Ao disponibilizar essas ferramentas de forma acessível e amigável, a Plataforma BrCris oferece recursos que permitem aos usuários navegar pelos dados, identificar tendências, padrões e visões relevantes. Isso proporciona uma melhor compreensão da pesquisa científica brasileira, contribuindo para a geração de conhecimento e a promoção de estudos inovadores.

4. Discussão

A Plataforma BrCris se insere em uma tendência global de criação de sistemas CRIS nacionais e regionais, com foco na consolidação da informação científica como recurso estratégico. A proposta brasileira apresenta diferenças relevantes ao adotar uma

abordagem federada, baseada em fontes abertas e com forte aderência aos princípios FAIR e à Ciência Aberta.

Um dos principais diferenciais técnicos da plataforma é sua capacidade de adaptar-se a diferentes padrões internacionais por meio de módulos de transformação de dados. Por exemplo, o BrCris é capaz de converter metadados oriundos da Plataforma Lattes em estruturas compatíveis com o OpenAIRE ou ROR, assegurando interoperabilidade em escala global. Isso permite, por exemplo, que dados sobre a produção científica brasileira sejam comparados com dados de outros países, apoiando análises transnacionais e estudos de colaboração científica internacional.

A partir de uma perspectiva computacional, a proposta se destaca pela combinação de diferentes camadas de representação: relacional, semântica e visual. A arquitetura modular e extensível permite a inclusão de novas entidades, fontes e relações conforme surgem novas necessidades e oportunidades. A escolha pela ontologia VIVO garante compatibilidade com sistemas como ORCID, Crossref, Wikidata e OpenAIRE.

Do ponto de vista dos estudos métricos, o BrCris amplia consideravelmente a capacidade de análise da produção científica brasileira. Torna-se possível, por exemplo, identificar padrões de colaboração internacional, avaliar a distribuição de pesquisadores por gênero, medir o impacto de políticas de fomento, mapear áreas emergentes e realizar análises longitudinais sobre produtividade e dispersão institucional.

Ademais, a exportação em formatos interoperáveis permite que os dados do BrCris alimentem diretamente outras ferramentas de análise e sistemas de gestão do conhecimento. Essa funcionalidade posiciona o BrCris como infraestrutura nacional para dados científicos, fomentando ecossistemas de reuso e inovação.

Considerando sua arquitetura e funcionalidades, o BrCris já pode ser considerado uma peça central da infraestrutura de dados científicos do Brasil. Sua adoção por instituições de ensino, pesquisa e agências de fomento tem o potencial de padronizar e qualificar os processos de avaliação da produção científica. Com isso, cria-se um ambiente mais transparente e eficaz para a formulação de políticas públicas e para o acompanhamento do impacto das pesquisas financiadas com recursos públicos.

5. Conclusões

O desenvolvimento da Plataforma BrCris representa uma importante contribuição para a integração de dados científicos no Brasil, oferecendo uma solução técnica robusta, escalável e aderente às melhores práticas internacionais. Ao reunir em um só ambiente dados historicamente dispersos, o BrCris viabiliza uma nova geração de estudos cientométricos, promovendo uma compreensão mais ampla e detalhada da ciência brasileira.

A consolidação do BrCris também tem implicações importantes no campo da ciência de dados aplicada à gestão da informação científica. Ao disponibilizar um repositório unificado, interoperável e atualizado, a plataforma facilita o uso de técnicas de mineração de dados, aprendizado de máquina e inteligência artificial para geração de insights estratégicos. Isso abre caminho para o desenvolvimento de *dashboards* inteligentes, sistemas de alerta para tendências emergentes e serviços automatizados de recomendação científica.

Como infraestrutura computacional, o BrCris atende às necessidades de diferentes perfis de usuários: pesquisadores, gestores institucionais, agências de fomento, jornalistas, tomadores de decisão e o público geral. Ao permitir a navegação visual, a exportação e o reuso dos dados, a plataforma consolida-se como pilar da Ciência Aberta e da gestão baseada em dados.

Em relação às atualizações da Plataforma BrCris, atualmente os dados são atualizados em períodos pré-definidos, nos quais é necessário realizar coletas dos conjuntos de dados utilizados. Espera-se que, no futuro, essas atualizações ocorram de forma automática, eliminando a necessidade de aguardar intervalos de tempo previamente estabelecidos.

Para o futuro, espera-se a expansão das fontes integradas, a inclusão de métricas altimétricas, a evolução da API REST e a incorporação de modelos preditivos e de aprendizado de máquina para identificar padrões emergentes. O BrCris não apenas sistematiza o passado da ciência brasileira, mas oferece instrumentos para planejar seu futuro.

Referências

Abbas, A.; Zhang, L.; Khan, S. U. A literature review on the state-of-the-art in patent analysis. World Patent Information, v. 37, p. 3-13, 2014.

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

Huang, Y.; Glanzel, W.; Zhang, L.. Tracing the development of mapping knowledge domains. Scientometrics, v. 126, p. 6201-6224, 2021.

Leta, Jacqueline; Glanzel, W.; Thijs, B.. Science in Brazil. Part 2: Sectoral and institutional research profiles. Scientometrics, v. 67, n. 1, p. 87-105, 2006.

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

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