Pecha kucha

A Brapci como ferramenta do fortalecimento da Ciência Aberta: inovações na organização e disseminação da informação

Brapci as a tool for strengthening Open Science: innovations in information organization and dissemination

Brapci como herramienta para el fortalecimiento de la Ciencia Abierta: innovaciones en la organización y difusión de la información

Rene Faustino Gabriel Junior*
Doutor em Ciência da Informação
Universidade Federal do Rio Grande do Sul
ORCID: 0000-0003-1021-3360
Lattes http://lattes.cnpq.br/5900345665779424
E-mail: rene.gabriel@ufrgs.br

Resumo

As bases de dados especializadas são fundamentais para a promoção da Ciência Aberta, especialmente na Ciência da Informação. A Brapci, inicialmente concebida como base de referências da produção científica em Biblioteconomia, evoluiu para uma plataforma de acesso aberto com foco na interoperabilidade e preservação digital. Este estudo descreve inovações tecnológicas adotadas na Brapci, como o uso de Inteligência Artificial (IA), os Modelos de Linguagem de Grande Escala (LLMs) e o modelo RAG, aplicados à organização e recuperação da informação. A pesquisa, de caráter descritivo, baseia-se em análise documental e estudo de caso. A Brapci automatizou 86 funções, desenvolveu APIs públicas, criou um repositório de PDFs (incluindo periódicos extintos) e implementou a padronização de nomes e estratégias de desambiguação. Consolidada como infraestrutura aberta e sustentável, a Brapci contribui para o acesso democrático à informação e para o avanço técnico-científico na área.

Palavras-chave: Bases de dados especializadas; Ciência Aberta; Modelos de Linguagem de Grande Escala; Inteligência Artificial.

Abstract

Specialized databases are essential for promoting Open Science, especially in the field of Information Science. Initially conceived as a reference database for scientific output in Librarianship, Brapci has evolved into an open-access platform focused on interoperability and digital preservation. This study describes the technological innovations adopted by Brapci, including the use of Artificial Intelligence (AI), Large Language Models (LLMs), and the Retrieval-Augmented Generation (RAG) model to enhance information organization and retrieval. The descriptive research is based on documentary analysis and case study. Brapci has automated 86 functions, developed public APIs, created a PDF repository (including defunct journals), and implemented name standardization and disambiguation strategies. Now established as an open and sustainable infrastructure, Brapci contributes to democratizing access to information and advancing technical and scientific development in the field.

Keywords: Specialized Databases; Open Science; Large Language Models; Artificial Intelligence.

Resumen

Las bases de datos especializadas son fundamentales para promover la Ciencia Abierta, especialmente en el ámbito de la Ciencia de la Información. Inicialmente concebida como una base de referencias de la producción científica en Bibliotecología, Brapci se ha transformado en una plataforma de acceso abierto centrada en la interoperabilidad y la preservación digital. Este estudio describe las innovaciones tecnológicas implementadas, como el uso de Inteligencia Artificial (IA), Modelos de Lenguaje de Gran Escala (LLMs) y el modelo RAG, aplicados a la organización y recuperación de la información. La investigación, de carácter descriptivo, se basa en el análisis documental y el estudio de caso. Brapci ha automatizado 86 funciones, desarrollado APIs públicas, creado un repositorio de PDFs (incluyendo revistas extintas) y aplicado estrategias de estandarización de nombres y desambiguación. Consolidada como infraestructura abierta y sostenible, Brapci contribuye a democratizar el acceso a la información y al avance técnico-científico en el área.

Palabras clave: Bases de Datos Especializadas; Ciencia Abierta; Modelos de Lenguaje de Gran Escala; Inteligencia Artificial.

Introdução

As bases de dados especializadas desempenham papel relevante na promoção da Ciência Aberta, especialmente no contexto profissional de pesquisa, estudos e aplicações em Ciência da Informação. Essas plataformas facilitam o acesso e a disseminação do conhecimento científico, promovendo a interoperabilidade e a exportação de dados, em conformidade com os princípios FAIR (*Findability, Accessibility, Interoperability, and Reusability*).

Com base na literatura, foram identificadas quatro bases de dados voltadas para a Ciência da Informação: a Brapci, que reúne a produção científica brasileira e latino-americana, com ênfase em artigos, periódicos e eventos acadêmicos, destacando-se pela interoperabilidade via OAI-PMH e APIs, desenvolvida na UFRGS; a base ABCDM, voltada à produção nacional, especialmente eventos e artigos, embora com baixa interoperabilidade, mantida pela UnB; a E-LIS (Eprints in Library and Information

Science), de caráter internacional, que permite o autoarquivamento de preprints e artigos na área, com ampla interoperabilidade e cobertura linguística; e a BDA (Biblioteca Digital de Arquivologia), que indexa documentos relevantes da arquivologia no Brasil, destacando-se pela adoção de padrões abertos de metadados via OAI-PMH, também sob responsabilidade da UnB. A Tabela 1 apresenta um quadro-resumo das bases identificadas.

Tabela 1 - Caracterização comparativa de bases de dados especializadas em Ciência da Informação

BASE DE DADOS	COBERTURA GEOGRÁFICA	FOCO TEMÁTICO	TIPO DE CONTEÚDO	INTEROPERABILIDADE	IDIOMA PREDOMINANTE	TAMANHO INDEXADO
BRAPCI	Brasil e América Latina	Ciência da Informação	Artigos, eventos, livros	Moderada (OAI-PMH e APIs	Português, Espanhol, Inglês	≈ 75.000 registros
BASE ABCDM	Brasil	Ciência da Informação	Artigos, eventos	-	Português	
E-LIS	Internacional	Biblioteconomia e Ciência da Informação	Preprints, artigos	Alta (self-deposit, OAI-PMH)	Inglês, Espanhol	≈ 25.000 registros
BDA	Brasil	Arquivologia	Artigos, eventos, livros	Alta (OAI-PMH)	Português	≈ 10.000 registros

Fonte: Autor (2025).

Diante do número reduzido de bases especializadas em Ciência da Informação e do fato de que, em sua maioria, são desenvolvidas em contextos acadêmicos com recursos orçamentários limitados, torna-se necessário otimizar os meios disponíveis. No caso da Brapci, essa necessidade se traduz na adoção de soluções automatizadas, no uso de ferramentas de código aberto e na incorporação de tecnologias como a programação orientada a regras e o uso de Inteligência Artificial. Tais estratégias visam garantir a sustentabilidade da plataforma, promover sua atualização contínua e ampliar sua cobertura, mantendo o compromisso com a qualidade na organização e disseminação da informação científica.

A Base de Dados em Ciência da Informação (Brapci) foi criada com o objetivo de reunir, em um único catálogo, a produção científica brasileira na área de Biblioteconomia. Inicialmente, sua proposta era funcionar como uma base de referências, dedicada apenas à catalogação de metadados e à geração de indicadores sobre a produção nacional, sem a intenção de oferecer acesso ao conteúdo completo dos documentos.

A Base de Dados em Ciência da Informação (Brapci)

Atualmente, a base de dados reúne aproximadamente 75.000 trabalhos indexados, provenientes de 79 revistas brasileiras e 25 revistas estrangeiras em língua portuguesa

e espanhola. Também disponibiliza a produção de cinco eventos brasileiros de relevância para a área da Ciência da Informação. Recentemente, a Brapci iniciou a indexação de livros e capítulos, consolidando-se como uma base unificada de consulta.

A Brapci atua como um laboratório de práticas em organização do conhecimento, promovendo o desenvolvimento de novas metodologias de organização, armazenamento e disseminação da informação, com foco em profissionais da informação, em pesquisadores, em estudantes e na sociedade em geral. Considerada uma iniciativa inovadora no contexto da Ciência Aberta (Bufrem et al., 2010), a Brapci destaca-se por promover o acesso gratuito e amplo à produção científica brasileira na área da Ciência da Informação, contribuindo significativamente para a democratização do conhecimento. Essa característica está diretamente alinhada aos princípios da Ciência Aberta, que preconizam o livre acesso à informação científica.

Além disso, a Brapci adota padrões de metadados que favorecem a interoperabilidade com outras plataformas (Arakaki & Arakaki, 2021), possibilitando o intercâmbio de informações e fortalecendo ecossistemas colaborativos de dados abertos. Projetos recentes voltados à aplicação de inteligência artificial na indexação automática de documentos reforçam seu caráter inovador, ao incorporar tecnologias que otimizam os processos de organização e recuperação da informação.

Por ser uma base de dados mantida por uma instituição acadêmica e sem fins comerciais, a Brapci exige a implementação de diversas automatizações que garantam sua sustentabilidade, manutenção e, sobretudo, sua constante atualização. Nesse contexto, configura-se como um laboratório de experimentação, ao proporcionar a estudantes de graduação e pós-graduação a oportunidade de desenvolver projetos inovadores voltados à melhoria dos processos de coleta, seleção, organização e recuperação da informação, contribuindo para o avanço técnico e científico da área.

Este estudo tem como objetivo descrever as inovações desenvolvidas na Brapci para promover a interoperabilidade e a organização eficiente da informação em repositórios digitais.

Metodologia

Trata-se de uma pesquisa descritiva que analisa as estratégias e tecnologias implementadas, incluindo a adoção de padrões de metadados, o desenvolvimento de APIs e o uso de ferramentas de inteligência artificial para a padronização de dados. A investigação será conduzida por meio de análise documental e estudo de caso, com o intuito de compreender de que modo essas inovações contribuem para aprimorar a recuperação e a disseminação da informação científica.

Resultados

Para aprimorar a estrutura de representação e organização dos dados, a Brapci desenvolveu, por meio da ferramenta Protégé, um modelo ontológico fundamentado na integração de diferentes vocabulários e padrões da web semântica, incluindo a Bibliographic Ontology (BIBO), o *Simple Knowledge Organization System* (SKOS), o *Friend of a Friend* (FOAF) e os modelos conceituais da IFLA, como o FRBR e sua evolução, o IFLA LRM (*Library Reference Model*).

Esse modelo ontológico é utilizado como base para a validação semântica das classes e propriedades associadas aos dados bibliográficos, permitindo a padronização e o enriquecimento das informações. A adoção dessa abordagem promove maior consistência, interoperabilidade e qualidade na organização dos registros, ao mesmo tempo em que facilita a integração com outras plataformas e iniciativas de dados abertos.

Todos os dados são descritos em conformidade com o padrão RDF (Resource Description Framework), o que permite a estruturação dos metadados em formato legível por máquina, viabilizando consultas semânticas via SPARQL e contribuindo para a construção de uma base de conhecimento vinculada ao ecossistema do Linked Data. Esse avanço consolida a Brapci como um ambiente de experimentação em ciência aberta, capaz de explorar tecnologias emergentes para a organização, disseminação e reutilização da informação científica.

Na versão atual da Brapci, 86 funções estão automatizadas por meio de scripts que são executados diversas vezes ao longo do dia, desempenhando tarefas como indexação e revisão de dados. Entre essas funcionalidades, destaca-se o controle de autoridades, que emprega variantes nominais para identificar corretamente os autores. Contudo, o uso de deduplicadores automatizados, como o algoritmo de Levenshtein, que mede o grau de similaridade entre cadeias de caracteres, revela-se inadequado, pois a base inclui nomes muito parecidos que pertencem a pessoas distintas.

Um exemplo de ambiguidade ocorre com o nome do Prof. Rene Faustino Gabriel Junior, da Universidade Federal do Rio Grande do Sul, que possui um homônimo, Renê Gabriel Junior, mestre em Gestão Pública e auditor fiscal no estado do Espírito Santo. Em sistemas baseados apenas em similaridade textual, há alta probabilidade de que ambos sejam identificados como a mesma pessoa, sobretudo quando nomes intermediários estão ausentes nos registros de autoria. No entanto, trata-se de indivíduos distintos.

Para mitigar esse tipo de erro, adotou-se o método proposto por Castanha e Silveira (2024), baseado na criação de grupos de acoplamento por coautoria. Essa estratégia organiza os autores em agrupamentos, permitindo que a desambiguação seja realizada manualmente dentro de cada grupo. Em um universo com mais de 25.000 autores, tais procedimentos são fundamentais para garantir a qualidade e a confiabilidade da base de dados.

No contexto da padronização dos nomes de autores na Brapci, foi necessário implementar ajustes para acomodar diferentes convenções de nomenclatura, especialmente após a inclusão de periódicos latino-americanos. Tradicionalmente, os autores eram apresentados pelo sobrenome seguido do nome, conforme as diretrizes do RDA (*Resource Description and Access*) (Resource ..., 2012). No entanto, em países hispânicos é comum o uso dos sobrenomes paterno e materno, o que pode gerar inconsistências na indexação.

Para resolver essa questão, optou-se por uniformizar os registros, apresentando os autores pelo primeiro nome seguido dos sobrenomes. Essa decisão exigiu a atualização dos registros existentes na base. Para automatizar o processo, foi desenvolvida uma função de padronização que realiza as seguintes operações: converte os nomes para letras minúsculas; capitaliza a primeira letra de cada nome e sobrenome; e mantém preposições em letras minúsculas. Além disso, foi criada uma API gratuita que oferece essa funcionalidade à comunidade, integrando-se à lista de serviços disponibilizados pela Brapci (Brapci, 2025). Essa função é executada diariamente, padronizando os nomes de novos autores incorporados à lista de autoridades.

Desde 2018, a Brapci mantém um repositório interno dedicado à coleta e preservação dos arquivos PDF dos documentos indexados, com o objetivo de salvaguardar a memória da Ciência da Informação. Essa iniciativa é particularmente relevante, considerando que periódicos extintos, como a DataGramaZero (DGZ) e a Revista do CRB-8, atualmente estão acessíveis apenas por meio da Brapci. Como parte desse esforço de preservação, também são realizadas buscas em bibliotecas para recuperar publicações disponíveis apenas em formato impresso, que são digitalizadas e incorporadas ao acervo. Um exemplo é a Revista de Biblioteconomia e Comunicação, predecessora da Em Questão.

Atualmente, a equipe da Brapci, composta por alunos de graduação e pós-graduação, está trabalhando na aplicação de ferramentas de Inteligência Artificial (IA), em especial por meio de modelos de linguagem de grande escala (*Large Language Models* – LLM) (CANELAS-PAIS, 2023), com ênfase no uso do Retrieval-Augmented Generation (RAG) ou Geração Aumentada por Recuperação (Tavares, et al., 2024). O objetivo é desenvolver, com o uso de softwares gratuitos, soluções que contribuam para a melhoria da qualidade da organização e indexação.

Essa necessidade decorre, sobretudo, do fato de que a maioria das revistas não exige o uso de vocabulários controlados na escolha das palavras-chave pelos autores e, mesmo quando essa exigência existe, os vocabulários disponíveis carecem de especificidade temática. Nesse contexto, a aplicação de Inteligência Artificial, por meio da combinação entre modelos de linguagem de larga escala (LLM) e técnicas de geração aumentada por recuperação (RAG), tem sido explorada para identificar conceitos mais especializados e propor a criação de um novo conjunto de metadados: os conceitos-chave.

Dado que na conversão de arquivos PDF para texto geralmente ocorrem problemas de formatação, foi necessário recorrer à literatura especializada em busca de uma ferramenta que permitisse estruturar o conteúdo do PDF em formatos como JSON ou Markdown, com agrupamento de parágrafos e separação adequada das seções. Para resolver esse problema, utilizou-se a biblioteca Docling (https://pypi.org/project/docling/), desenvolvida em Python, que organiza o texto de forma a viabilizar sua leitura e seu processamento, sendo especialmente útil na extração de referências, resumos e palavras-chave.

A base também se destaca por valorizar a produção científica nacional e regional, dando visibilidade a periódicos e eventos muitas vezes ignorados por bases internacionais. Isso contribui para a equidade no acesso à informação e para a valorização de saberes locais. Além disso, ao oferecer possibilidades para análises bibliométricas e cientométricas, a Brapci apoia práticas de avaliação científica mais transparentes e participativas, promovendo a responsabilização e a reprodutibilidade, aspectos fundamentais da Ciência Aberta.

Considerações finais

A análise das bases de dados especializadas em Ciência da Informação evidencia a importância dessas plataformas na promoção da Ciência Aberta e na democratização do acesso à informação científica. Em especial, a Brapci se destaca não apenas por consolidar a produção científica brasileira e latino-americana, mas também por adotar estratégias inovadoras que visam otimizar processos e ampliar sua relevância. A utilização de soluções automatizadas e de ferramentas de código aberto, além da integração de tecnologias avançadas – como a programação orientada a regras e a aplicação de modelos de linguagem de grande escala (LLM) combinados com o Retrieval-Augmented Generation (RAG) –, demonstram um compromisso robusto com a sustentabilidade, a atualização contínua e a qualidade dos registros indexados.

Uma iniciativa desse tipo reforça a importância de parcerias entre o meio acadêmico e as práticas de Ciência Aberta, promovendo uma gestão colaborativa dos saberes científicos. Nesse sentido, a Brapci se consolida como uma ferramenta estratégica para o fortalecimento da Ciência Aberta no Brasil e na América Latina, atuando como um elo entre a produção e as novas práticas de circulação e uso da informação no ambiente digital.

Agradecimento

Este trabalho contou com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, por meio da Bolsa de Produtividade em Pesquisa (Processo nº 312975/2022-8), o que tem contribuído significativamente para o desenvolvimento desta linha de investigação.

Conflito de Interesses

Os autores declaram não haver conflitos de interesses.

Disponibilização dos Dados de Investigação

Os dados de pesquisa estão disponíveis em Hdl: 20.500.11959/brapci/dataset_dump

CRediT – Contribuições dos Autores

Rene Faustino Gabriel Junior | Concetualização, Curadoria de dados, Escrita

Referências

BRAPCI. (2025). Lista de API da Base de Dados em Ciência da Informação. https://brapci.inf.br/api

Arakaki,F. A., & Arakaki, A. C. S. (2021). Metadados e tipos de metadados: conceitos, categorias e relações. Inclusão Social, 14(2). https://doi.org/10.18225/inc.soc.v14i2.6407

Bufrem, L. S., Costa, F. D. O., Gabriel Junior, R. F., & Pinto, J. S. P. (2010). Modelizando práticas para a socialização de informações: A construção de saberes no ensino superior. Perspectivas em Ciência da Informação, 15(2).

Canela-Pais, M., Silva, R. F. Da, Duarte, I., Rodrigues, P. P., & Cruz-Correia, R. (2023). Integration of artificial intelligence in medical education: Developing a framework for curriculum enhancement. Revista Fontes Documentais, 6, 70–72.

Castanha, R. G., & Silveira, R. C. (2024, julho). Acoplamento bibliográfico entre autores enquanto critério adicional para desambiguação de nomes de pesquisadores: Primeiras aproximações, IX Encontro Brasileiro de Bibliometria e Cientometria, Brasília, Brasil. https://doi.org/10.22477/ix.ebbc.236

Resource Description and Access (RDA). (2012). In RDA Toolkit. American Library Association; Canadian Library Association; Chartered Institute of Library and Information Professionals. http://access.rdatoolkit.org

Tavares, H. L., Coneglian, C. S., Torino, E, Vidotti, S. A. B. G, & Santarem Segundo, J. E. (2024, 4-8 de novembro). Recuperação da informação e inteligência artificial generativa com large language model e retrieval-augmented generation, XXIV Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação, Vitória, Brasil, 1–6. https://enancib.ancib.org/index.php/enancib/xxivenancib/paper/viewFile/2690/1644