8º EBBC
Encontro Brasileiro de Bibliometria e Cientometria
MACEIÓ2022
20, 21 E 22 · JULHO

O PROJETO BRCRIS:
uma plataforma computacional para integração, visualização e prospecção de dados científicos

Thiago Magela Rodrigues Dias[1]
Jesus Pascual Mena Chalco[2]
Washington Luís R. de Carvalho Segundo[3]
Adilson Luiz Pinto[4]
Luc Quoniam[5]
Tales Henrique José Moreira[3]
Viviam dos Santos Silva[3]

Resumo: A disponibilização de dados sobre a produção científica nacional e internacional tem crescido expressivamente e, em perspectiva às especificidades de campos disciplinares distintos, esta produção se revela diversa quanto à sua tipificação, tanto em termos quantitativos como qualitativos. Neste trabalho apresentamos o processo de desenvolvimento da Plataforma BrCris cujo objetivo é de fornecer ferramentas tecnológicas para munir à comunidade acadêmica com dados consolidados da produção científica do Brasil. Esta iniciativa se apresenta como mecanismo ímpar de agregação de dados, possibilitando visualizações e análises cientométricas sobre o presente, passado e futuro da produção científica brasileira.

Palavras-Chave: BrCris. Recuperação de Informação. Dados científicos/tecnológicos.

1 INTRODUÇÃO

O ecossistema da pesquisa envolve a presença de diversos atores que interagem entre si. Desde o financiamento obtido por meio de um projeto de pesquisa, passando pela figura do próprio pesquisador, que faz uso de infraestrutura para efetuar seus trabalhos, tais como laboratórios e equipamentos físicos (LEE; BOZEMAN, 2005). Os pesquisadores, por sua vez, estão associados e são mantidos por Instituições onde as pesquisas são desenvolvidas. Uma das principais contribuições do trabalho de um pesquisador é a produção de conhecimento, expressa em termos de artigos científicos que comumente são indexados nos repositórios acadêmicos como por exemplo Web of Science, Scopus, e Dimensions (SINGH *et al.*, 2021).

[1] Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
[2] Universidade Federal do ABC (UFABC)
[3] Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
[4] Universidade Federal de Santa Catarina (UFSC)
[5] Universidade Federal de Mato Grosso do Sul (UFMS)

Há ainda, produtos de pesquisa que geralmente não são visíveis nos repositórios acadêmicos tradicionais, como, por exemplo os produtos associados à literatura cinzenta representada por teses, dissertações, relatórios técnicos, banco de dados com resultados de experimentos, manuais, protocolos (PAEZ, 2017). Aqui é importante frisar que esses produtos, não tradicionais, são tão importantes quanto os artigos científicos dado que permitem observar a produção de todo o ecossistema de pesquisa (RAMSDEN, 1994).

Para o campo da Ciência da Informação, e em especial da Cientometria, quantificar essa produção é uma tarefa árdua pois a disponibilização de bases de dados abertas muitas vezes é restrita ou simplesmente inexistente em determinados contextos. Bases de dados proprietárias como a Scopus, Web of Science, Google Acadêmico e Microsoft Research Data permitem o acesso mas este é sempre limitado ao número de registros que podem ser obtidos, contemplam poucos repositórios e periódicos nacionais e ainda existe o grave problema da opacidade dos algoritmos utilizados por estas plataformas que determinam o que é ou não relevante.

A partir desse cenário, começaram a surgir iniciativas que visavam a criação de sistemas que gerenciam a produção acadêmica de uma instituição, país ou área de conhecimento. Tais sistemas são conhecidos pela sigla CRIS (Current Research Information Systems) e têm como objetivo agregar informações de bases de dados diversas com intuito de fornecer relatórios e dados consolidados para que pesquisadores da área possam analisar como se dá a produção em seus países ou áreas de conhecimento.

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 contexto, a idealização do projeto do Sistema BrCris (PINTO *et al., 2021*), que é o CRIS no contexto da Ciência Brasileira, data de 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 de um projeto de pesquisa para a construção do BrCris. O intuito é fornecer ferramentas

tecnológicas visando munir dados consolidados da produção científica brasileira para toda a comunidade acadêmica.

Logo, o BrCris tem por objetivo estabelecer um modelo único de organização da informação científica de todo o ecossistema da pesquisa brasileira. 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).

2 DESENVOLVIMENTO

O BrCris concentra um amplo ecossistema de dados, de diversas fontes, como por exemplo, dados curriculares de indivíduos, sobre organizações, programas de pós-graduação, publicações, orientações acadêmicas, revistas científicas, dentre outros, sendo necessário todo um esforço para tratamento dos dados de interesse. Neste contexto, tendo em vista as diversas fontes de dados que irão compor o BrCris se faz necessário a transformação dos dados em formato padronizado. Em se tratando do modelo de dados do BrCris, iniciou-se pela adoção de nove entidades de dados, seguindo padrões amplamente utilizados na comunidade científica internacional:

- **Project**: projetos de pesquisa executados, ou em execução;
- **Service**: revistas científicas, repositórios digitais, bibliotecas digitais e outras fontes de informação científica;
- **Program**: programas de pós-graduação brasileiros;
- **Course**: cursos de pós-graduação stricto ou lato sensu;
- **OrgUnit**: instituições, faculdades, departamentos de pesquisa;
- **Person**: pesquisadores, assistentes de pesquisa e de apoio técnico à pesquisa;
- **Patent**: patentes como resultado da pesquisa;
- **Dataset**: conjuntos de dados de pesquisa de um projeto ou pesquisa científica;
- **Publication**: artigos científicos, teses, dissertações, livros, capítulos de livro.

O modelo de dados é definido por um conjunto de entidades e relações, que por sua vez possuem identificadores e atributos pré-definidos. A utilização de um descritivo visa facilitar a identificação de atributos de cada entidade e suas relações, possibilitando que o modelo

possa incorporar todas as mudanças realizadas diretamente no modelo. Esta estratégia visa facilitar de forma significativa a incorporação de novos atributos e relações, sem a necessidade de alterações diretamente no modelo de dados.

Para o tratamento dos dados foi desenvolvida uma biblioteca computacional contendo uma estrutura de dados preparada para facilitar o processamento de dados originários de todas as fontes para o formato exigido pela plataforma LA Referencia. Nesse sentido, a biblioteca desenvolvida é responsável por toda a transformação e exportação dos dados, utilizando como base o "Modelo de Dados" da plataforma LA Referencia, validando as entidades, campos e relacionamentos aceitos pelo modelo.

3 RESULTADOS

O principal repositório de dados para o BrCris são os currículos cadastrados na Plataforma Lattes do CNPq. Acessível em < http://lattes.cnpq.br/ >, a Plataforma Lattes foi criada e é mantida pelo CNPq, contando atualmente com mais de 7,4 milhões de currículos cadastrados, além de grupos de pesquisa e diretórios de instituições.

Além dos dados curriculares da Plataforma Lattes que subsidiaram informações para as entidades Person, OrgUnit, Publication, Patent, Event, Program, Course e Service, também são integrados dados dos seguintes repositórios:

- **OasisBR**: mantido pelo IBICT, fornece dados confiáveis sobre publicações científicas em acesso aberto. Os dados foram mapeados para as entidades Publication, Service, Person.
- **BDTD**: a exemplo do OasisBR, a BDTD também é mantida pelo IBICT. Fornece dados confiáveis sobre teses e dissertações brasileiras. Os dados foram mapeados para as entidades Publication, Course e Person.
- **Plataforma Sucupira**: concentra dados dos Programas de Pós-graduação do Brasil, fornecendo um conjunto de informações sobre os programas e cursos de pós-graduação. Todos os dados dos programas foram mapeados para as entidades Program , Course e OrgUnit.
- **Instituições do INEP**: assim como os programas de pós-graduação da Plataforma Sucupira, o INEP fornece uma base confiável, sobre as instituições de

ensino do país em outros níveis de capacitação, sendo mapeados para a entidade OrgUnit.

- **Dados Abertos da CAPES**: fornece dados como publicações, orientações acadêmicas, entre outros que são mapeados para diversas entidades, como Person, OrgUnit, Publication, Program e Course.

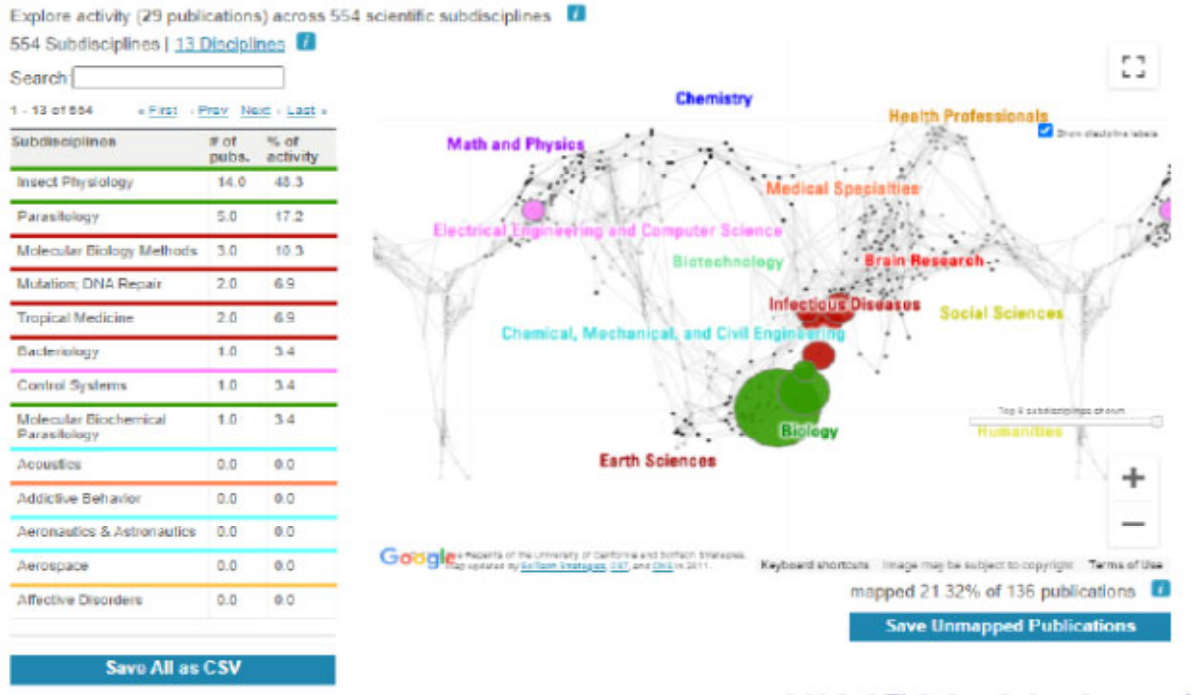
- **Revistas Científicas**: o processamento de dados do conjunto de revistas científicas fornece informações diversas sobre elas, sendo mapeadas para a entidade Service.

Como pode ser observado, as diversas fontes de dados mapeadas, se completam, possibilitando a criação de um conjunto padronizado e consistente, validado através de dados provenientes de diversas entidades brasileiras amplamente consolidadas e utilizadas. Ao se agregar todos os repositórios apresentados, é possível a adoção de técnicas que visam permitir a vinculação de conjuntos que inicialmente não era possíveis de se comunicarem, possibilitando dessa forma, a construção de um grande conjunto de dados, interligados, que facilitam a aplicação de consultas que inicialmente não seriam possíveis.

Os resultados da execução do projeto já incluem o desenvolvimento da arquitetura do BrCris, o mapeamento das fontes de dados a serem agregadas pelo Sistema, a implementação de provas de agregação dos recursos mapeados, a definição e realização de testes de serviços a serem disponibilizados.

Diversas visualizações já estão implementadas o que viabiliza obter retratos da produção científica nacional de forma inédita. Tais visualizações possibilitam a utilização de filtros e outros métodos para agregar elementos de personalização como por exemplo análises temporais ou por áreas (ver Figura 1).

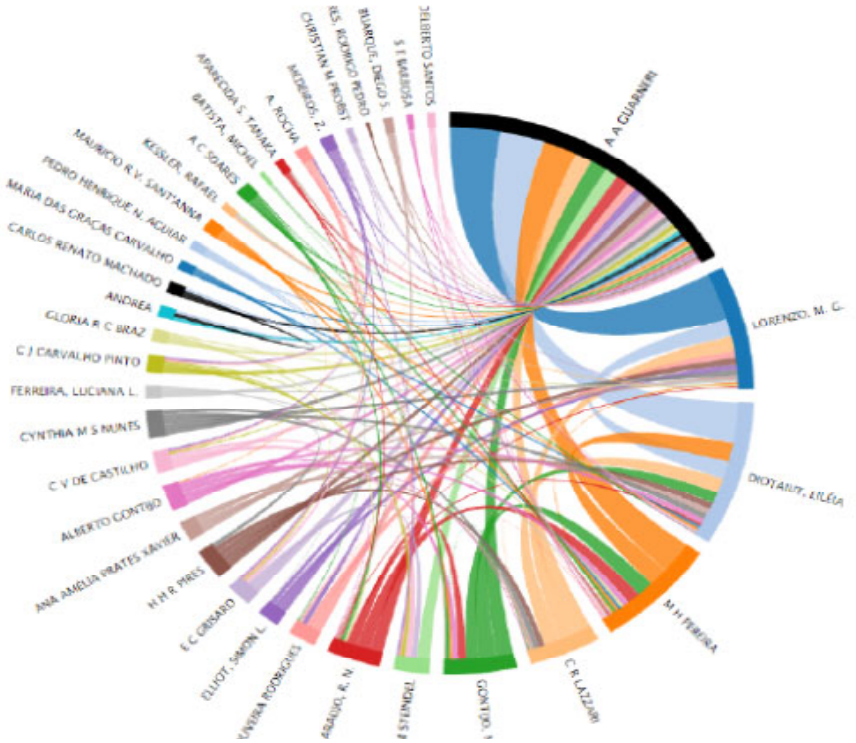
Figura 1 – Dashboard para Análises da Produção Científica



Fonte: Os autores.

Também é possível realizar a visualização de redes de colaboração o que viabiliza diversos tipos de análise sobre como tem ocorrido o processo de co-autoria nas diversas áreas do conhecimento (Figura 2).

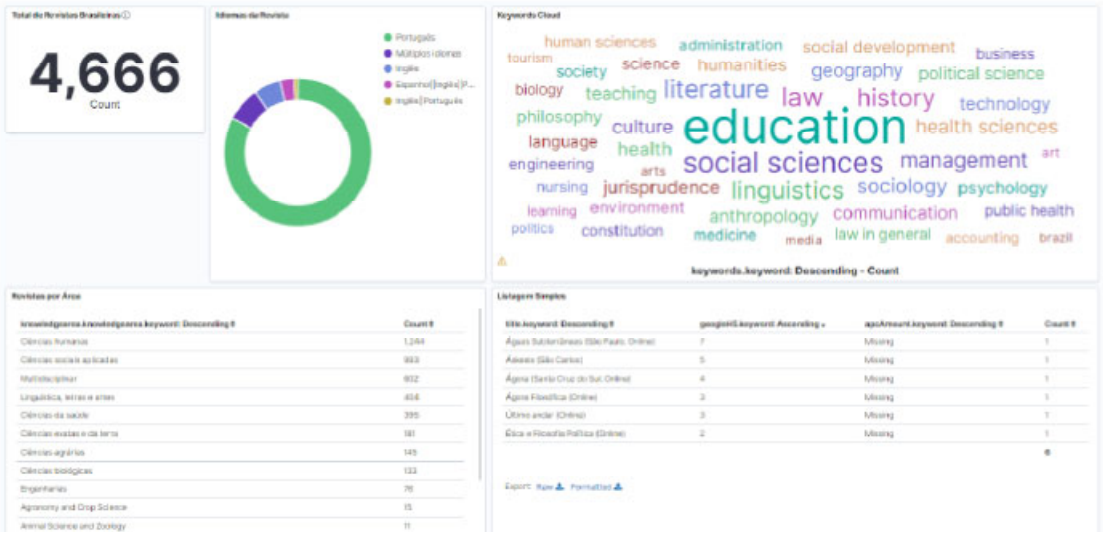
Figura 2 – Grafo de Colaboração Científica



Fonte: Os autores.

Além disso visalizações que integram grandes conjuntos de dados também já estão disposíneis como por exemplo, de periódicos científicos (Figura 3).

Figura 3 – Dashboard para Análises sobre Periódicos Científicos



Fonte: Os autores.

Todas as visualizações podem ser manipuladas de forma flexível e ainda, com a possibilidade de exportação dos dados em formatos tabulares ou consumidos por API´s.

4 CONSIDERAÇÕES FINAIS

O BrCris é um importante espaço para 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, que pretende ser o alvo da proposta apresentada. O BrCris é uma iniciativa que coleta e enriquece dados de repositórios e bases de dados abertas sendo uma proposta, impar no mundo, que facilita obter um Panorama Brasileiro da Produção e Atuação de todos os seus atores acadêmicos/científicos.

REFERÊNCIAS

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.

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.

PAEZ, A. Gray literature: An important resource in systematic reviews. **Journal of Evidence-Based Medicine**, Richmond. v. 10, n. 3, p. 233–240, 2017.

RAMSDEN, P. Describing and explaining research productivity. **Higher Education**, Dordrecht, v. 28, n. 2, p. 207–226, 1994.

PINTO, A. L. *et al.* BrCris como um sistema de recomendação cintífico-tecnológica. *In*: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - ENANCIB, 21., 2021, Rio de Janeiro. **Anais** [...]. Rio de Janeiro: [*s.n.*], 2021.

SINGH, V. K. *et al.* The journal coverage of web of science, scopus and dimensions: A comparative analysis. **Scientometrics**, Budapest, p. 1–30, 2021.

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, 2019. p. 667-683.