

BrCris: A Catalyst for Advancing Open Science and Scientific Data Processing in Brazil

Authors: Marcel Souza, Washington Segundo, Vivian Silva, Rene Gabriel Junior, Thiago Dias and Alexandre Oliveira.

Abstract

Brazil currently has several institutional platforms, such as the Lattes Platform, the Sucupira Platform, the Transparency Portal, the Theses and Dissertations Catalog/Library and the Industrial Property Bank. These platforms were created with the aim of providing centralized and up-to-date data on the academic, scientific and technological activities of their researchers. In this sense, BrCris aims to establish a single model for organizing scientific information throughout the Brazilian research ecosystem, in line with the FAIR (Findable, Accessible, Interoperable and Reusable) principles. Among the entities in this ecosystem are researchers, projects, infrastructures, laboratories, research institutions, funders and research results expressed mainly through scientific publications, theses, dissertations, scientific datasets, software and patents. In addition, BrCris has contributed to the process of aggregating data platforms and repositories, with the aim of promoting access to and evaluation of science. This work significantly highlights the importance of the BrCris Platform as an indispensable resource for scientific research in Brazil, acting as a catalyst for Open Science and driving the continuous advancement of knowledge.

Introduction

The development of knowledge in science occurs slowly and progressively (HUANG et al., 2021). Often, identifying the current state of knowledge in a specific area is done through a thorough analysis of specialized publications, sources that add important information related to the topic under discussion. These publications, usually from renowned researchers and academic institutions, provide a comprehensive and up-to-date view of the academic landscape, exploring recent discoveries, technological advances and emerging trends in the field of study.

The production of scientific knowledge, a gradual and incremental process, requires a detailed understanding of the state of the art in different areas. Traditionally, this understanding derives from a meticulous analysis of specialized publications, scientific collaborations, as well as patent records, among other sources. A Plataforma BrCris surge em resposta a essa demanda, com o objetivo de integrar dados de pesquisadores, projetos, instituições e resultados de pesquisas em um sistema abrangente conhecido como Current Research Information System (CRIS).

CRIS defines an information system that covers the entire ecosystem of the scientific process. All the information relating to the Scientific Research cycle is organized in one place, from funding, through projects, researchers, research institutions and laboratories, to the products of scientific research, such as scientific articles, theses, dissertations, books, book chapters, patents and scientific datasets (SIVERTSEN, 2019).

The work led by the Brazilian Institute of Information on Science and Technology (IBICT) to gather information on scientific production at national level has yielded concrete results. The BrCris platform materializes the efforts to create the Brazilian research ecosystem, a unified

database that aims to provide the academic community with consolidated and easily accessible data on national scientific production.

The BrCris project makes this information ecosystem possible by collecting, processing and integrating heterogeneous data from different sources into a consolidated database, creating homogeneous data sets that can be explored using different visualization and analysis tools. The BrCris platform brings together several of these tools in an integrated way, allowing data search, navigation and visual analysis in a single environment.

By bringing all these elements together in a unified and organized system, BrCris promotes integration, accessibility and interoperability between the various players in the Brazilian research ecosystem. This not only facilitates collaboration and the exchange of knowledge, but also boosts scientific progress in all areas of knowledge.

BrCris' contributions to Open Science and its alignment with the FAIR principles are fundamental to strengthening mapping, transparency and accessibility in the Brazilian scientific landscape, as well as promoting the effective use of open data.

By adopting the FAIR principles, the BrCris platform ensures that scientific data is easily located, accessible, interoperable between different systems and, finally, can be reused for future research.

This approach not only promotes openness and data sharing, but also increases the visibility of research stakeholders and collaboration between them. By strengthening Open Science, BrCris reinforces its commitment to the democratization of knowledge, contributing significantly to building a more dynamic, collaborative and impactful scientific community.

Development

BrCris aims to establish a single model for organizing scientific information from the entire Brazilian research ecosystem. The players in this ecosystem include researchers, projects, infrastructures, research groups, laboratories and research institutions, funders, as well as the results of scientific and technological production.

The consolidation of data in a standardized repository, combined with the use of dashboards and the application of filters for visualization, has several explicit benefits. Integrating data simplifies its complexity, making it easier to recognize its intrinsic properties. It also helps to identify variations and detect obvious errors. This approach makes it possible to condense information into more compact formats, which in turn increases the capacity for cognitive understanding, among other benefits.

With regard to the data model adopted by BrCris, the approach began with the incorporation of nine data entities aligned with widely recognized standards in the international scientific community. This choice aims to ensure consistency and interoperability, facilitating the integration of information from different sources. The transformation process therefore represents a crucial step in the harmonization of heterogeneous data, promoting a solid foundation for the construction and effective use of BrCris.

BrCris adopts a two-level data representation scheme. The first level consists of the logical level, materialized as an entity and relationship model based on CERIF (Jörg, 2010), also called a metamodel. This model is then translated into a physical relational schema on the LA Referencia platform, where the data is loaded and processed. This metamodel, known as a template, meets

the internal demands of data storage, playing a crucial role in organizing and integrating the information collected, transforming it into inputs for the project's outlined objectives (Pinto et al., 2021).

One of the biggest challenges in consolidating this data is finding viable computational strategies so that the various data sources to be collected can be integrated in a consistent way, without ambiguity in the records. There are several other initiatives in the literature that integrate data sets from different sources, such as the OpenAIRE Research Graph. In this context, adopting best practices to properly integrate these repositories becomes a significant challenge.

After the entire collection and processing process, regardless of origin, the data generated as output is aggregated into a single database, and using and/or generating unique identifiers, the datasets are linked and de-duplicated, thus allowing for data interoperability, regardless of its origin and format.

In addition, adherence to international standards for data modeling and representation is one of the keys to enabling effective information sharing. BrCris adopts a two-level data representation scheme. The first is the logical level, materialized as an Entity-Relationship (ER) model based on CERIF (JÖRG, 2010). The second is the semantic level, implemented as an ontology to allow data to be visualized and navigated as a knowledge graph.

The use of FAIR (Findable, Accessible, Interoperable, Reusable) principles plays an important role in the management and sharing of data in Scientific and Research Information (CRIS) repositories.

Applying these principles ensures that the data stored in CRIS repositories is easily locatable, accessible, interoperable and reusable. This is achieved by assigning unique identifiers, rich metadata and interoperability standards, thus facilitating data discovery by researchers, machines and algorithms, promoting a more open and collaborative science.

The FAIR principles facilitate reuse by promoting practices that guarantee its discovery and effective use by different users and systems. In the context of CRIS repositories such as BrCris, their adherence to the FAIR principles allows research data to be integrated more efficiently with other databases and analytical tools, with as little human interference as possible, expanding their potential for use in different scientific, academic and social contexts. Following these principles not only improves the quality and efficiency of research, allowing scientists to access and reuse pre-existing data, but also increases the visibility and impact of research. In this way, CRIS repositories, by adhering to the FAIR principles, make a significant contribution to transparency, innovation and progress in the field of scientific research.

In BrCris there are two types of metadata aligned with the FAIR principles, one on descriptive metadata for providing correlated information on the context, quality, condition and characteristics of the data, such as its classes, facilitating its identification and discovery. This includes data such as Publications, People, Journals, Organizations, Patents, Postgraduate Programs, Research Groups and Software (registries). All this information is represented by the second type of metadata, structural metadata, which ensures that the data can be integrated with other data by means of identifiers and represented by the Vivo ontology. This ontology was chosen due to its recognized international projection and because, by adopting its technical specifications, the data can be made compatible and interoperable with other data sets.

Results

The platform makes a difference with its ability to export subsets of data in standardized formats, such as CSV and RDF files. This feature not only highlights the platform's versatility, but also expands its possibilities of use. By exporting data in these formats, users benefit from easy integration with various other data analysis and visualization tools. The variety of formats available, especially CSV and RDF files, provides a flexible experience, allowing users to explore the data in greater depth and adapt it to their specific needs, as well as correlate this data in Linked data.

Moreover, the relevance of this functionality goes beyond mere export, as it is a significant means of strengthening the interoperability and practical usefulness of the data. The ability to export subsets in standardized formats not only facilitates the incorporation of this data into other tools, but also promotes a personalized approach to analysis, ensuring that users can extract results that are relevant to their respective areas of research or interests. This robust functionality reinforces the platform as an essential tool for the scientific community and other users involved in analyzing scientific data.

However, integrating various sources of information into a single repository, while maintaining compliance with the FAIR principles, represents a considerable challenge. A significant investment in processing hours and human analysis is required to guarantee the quality and enrichment of BrCris data and metadata.

However, it can be seen that with each new addition of data, confidence in the information provided grows. Despite this, research is still being carried out into the reuse of this data, as well as the development of tools to validate and consolidate it. We are increasingly seeking to be representative of the Brazilian research ecosystem and aligned with the FAIR principles.

Conclusion

BrCris stands out as a fundamental space for research and analysis of scientific data, presenting information aggregated and organized in a robust semantic data model. This approach makes it possible to create services aimed at different stakeholders, covering both management and academic research. By providing access to reliable and personalized information, the platform simplifies decision-making based on solid data.

With its flexible proposal, BrCris enables detailed analysis of trends, metrics and opportunities for collaboration, playing a catalyzing role in the advancement of scientific research. By facilitating the interpretation and extraction of results, the platform contributes significantly to the expansion of knowledge and creates an environment conducive to innovation and effective collaboration between the various players in the national and international scientific scene. Ultimately, BrCris not only adheres to the FAIR principles, but also consolidates itself as a catalyst for Open Science, promoting a more transparent, accessible and collaborative scientific ecosystem.

Bibliography

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

DE MEIS, Leopoldo 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, Ying; GLÄNZEL, Wolfgang; ZHANG, Lin. Tracing the development of mapping knowledge domains. Scientometrics, v. 126, p. 6201-6224, 2021.

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

LETA, Jacqueline; GLÄNZEL, Wolfgang; THIJS, Bart. Science in Brazil. Part 2: Sectoral and institutional research profiles. Scientometrics, v. 67, n. 1, p. 87-105, 2006.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008, August). Arnetminer: extraction and mining of academic social networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 990-998).

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