ISSN 1678-765X doi 10.20396/rdbci.v21i00.8673171/en

Article

RDBCI
Revista Digital de Biblioteconomia e Ciência da Infor
Digital Journal of Library and Information Science

BrCris: tools for treatment, analysis, and dissemination of scientific information in support of Open Science in Brazil

Vivian dos Santos Silva [1] Jesiel Viana da Silva [2] Thiago Magela Rodrigues Dias[3] Renê Faustino Gabriel Junior[4] Washington Luís Ribeiro de Carvalho Segundo[5]

ABSTRACT

Introduction: CRIS systems constitute comprehensive information systems over the entire ecosystem of the scientific process. BrCris aims to integrate and organize information regarding research activities, projects, publications, researchers, institutions, financing and other relevant data in the Brazilian scientific context. Objective: This study aims to discuss data processing on the BrCris Platform and analyze the computational tools used for this purpose, exploring three main approaches: data integration and consistency, visualization and validation, in addition to data certification. Methodology: The study is descriptive, presenting in detail the BrCris data processing steps, discussing the challenges encountered in handling large volumes of information. Furthermore, it describes the computational tools used to process scientific and technological information. Results: The study reveals the procedures for data processing and the computational tools developed for information systems, as well as the integration and analysis of the data obtained. The results of the processing and modeling of information based on VIVO, with data and graphic panels, and the opportunities for reusing the generated data are presented. The integration of data into a self-declared repository (Lattes) and the theses and dissertations aggregator repository (Oasisbr) is also detailed, culminating in the issuance of a certification seal. Conclusion: The results show that the adoption of these computational tools provides easy and agile access to an extensive set of consolidated information, previously dispersed across various sources, especially due to the diversity of repositories and individual access limitations. Thus, this study presents a set of computational tools whose functionalities are aligned with the guidelines of Open Science in Brazil.

KEYWORDS

Open science. Data repositories. Information processing. BrCris. Scientific data.

Authors' correspondence

[1] Universität Passau, Passau - Alemanha
vivian.ss@gmail.com

[2] Instituto Federal do Piauí
Teresina, PI - Brasil
jesiel.analista@gmail.com

[3] Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, MG - Brasil
thiagomagela@gmail.com

[4] Universidade Federal do Rio Grande do Sul
Porto Alegre, RS - Brasil
renefgj@gmail.com

[5] Instituto Brasileiro de Informação em Ciência e Tecnologia
Brasília, DF - Brasil
washingtonsegundo@ibict.br

BrCris: desenvolvimento de ferramentas no tratamento, análise e disseminação da informação em apoio à Ciência Aberta no Brasil

RESUMO

Introdução: Os sistemas CRIS constituem-se em sistemas de informação abrangentes sobre todo o ecossistema do processo científico. O BrCris tem como propósito integrar e organizar informações referentes a atividades de pesquisa, projetos, publicações, pesquisadores, instituições, financiamentos e outros dados relevantes no contexto científico brasileiro. Objetivo: Este estudo visa discutir o processamento dos dados na Plataforma BrCris e analisar as ferramentas computacionais empregadas para essa finalidade, explorando três abordagens principais: integração e consistência dos dados, visualização e validação, além da certificação dos dados. Metodologia: O estudo se configura

como descritivo, apresentando em detalhes as etapas de tratamento de dados do BrCris, discutindo os desafios encontrados no manuseio de grandes volumes de informações. Além disso, descreve o ferramental computacional utilizado para o tratamento das informações científicas e tecnológicas. Resultados: O estudo revela os procedimentos para o tratamento de dados e as ferramentas computacionais desenvolvidas para os sistemas informacionais, bem como a integração e análise dos dados obtidos. São apresentados os resultados do tratamento e modelagem das informações baseadas no VIVO, com dados e painéis gráficos, e as oportunidades de reutilização dos dados gerados. Também é detalhada a integração dos dados em um repositório autodeclarado (Lattes) e no repositório agregador de teses e dissertações (Oasisbr), culminando na emissão de um selo de certificação. Conclusão: Os resultados evidenciam que a adoção dessas ferramentas computacionais proporciona um acesso facilitado e ágil a um extenso conjunto de informações consolidadas, previamente dispersas em várias fontes, especialmente devido à diversidade de repositórios e limitações de acesso individualizados. Assim, este estudo apresenta um conjunto de ferramentas computacionais cujas funcionalidades estão alinhadas com as diretrizes da Ciência Aberta no Brasil.

PALAVRAS-CHAVE

Ciência aberta. Repositórios de dados. Tratamento da informação. BrCris. Dados científicos.

CRediT

- Acknowledgments: Not applicable.
- Funding: This study was partially funded by the Brazilian agencies Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF) - process 00193-00000788/2021-31; National Council for Scientific and Technological Development (CNPq) - process 400038/2023-4; Financier of Studies and Projects (FINEP) - agreement 01.16.0051-00.
- Conflicts of interest: Authors certify that they have no commercial or associative interest that represents a conflict of interest in relation to the manuscript.
- Ethical approval: Not applicable.
- Availability of data and material: The datasets generated and/or analysed during the present study are available in the Zenodo Scientific Data Repository.
- Authors' contributions: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization; Writing – original draft: SILVA, V. S.; Data Curation, Formal Analysis, Research, Software, Visualization; Writing – original draft: VIANA, J.; Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Management, Supervision, Validation; Writing – original draft; Writing – proofreading & editing: DIAS, T. M. R.; Data Curation, Formal Analysis, Research, Methodology, Software, Validation; Writing – original draft; Writing – proofreading & editing: GABRIEL-JUNIOR, R. F.; Conceptualization, Formal Analysis, Funding Acquisition, Research, Methodology, Project Management, Resources; Writing – original draft: CARVALHO-SEGUNDO, W. L. R.

JITA: IN. Open science.
ODS: 9. Industry, Innovation and Infrastructure

turnitin

Article submitted to the similarity system

Submitted: 21/03/2023 – Accepted: 23/11/2023 – Published: 17/12/2023

Editor: Gildenir Carolino Santos

1 INTRODUCTION

Humanity has shown itself to be capable of facing the most important challenges, in different contexts, that impact on the way we live and hinder our development. At the same time, we are witnessing the emergence of treatments, drugs, and vaccines for the most diverse diseases, new food production processes that guarantee the perpetuation of the species and technological developments that transform people's daily lives. These advances have a common origin, i.e., they are the result of the practical application of some theoretical scientific discovery which, in turn, derives from the research effort undertaken by researchers in the most diverse areas of knowledge (TANG *et al*., 2008).

The production of scientific knowledge is a process that takes time and is incremental (Huang; Glänzel; Zhang, 2021). Researchers look for the basis for their research in the observed development of their area of knowledge (state of the art). Identifying the state of the art in a given area can be done mainly by analyzing publications that concentrate information on that topic of knowledge.

Researchers who work collaboratively often publish the results of their research in the same way. Identifying co-authorship relationships in scientific/technological papers and research projects makes it possible to structure networks of researchers and institutions that are linked by their academic output (Abbasi; Altmann; Hossain, 2011).

Currently, everything that is known about the emergence and development of disciplines, the dissemination of knowledge and the evolution of science and technology is predominantly the result of analyzing scientific publications (De Meis et al. 2003; Leta; Glänzel; Thijs, 2006), analyzing scientific collaboration (Yoshikane; Kageura, 2004) and analyzing patent registrations (Abbas; Zhang; Khan, 2014).

Brazil has a significant share of international scientific production, mainly in specific niches in its economy. The country is a leader in the production of knowledge in Latin America (Collazo-Reyes, 2014) and an attractor of talent in the regional context (Saraiva; Miranda, 2004). It stands out for its implementation of digital platforms for national registration of the work of its researchers.

The international highlight is the Lattes Platform, which emphasizes the strategic importance of having scientific curricular information widely available (Lane, 2010). Similar to the Lattes Platform, there are other national platforms that record part of academic and technological activity, such as CAPES' Sucupira, Ibict's Theses and Dissertations Database, INPI's Industrial Property Database and the Federal Government's Transparency Portal, for example. Although all the information in these data sources is open and free to consult, they are restricted to their databases and there is no integration to enable this information to be widely exploited by science and society in Brazil.

In this context, it is important to emphasize that information from research into Science, Technology and Innovation (ST&I) represents one of the main pillars for a country's economic and social development. This information is a fundamental guiding element for the systematic generation of knowledge, both theoretical and applied.

The information generated within the scope of Science, Technology and Innovation (ST&I) is highly specialized, distinguishing it from other types of information. Its production is based on a specific method - the scientific method - and the dissemination of its

and the dissemination of its results follows distinct procedures, including evaluation, validation, publication, and access through specialized sources (Meadows, 1999).

The diversity of information sources, together with the variety of their data models (metadata), resulting from research in Science, Technology and Innovation (ST&I) are often stored in databases and repositories that have different data structures, aimed at ensuring the preservation, visibility and effective retrieval of the information contained, and in a few cases

aim for integration.

Complementing Brazilian information sources, there are international ones, which enable data exchange for both humans and computers. These databases include Wikidata, CrossRef, OpenCitations, OpenAIRE Research Graph, Latindex and DOAJ, among other open sources of STI information.

Gathering and integrating data from these and other sources is a challenge that requires a great deal of intellectual effort and computational power. The storage capacity required exceeds what a traditional database management system can handle, requiring advanced technological solutions from the so-called Web 4.0, as well as analyses involving computational intelligence techniques.

Against this backdrop, initiatives have emerged to create systems to manage academic production, whether institutional, national or thematic. These systems are known by the acronym CRIS (Current Research Information Systems). Their main aim is to integrate information from various databases, providing reports, indicators and consolidated data for managers, researchers and other users, allowing them to analyze production in their respective countries or areas of knowledge.

CRIS systems facilitate the promotion of Open Science by making data related to projects, results, publications, patents, research groups, researchers, and institutions visible and accessible. This not only facilitates interoperability between different systems, but also resonates with the core of Open Science, which emphasizes collaboration and the sharing of resources. Thus, CRIS can be adjusted to foster interconnection between researchers, enabling the identification of areas of common interest, boosting collaborations and the exchange of information (Singh *et al*., 2021).

In this context, the research problem seeks to report on the challenges faced and strategies adopted during the construction of BrCris and the development of computational tools for processing, analyzing and disseminating Brazilian scientific information, and what results and impacts these initiatives have achieved.

In this way, the research aims to discuss the treatment and processing of data from the BrCris Platform, as well as the computational tools used for this purpose. More specifically, the general objective of the research is subdivided into discussing the integration and consistency of the data for BrCris; validating the data by visualizing the data; and finally, discussing the data generated and the certification of the data.

2 THEORETICAL REFERENCE

The research ecosystem involves the participation of multiple actors who interact with each other. These range from obtaining funding for a research project to the fundamental role of the researcher, who uses infrastructure such as laboratories and physical equipment to conduct his or her investigations (Lee; Bozeman, 2005). In turn, researchers are linked to institutions where they conduct their research to generate knowledge, often documented in scientific articles and/or technical reports that are made available in databases (Singh *et al*., 2021) or research repositories.

The CRIS model defines an information system for the entire ecosystem of the scientific process. All information in the scientific research cycle is organized in one place, from funding, through projects, researchers, research institutions and laboratories, to the outputs of scientific research, such as scientific articles, theses, dissertations, books, book chapters, patents and scientific datasets (Sivertsen, 2019).

For Joint (2008), CRIS can be defined as an information system capable of managing all relevant research information, from the initial stage of identifying funding opportunities, through the writing and submission of proposals, to the follow-up of successful proposals that become active projects that are managed until their completion. At this point, results are

produced, including various publications or other artifacts related to the research activity.

Torino, Coneglian and Vidotti (2020) emphasize that setting up an institutional CRIS requires the integration of institutional representation structures, knowledge of the information systems that store them, the ways in which data are displayed and the communication protocols available, to define the conversion structure.

Ecosystem mapping initiatives are underway in Europe with the Directory of Research Information Systems (DRIS) run by euroCRIS (Eurocris, 2023). In Brazil, BrCris is the mapping of the Brazilian scientific research information ecosystem. It takes the form of a platform that aggregates different sources of information, making it possible to retrieve, certify and visualize data and information related to the different actors involved in scientific research in Brazil. Among the main sources are the curricular data of individuals, organizations, postgraduate programs, publications, academic guidelines, scientific journals, patents, research groups, software, and other sources to be added (Dias *et al*., 2022).

BrCris provides a unified information search interface, visualization of collaborative networks and dashboards of science, technology, and innovation indicators. And it establishes a single model for organizing scientific information from the entire Brazilian research ecosystem (Dias; et al., 2022). The agents of this ecosystem include researchers, projects, infrastructures, laboratories and research institutions, funding agencies, as well as the results of research, mainly expressed through scientific publications, theses, dissertations, scientific datasets and patents (Kong *et al*., 2019).

In this context, the idealization of the BrCris System project (Pinto et al., 2021), which is the CRIS in the context of Brazilian Open Science, was conceived in 2014, when inspired by the model proposed by Portugal for a national CRIS (PTCRIS - https://ptcris.pt), the Brazilian Institute of Information in Science and Technology (Ibict) began a sequence of studies and inter-institutional partnerships for the execution of the project. In 2020, the research project to build BrCris was formally implemented. The aim of the project was to make technological tools available to provide consolidated Brazilian scientific and technological data to the entire academic community.

To make this available, BrCris adopts a two-level data representation scheme. The first is the logical level, materialized as a model of entities and relationships based on CERIF (Jörg, 2010), also known as a metamodel. This model is translated into a physical relational schema on the La Referencia platform, where the data is loaded and processed. It is therefore a model that meets internal data storage needs and is responsible for organizing and integrating the information collected so that it becomes input for the project's objectives (Pinto *et al*., 2021).

The VIVO-ISF (Integrated Semantic Framework) ontology, used by the VIVO platform, can be used to represent the academic and scientific domain. The VIVO-ISF ontology is based on the high-level Basic Formal Ontology (BFO), which provides a well-founded conceptual basis. The ontology also allows for extensions that incorporate local institutional features (Rathke; Rocha, 2019). The VIVO-ISF ontology was developed by integrating several other ontologies and vocabularies, which makes it a comprehensible tool for several other platforms. This ontology is versatile and can be used in various applications, as it represents the domain of academic information, covering aspects such as publications, teaching, guidance and other relevant areas. It is also worth noting that the use of other ontologies and vocabularies facilitates the connection with linked data databases, as there is greater compatibility between resources and other databases (Lyrasis, 2016).

As a model that describes the academic research domain, the VIVO-ISF ontology is made up of classes and properties that represent a network of researchers, the institutions, and projects to which they are linked, and the publications, patents, software and any other products of their research. Its main advantage is the reuse of other well-established ontologies, such as Bibliographic Ontology (BIBO), Event Ontology (EO), Friend of a Friend (FOAF), Geopolitical Ontology (GEO), Software Ontology (SWO), Simple Knowledge Organization

System (SKOS) and vCard, among others. Also, noteworthy is the integration of the Basic Formal Ontology (BFO), a foundational ontology that provides a solid conceptual basis for the classes and properties of the model.

The BrCris semantic model is made up of a subset of the VIVO ontology, represented by classes and properties equivalent to the entities, attributes, and relationships of the logical metamodel, plus a local extension that covers information specific to the Brazilian context. Each entity in the model is assigned a unique identifier, i.e. each entity within the research ecosystem has a permanent link on the web, guaranteeing the absence of ambiguities.

The use of a semantic model based on ontology allows data to be represented as a knowledge graph, which allows this data to be published and consumed as Linked Open Data (LOD). Bauer and Kaltenböck (2011) emphasise that to really take advantage of open data, it is crucial to put information and data into context, creating new knowledge that feeds efficient services and applications. They also add that, as it is an important mechanism for integrating and managing information, making LOD available facilitates innovation and the multiplication of knowledge from linked data, which is in line with the principles and goals of BrCris and CRIS platforms in general.

Publishing data as LOD is a good practice for sharing data in Open Science, mainly because it allows for the traceability of the information made available. This is particularly important in the context of BrCris due to the volume and diversity of data sources. By generating linked data, it is possible, for example, to link entities such as "Person" with their profiles on the Lattes or Orcid platforms; "OrgUnits" with their records in the Research Organization Registry (ROR) or the Global Research Identifier Database (GRID); “Publications” with their entries in BDTD, Oasisbr, or any repository where they are identified by their DOI; and so on.

3 METHODOLOGY

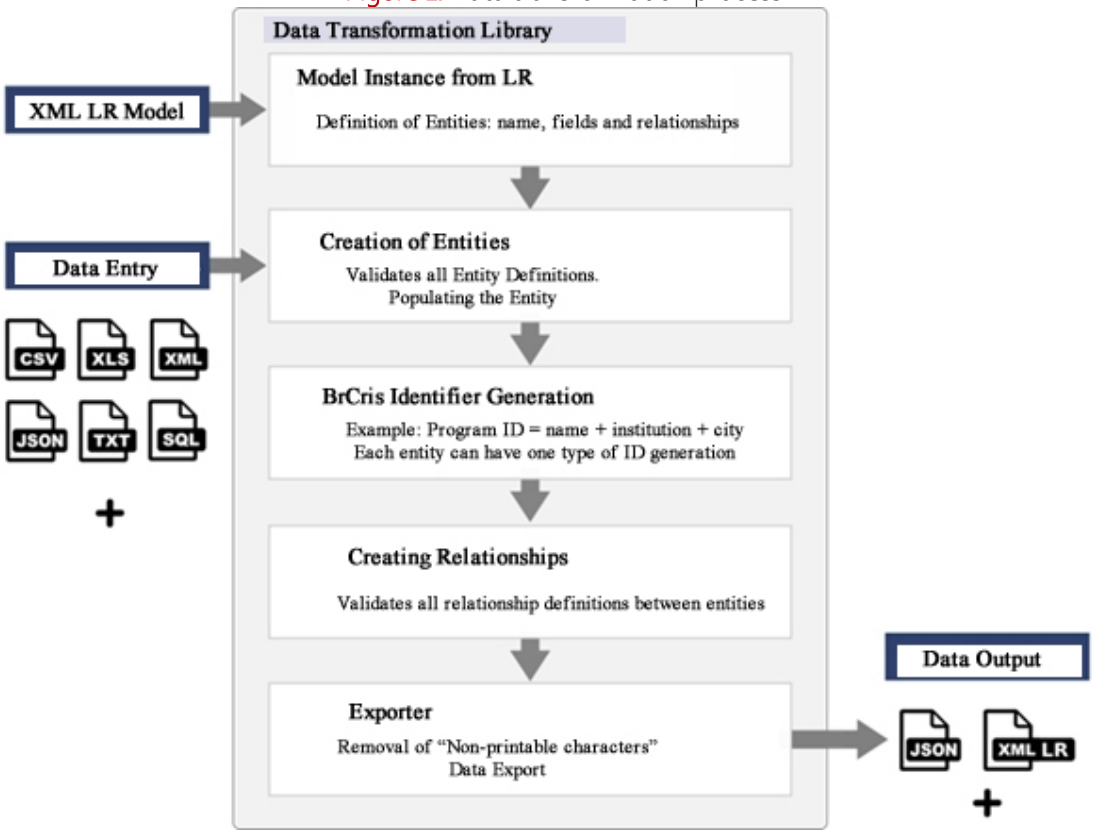
The study is characterized as descriptive in that it presents the BrCris data processing stages in detail, as well as discussing the challenges faced when processing large volumes of data. To develop the BrCris Platform, it was necessary to develop tools for processing information. In this context, and to meet the first objective of discussing data integration and consistency for BrCris, it was necessary to develop and test a methodology for transforming structured and unstructured data into a standardized format. For this standardization, the BrCris data model began by structuring nine data entities, following standards widely used in the international scientific community, namely: publications, theses and dissertations, people, journals, patents, research groups, software, expert networks and themes (grouping of concepts).

The data transformation process for BrCris is shown in Figure 1. This process begins with collecting data from the sources and mapping their metadata to the La Referencia (LR) model, to define the types of entities in BrCris. The intention is to reuse existing ontologies, avoiding the creation of new classes and/or relationships. Each entity is identified and validated through a combination of descriptive elements, ensuring the absence of ambiguity. For example, in the case of a postgraduate program, the name of the program, the institution, and the city are used to generate a unique identifier in BrCris.

Once the entities have been created, we move on to creating and validating the relationships between them, as defined in the ontology. Once the information has been validated, the content is processed, eliminating "non-printable characters" such as control characters, non-visible spaces and formatting characters such as bold. After this treatment, the data is ready to be exported to the BrCris Platform.

When creating the semantic model, the premises adopted were maximizing the reuse of existing resources and using international standards for data representation in the area, so that BrCris would be compatible with similar systems around the world.

Figure 1. Data transformation process



Source: Authors (2023)

After the entire processing process, regardless of the data source, the data generated as output is imported into a single database, and using the unique identifiers generated or identified, the data sets are linked and deduplicated, thus enabling data interoperability, regardless of its source and format. The output data, which can be characterized in various data formats, also makes it possible to import previously processed sets in such a way that they can be incorporated by other analysis tools.

To visualize and validate the data, the data generated by VIVO and the consolidation of this information is analyzed. This analysis was carried out by Ibict's information development and processing team.

BrCris data is certified through integration with the Lattes platform and Oasisbr, with a description of the process for validating theses and dissertations between these systems.

4 RESULTS

This section shows the results obtained with all the tools developed as described above. This set of tools has helped access and evaluate Brazilian scientific production. It can also provide inputs such as data sets in standardized formats that facilitate integration with other sets and tools. These actions aim to provide mechanisms to boost access to scientific data easily, contributing significantly to the advancement of Open Science in Brazil.

4.1 Data Integration and Consistency

Considering the whole process of curating the data to be collected, integrated and analyzed in this project, a strategy for generating identifiers is necessary. These identifiers are important because all the data collected will be mapped to previously identified entities, which must be uniquely identified considering the entire processing process to be carried out, especially the data disambiguation and deduplication process.

One of the tools developed was Ocean Dragon, which allows obtaining individual data, institutions, education, publications, orientations, among others. Ocean Dragon has a data structure made up of entities, relationships, nested fields with the possibility of internationalization, making an exchange with the La Referencia platform, used by Ibict to collect data and index it on various other platforms.

To this end, a computer library has been proposed using the Python programming language, which is responsible for generating BrCris identifiers, created to pre-disambiguate the data, avoiding duplicate entities in the set to be analyzed.

For each data set to be analyzed, a strategy for generating the identifiers has been considered. This strategy aims to use as little extracted information as possible, but which can generate, with a satisfactory level of confidence, unique identifiers to be used in various future stages.

The library developed to process the data is also responsible for the generation of BrCris identifiers, created to pre-disambiguate the data, avoiding duplicate entities on the La Referencia platform. The library also allows the validation of the data against the model used to structure the La Referencia platform. In other words, it checks that the names of entities, fields, and relationships are consistent with the platform, avoiding problems during loading.

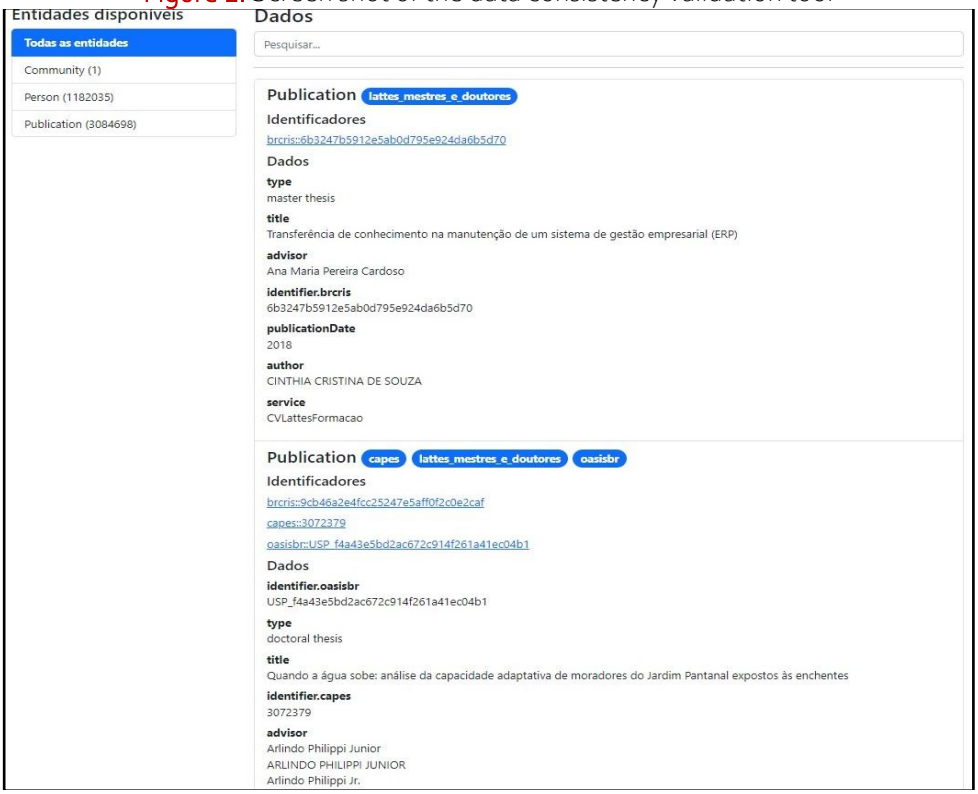
The creation of the entity and relationship metamodel considered the characteristics of the different data sets collected, to facilitate data processing routines, especially entity deduplication. However, in this format, the resulting information would not be easily accessible and reusable by external agents. To solve this problem, the second layer of representation, the semantic layer, was developed and implemented as an ontology to allow the data to be visualized and navigated as a knowledge graph.

With all the steps described above, it was possible to perform the entire data collection, transformation, and integration process consistently. These strategies, developed by applying different techniques validated in different studies, made it possible to characterize a single, duly validated data set.

When loading data onto the La Referencia platform, it is important to ensure that all information is correctly deduplicated. To achieve this, the platform uses unique identifiers, as mentioned above, to identify similar entities and merge the data. These identifiers are known as "brcrisId" and are created differently for each type of entity.

To facilitate the visualization, verification, and analysis of what has been sent to the La Referencia platform and how the data has been processed, a tool (Figure 2) was developed that allows the user to navigate between entities and visualize the result of the loading and deduplication in a clear and intuitive way, in this case acting as a data curator.

Figure 2. Screen shot of the data consistency validation tool



Source: Authors (2023)

By grouping entities with unique identifiers, the platform can reduce the size of the dataset and eliminate redundant information, making it easier to manage and analyze.

In the context of BrCris, this deduplication process is also essential to guarantee data interoperability. When selecting and processing data for integration from different data sources, the platform uses implicit identifiers or those generated during the data processing process to resolve possible deduplications and guarantee data consistency.

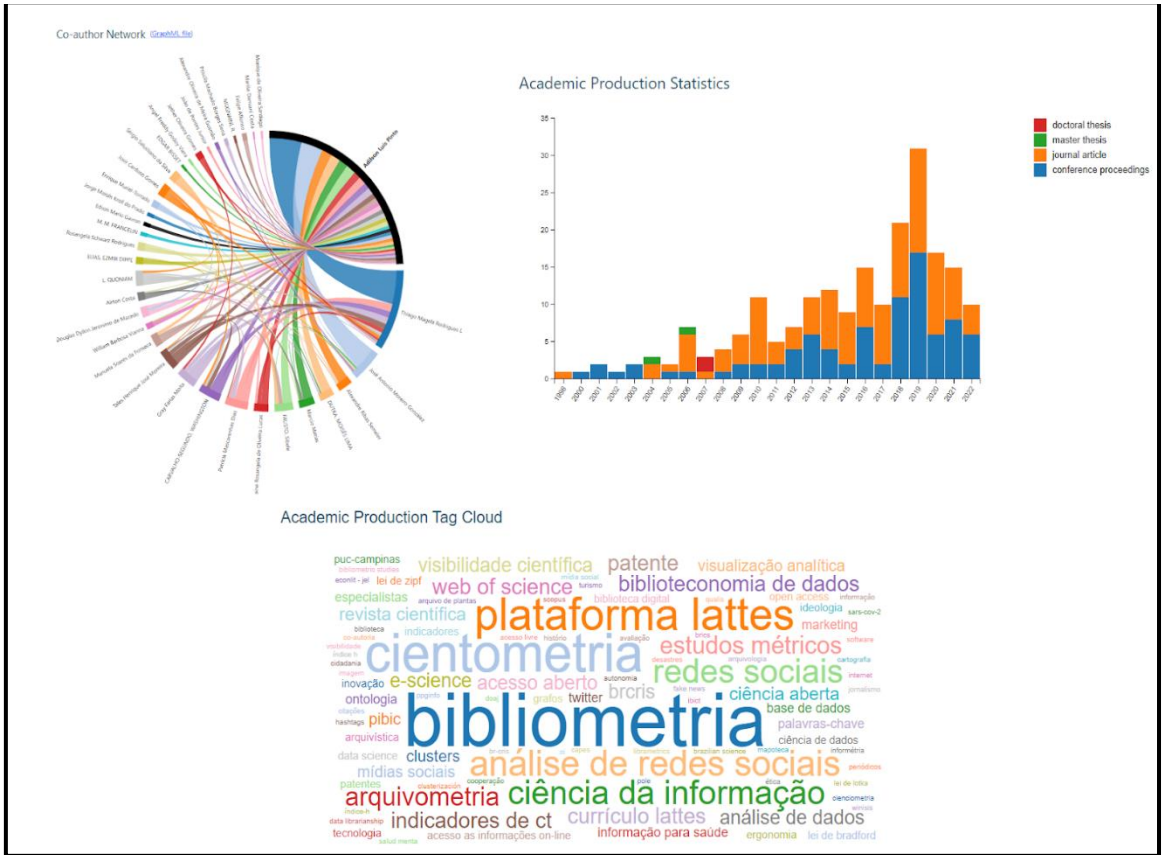
4.2 Data Visualization

With the data aggregated and deduplicated, the platform can offer various bibliometric analyses, allowing users to carry out a wide variety of studies and research based on the information available. In short, the deduplication process is a fundamental step in guaranteeing the accuracy, reliability, and interoperability of the data, making it even more valuable for analysis and research.

After all the stages of collecting, processing and integrating the data, it is possible to access the dataset with the help of graphical interfaces that make it easier to access and certify the sets. The VIVO ontology, in particular, allows the data to be visualized on the VIVO Platform, a tool for browsing data in the academic domain that allows BrCris to serve Linked Open Data to external agents, as well as facilitating the exploration of the knowledge graph. Another important feature offered by the VIVO platform is the graphical visualizations, which provide a broader overview of a given individual. In addition to the predefined visualizations, it is also possible to implement and include custom graphics in the interface in a simple way.

Figure 3 illustrates some of these visualizations: a researcher's co-authorship network, which is already available natively on the platform, and two custom graphs implemented for BrCris: the total number of publications by type and the word cloud given by the keywords of a given researcher's publications.

Figure 3. Visualizations of the VIVO Platform



Source: BrCris (Ibict, 2023)

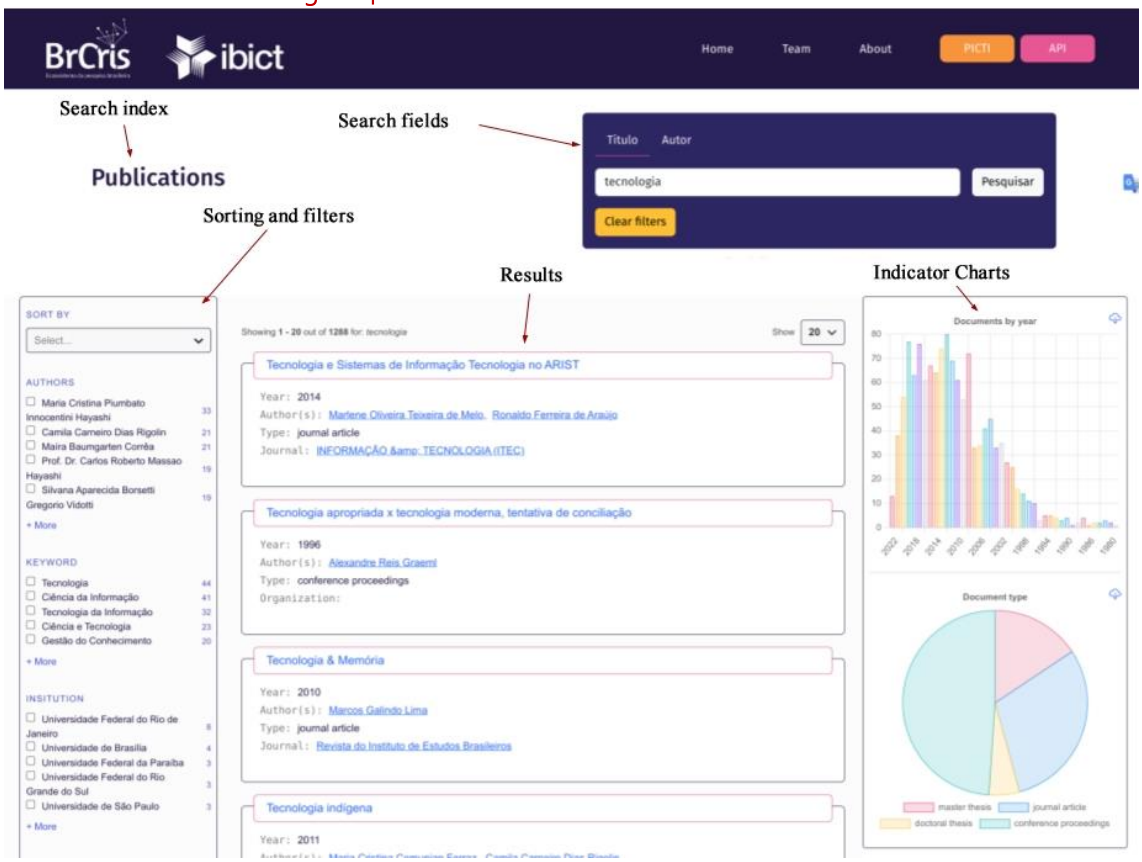
Reinforcing the theme of good practices for publishing and sharing data, it is important to note that by following international standards and adopting the same semantic model as similar systems, we ensure that the data collected and processed in BrCris is accessible, reusable, and interoperable, thus also adhering to the FAIR principles (Findability, Accessibility, Interoperability, Reusability). The RDF representation based on the VIVO ontology enables both the unambiguous description of data and the automatic retrieval and cross-referencing of information, which is in line with the FAIR principles' emphasis on the automatic discovery and use of data, in addition to its reuse by individuals (Wilkinson *et al*., 2016).

As an information retrieval component, a graphical web interface was developed (Figure 4) based on Elastic's Search-UI. Search-UI is a free, open-source library written in Typescript that provides various customizable web components compatible with desktop and mobile applications, which integrates with Elasticsearch to provide an interface for searching and viewing information on the web.

Search-UI[1] features a set of configurable search components, making it easy to build rich, fully customizable search interfaces with advanced filtering features that help users find exactly what they need. It offers pagination, predictive typing, autocomplete, filtering, sorting and indicator graph generation.

[1] Search-UI is responsible for presenting the user with an intuitive, user-friendly and efficient interface for searching for information in a given system.

Figure 4. BrCris Search and Visualization Interface



Source: BrCris (Ibict, 2023)

Using Search-UI, it was possible to develop an interface to perform text searches in natural language with the addition of Boolean operators. Users' search requests are cross-referenced with an Elasticsearch index, allowing retrieval and ranking to be carried out quickly.

Once the search has been carried out, the BrCris retrieval interface allows the results to be sorted, such as alphabetical sorting, chronological sorting, which is very useful for researchers as it allows them to select the most recent results, sorting by relevance according to the Elasticsearch ranking. It is also possible to apply filters to refine the search by some attribute of the documents, for example: author's name, type of document.

All of these visualizations, which are currently available, allow for easy interaction with the entire data set that was initially collected, processed and integrated. In addition, various filters can be applied and the results of searches exported to standardized formats, making a significant contribution to the advancement of Open Science, since it provides analysis of certified datasets that were previously inaccessible.

4.3 Data certification

By integrating the data into a standardized and properly evaluated data repository, a whole process of certifying data originating from other sources that have not yet been validated can be carried out, providing a real view of Brazilian scientific and technological production.

Data certification systems are characterized as mechanisms for verifying the origin, veracity, integrity, and reliability of data sets stored in different systems (Dias *et al.*, 2023).

The self-declaratory actions of a system can be validated by an agent called a "trusted third party" (self-declaratory signature versus certified signature), providing security and veracity to the information provided.

In science and technology information sources, however, data certification systems are still rare in Brazil, given the various challenges involved in the certification process. The lack of persistent identifiers for the various entities that make up the entire scientific research ecosystem stands out as one of the main limitations.

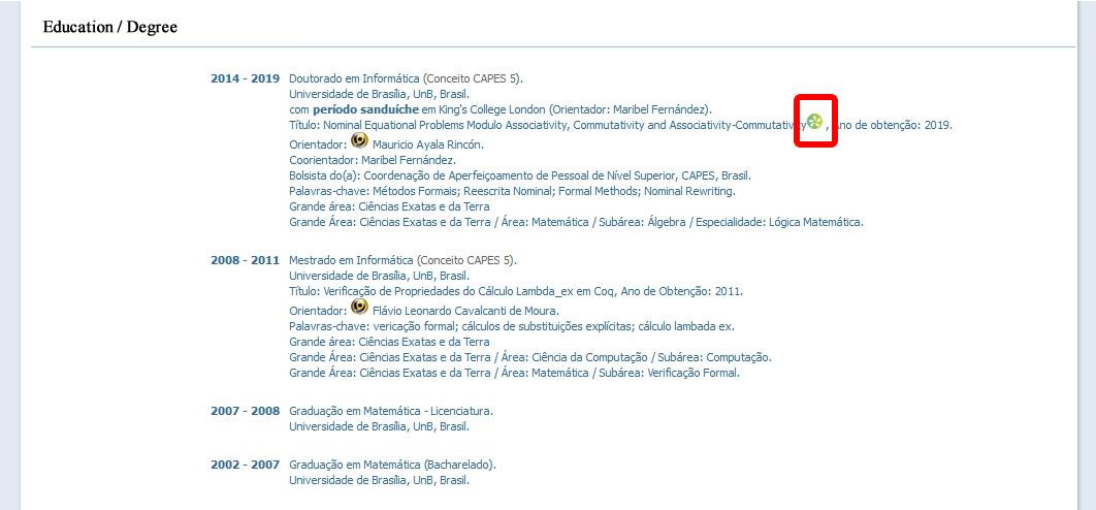
With the data collected and already deduplicated, classified and categorized, it can later be adapted and validated, establishing relationships with records from other sources. A record collected from source "A" has a common attribute with the record collected from source "B", and a link can be established between the two with a certain degree of reliability. The other record attributes can be merged to result in a single enriched record, eliminating replicates. A validation scheme can be created to discard malformed, redundant, inconsistent or ambiguous records.

In the context of this work, the first certification model tested is the integration of the Lattes Platform with Oasis.br (https://oasisbr.ibict.br/). Through developments within the scope of the BrCris Project, it was possible to create an intelligent mechanism for identifying theses and dissertations declared in the academic training and completed orientation sessions of a given Curriculum registered on the Lattes Platform, which were also included in the registry aggregated by Oasisbr.

In this way, Oasisbr becomes the "trusted third party" in this process, without the need for the pre-existence of a persistent identifier explicitly assigned to the thesis or dissertation.

The entire certification process is based on computational strategies that have been tested and validated in various studies, by analyzing self-declared information and comparing it with information entered into repositories, digital libraries and portals aggregated by Oasisbr. The certification seal is displayed next to the titles of the theses or dissertations in the user's curriculum (Figure 5).

Figure 5 . Fragment of a Curriculum with Certification Included



Source: Lattes Platform (CNPq, 2023)

With the seal, you can quickly and easily obtain documentary proof of the title you have entered and access the document on Oasisbr. The certificate can be issued automatically by the Lattes Platform or requested manually by the user.

Through developments under the BrCris Project, it was possible to create an intelligent mechanism to identify theses and dissertations declared in the completed training and orientation sections of a given Lattes CV, which were also included in the set of records aggregated by Oasisbr. In this way, Oasisbr becomes the "trusted third party" in this process, without the need for the pre-existence of a persistent identifier explicitly assigned to the thesis or dissertation. Among the CVs on the Lattes Platform, there are approximately 1.1 million

records of theses and dissertations, 65% of which (approximately 700,000) can already be certified. Another 10,000 theses and dissertations defended abroad have also been mapped to receive the certification seal.

The advantages of the certification process are many. Through certification, it is possible to verify that the scientific work, guidance, participation on boards, among other elements from self-declared sources are really true, avoiding false information. Therefore, certification is a factor that gives the researcher greater credibility and authority. In this way, BrCris is an important contribution to understanding Brazilian open science.

5 CONSIDERATIONS

BrCris is an important space for research and data analysis. The information aggregated and organized according to a semantic data model enables the generation of services for various players in the contexts of management and academic research, as well as in the area of information for innovation.

The platform is an initiative that collects and enriches data from repositories and open databases and is a unique proposal in the world that makes it easier to obtain a Brazilian Panorama of the Production and Performance of all its academic/scientific actors. However, it requires a lot of computer and human resources to process and standardize this data.

With BrCris, those involved in the Brazilian scientific research ecosystem will have easy access to a large aggregate of scientific data, according to the best practices of the FAIR principles, obtaining data sets or information of interest in an accessible way.

It should be noted that Lattes CVs are self-declaratory in nature, thus highlighting the importance of the certification processes applied to this database. A thesis or dissertation is only considered an official degree document if the final version, with the respective corrections suggested by the evaluators, is deposited in an official and publicly accessible repository. This strategy can be replicated in other datasets with a view to certification.

REFERENCES

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

ABBASI, A.; ALTMANN, J.; HOSSAIN, L. Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. **Journal of informetrics**, Amsterdam, v. 5, n. 4, p. 594-607, 2011.

BAUER, F.; KALTENBÖCK, M. **Linked open data**: the essentials. Vienna: Edition mono/monochrom, 2011. p. 21, v.710.

COLLAZO-REYES, F. Growth of the number of indexed journals of Latin America and the Caribbean: the effect on the impact of each country. **Scientometrics**, Budapest, v. 98, p. 197-209, 2014.

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO (CNPq). Plataforma Lattes. Brasília. Available at: https://lattes.cnpq.br. Access on: 12 out. 2023.

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

DIAS, T. M. R. *et al*. Brcris: plataforma para integração, análises e visualização de dados técnicos-científicos. , p. 622-638, **Informação e Informação**, Londrina, v. 27, n. 3, 2022. DOI: https://doi.org/10.5433/1981-8920.2022v27n3p622.

DIAS, T. M. R. *et al*. *In*: WORKSHOP DE INFORMAÇÃO DADOS E TECNOLOGIA, 6., 2023, Brasília, DF. **Anais**... Brasília: Ibict, 2023. DOI: https://doi.org/10.22477/vi.widat.53.

EUROCRIS. **Directory of Research Information Systems** (DRIS). Nijmegen, Nertherlands. Available at: https://eurocris.org/services/dris. Access on: 23 out. 2023.

HUANG, Y.; GLÄNZEL, W.; ZHANG, L. Tracing the development of mapping knowledge domains. **Scientometrics**, Budapest, v. 126, p. 6201-6224, 2021.

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

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.

LANE, J. Let's make science metrics more scientific. **Nature**, London, v. 464, n. 7288, p. 488-489, 2010.

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. Available at: https://elibrary.ru/item.asp?id=11423996. Access on: 27 mar. 2023.

LETA, J.; GLÄNZEL, W.; THIJS, B. Science in Brazil. Part 2: Sectoral and institutional research profiles. **Scientometrics**, Budapest, v. 67, n. 1, p. 87-105, 2006.

MEADOWS, A. J. **A comunicação científica**. Trad. A. A. B. de Lemos. Brasília, DF: Briquet de Lemos, 1999.

PINTO, A. L. *et al*. The Brazilian current research information system: BrCris. *In*: SILVA, Carlos Guardado da Silva; REVEZ, Jorge; CORUJO, Luis (coord.).Organização do conhecimento no horizonte 2030: desenvolvimento sustentável e saúde. Lisboa: Universidade de Lisboa, 2021. p. 319. (Colecção CA–Ciência Aberta). ISBN 978-989-566-137-4. DOI: https://doi.org/10.51427/10451/50067.

RATHKE, S. B.; ROCHA, R. P. Sistema de informação de pesquisa: uso da ontologia de VIVO no contexto das instituições brasileiras. **Brazilian Journal of Information Science**, Marília, v.13, n.4, 2019. DOI: https://doi.org/10.36311/1981-1640.2019.v13n4.08.p132.

SINGH, V. K. The journal coverage of web of science, scopus and dimensions: A comparative analysis. **Scientometrics**, Budapest, v. 126, Jun., 2021. DOI: https://doi.org/10.1007/s11192-021-03948-5.

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, Cham. 2019. p. 667-683. DOI: https://doi.org/10.1007/978-3-030-02511-3_25.

TANG, J. *et al.* Arnetminer: extraction and mining of academic social networks. *In*: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 14th, 2008. **Proceedings of the…** Las Vegas, Nevada: ACM, 2008. p. 990-998. DOI: https://doi.org/10.1145/1401890.1402008.

TORINO, E.; CONEGLIAN, C. S.; VIDOTTI, S. A. B. G. Estruturas de representação para reuso de dados no contexto da ecologia de pesquisa: Cris institucional. Informação e Informação, Londrina, 2020, v.25, n. 3. DOI: https://doi.org/10.5433/1981-8920.2020v25n3p1.

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