ADVANCES IN KNOWLEDGE ORGANIZATION
Proceedings of International ISKO Conference - ISSN: 0938-5495 - 2026 - São Paulo

Metadata Representation and Quality in Multidisciplinary Research Data Repositories: A FAIR Assessment of the Deposita Dados Repository

Raimunda Fernanda dos Santos

Federal University of Rio Grande do Norte, Brazil

Carla Beatriz Marques Felipe

Federal University of Rio de Janeiro, Brazil

Abstract

This study examines issues related to the representation of information in research data repositories. Its objective is to assess the compliance of information representation practices and the quality of metadata in the Deposita Dados repository, based on the FAIR principles, to identify the platform's level of maturity in promoting Open Science within the Brazilian context. The methodology employed is descriptive, bibliographic and documentary research. It adopts a qualitative approach, outlined through a case study focused on the Deposita Dados repository. To analyze the repository based on the FAIR principles, the F-UJI (FAIRsFAIR Research Data Object Assessment Service) tool was applied to five datasets from the investigated repository. The selection of these datasets followed a temporal criterion, with the five most recent deposits available at the time of collection being chosen. Regarding the repository's adherence to the FAIR principles, the evaluation of the selected datasets revealed a compliance rate of 92%. The consistent results across all samples indicate the maturity and consistency of the repository's curation, organization, and metadata policies.

Keywords: research data, Deposita Dados, information representation, metadata.

1.0 Introduction

Research data constitute the essence of scientific inquiry, serving as the fundamental evidence that underpins the validity of a study's conclusions. Beyond their evidential function, such data are pillars of reproducibility, since their availability allows for the validation of results and the promotion of new peer-reviewed research. In this sense, the effectiveness of transparency and collaboration within the academic community becomes intrinsically dependent on how these data are organized and made available.

From this perspective, research data are characterized as records processed and collected during scientific inquiry, encompassing a variety of types and formats, ranging from structured spreadsheets and interview transcripts to visual, audio, and computational records (Santos and Felipe 2025). More than mere by-products of the experimental process, these data constitute complex informational objects that, to be effectively

integrated into the Open Science ecosystem, require an organization that preserves their evidential value and semantic integrity over time.

The growing production of data has driven the creation of specialised digital repositories. These platforms aim to store, organise, preserve, and disseminate research data in an accessible manner, enabling its reuse in future scientific investigations (Bonetti and Arakaki 2022). This data is represented in these repositories through metadata, which is structured information that describes an information resource. It functions as a kind of "identity card" for the data, providing essential details about its content, context, authorship, and format.

We agree with the authors Santos and Felipe (2025) when they state that the proper representation of data through standardized metadata and clear descriptions is crucial for its retrieval and reuse in new research. Good representation not only helps users find the data most relevant to their needs but also ensures the conditions for it to be understood and applied in different contexts. Without an accurate description, a dataset becomes practically useless, even if it is technically available online. Furthermore, standardization in representation contributes significantly to interoperability between systems and platforms.

For data sharing to go beyond mere technical availability and achieve full scientific utility, it is imperative that data management be guided by rigorous regulatory guidelines. In this context, the FAIR Principles emerge as a framework of strategic guidelines designed to ensure that data is effectively findable, accessible, interoperable, and reusable.

Based on this perspective, the following question arises: What is the level of maturity of the Deposita Dados repository regarding the implementation of the FAIR Principles, and how does the current structure of its metadata impact the preservation of the semantic integrity of research data? Considering the above, this study aims to assess the compliance of information representation practices and the quality of metadata in the Deposita Dados repository, based on the FAIR principles, to identify the platform's level of maturity in promoting Open Science in the Brazilian context.

The relevance of this study is justified by the pressing need to overcome the paradox of "availability without accessibility." Although the volume of data deposited in repositories has grown exponentially, the lack of technical rigor in describing these assets can turn them into "orphan data"—files that, despite being stored, cannot be located or interpreted by humans or machines. By evaluating the Deposita Dados repository through the lens of the FAIR Principles (Findable, Accessible, Interoperable, Reusable), this research seeks to determine whether Brazil's information infrastructure is, in fact, fulfilling its role of facilitating scientific discovery or whether it remains merely a passive repository of files.

From the perspective of Organization and Representation of Information and Knowledge, this research is justified by the need to investigate how representation structures act as semantic mediators in complex digital environments. Knowledge Organization (KO) also involves the creation of systems that guarantee the integrity of the scientific context. By analyzing the Deposita Dados repository, this study contributes to the theoretical corpus of

the field by discussing how the standardization and comprehensiveness of information representation influence the construction of interoperable knowledge networks.

2.0 Research data repositories and FAIR principles

According to Gäal and Martins (2022), Open Access serves as a mechanism for scientific communication aimed at democratizing knowledge, working to remove financial and legal barriers to research findings. It is imperative to emphasize, however, that openness is not absolute: in data repositories, access may be limited by safeguards related to confidentiality, privacy, or intellectual property. Once these ethical and legal restrictions are overcome, the model enhances the circulation of information across different fields, fostering interdisciplinary practices that allow researchers to consult and apply data from diverse areas of knowledge.

In this context, data sharing is based on the premise that such assets have potential value for reuse, both within their original disciplinary field and in interdisciplinary research (Borgman 2010). However, from the perspective of Knowledge Organization (KO), the utility of such data depends on the quality of its representation. For information to transcend mere technical availability and achieve scientific interpretability, it is necessary for data to be mediated by robust metadata, which acts as the semantic link between production and retrieval of knowledge.

For this data sharing to be effective and fulfill its purpose, it is essential that the data be managed according to specific guidelines. To this end, the FAIR principles were created, a set of essential guidelines that focus on the findability, accessibility, interoperability, and reusability of data (Diniz et al. 2018).

The FAIR principles (Findable, Accessible, Interoperable, Reusable) aim to improve data reuse through discoverability, accessibility, interoperability, and reusability. Their sub-principles address metadata, persistent identifiers, semantics, controlled vocabularies, provenance, and other elements that support information retrieval by humans and machines, reinforcing the role of descriptive and thematic metadata in the organization and retrieval of data.

The central articulation between Knowledge Organization and the FAIR principles lies in Semantic Representation. The pillar of Interoperability (I), for instance, advocates that data should employ formal, accessible, and shared languages for knowledge representation. This implies the use of Controlled Vocabularies, thesauri, and ontologies, which function as Knowledge Organization Systems capable of neutralizing terminological ambiguities. According to Guizzardi (2020), the interoperability proposed by the FAIR principles depends on semantically consistent informational structures grounded in domain ontologies, going beyond the mere use of controlled vocabularies. Without the use of these instruments, interoperability becomes fragile, since machines rely on rigid and standardized conceptual schemes to “understand” and relate datasets from different sources.

Furthermore, the increasing volume and variety of (meta)data impose challenges on traditional methods of information processing, management, and retrieval. In this context, new structures such as knowledge graphs, associated with ontologies and semantic metadata schemas, emerge as a promising approach for the implementation of the FAIR principles, as they enable explicit semantic representation, machine-readable structures, and the adoption of interoperable standards (Vogt et al. 2025). The FAIR sub-principles reinforce the foundational role of descriptive and thematic metadata, requiring persistent identifiers, clear semantics, and documented provenance. Thus, the FAIR principles serve as strategic guidelines to ensure the long-term integrity and reuse of scientific knowledge in digital environments. To follow up on these considerations, the methodological approach taken in this research is presented below.

3.0 Methodological procedures

This investigation is characterized as a bibliographic and documentary research study of a descriptive nature. It adopts a qualitative approach, structured through a case study focused on the Deposita Dados repository.

To establish the theoretical foundation for this research, a literature review was conducted in the BRAPCI, Scopus, and Web of Science databases in February 2026. The search strategy was designed to retrieve scientific publications addressing the topics of “research data,” “research data repositories,” “FAIR principles,” and “representation in research data repositories.” To ensure adherence to the topic, only those works that were directly relevant to the scope and objectives of this study were carefully selected and cited.

With regard to the analysis of organization and representation in research data repositories, Deposita Dados was selected as the subject of this study. This Repositório Comum de Dados de Pesquisa aims to collect and make available datasets generated by Brazilian researchers affiliated with institutions that do not yet have their own infrastructure for data deposit (Deposita Dados 2026). Additionally, the platform includes scientific output resulting from collaborations between Brazilian researchers and foreign teaching and research institutions, acting as a central hub for the integration and dissemination of scientific knowledge produced in the country.

The selection of Deposita Dados as the empirical field for this study is justified by its role as an institutional repository dedicated to the management, preservation, and dissemination of research data. The platform brings together datasets from different fields of knowledge, providing a structured environment that facilitates the observation of practices related to the organization, description, and representation of information.

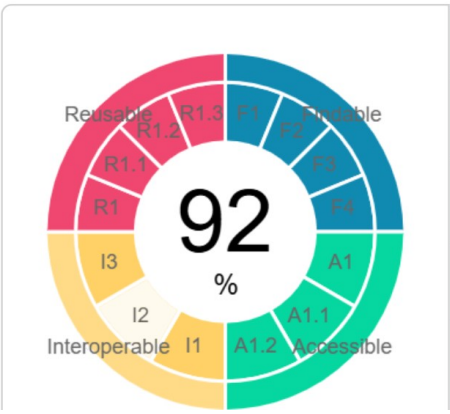
To analyze the repository based on the FAIR Principles, we used the F-UJI (FAIRsFAIR Research Data Object Assessment Service) tool, which is designed to programmatically assess the compliance of research data objects with the FAIR Principles, based on metrics developed by the FAIRsFAIR project. This tool is currently under development. It allows the assessment to be performed using a dataset’s URL or DOI.

The evaluation analyzed five datasets selected from the most recent deposits available in February 2026, to represent current practices regarding data representation and organization in the repository. Since this is a non-thematic repository, the sample included datasets from the fields of Medicine, Chemistry, Computer Science, and Information Science, allowing for the observation of the evaluation criteria across different disciplinary contexts. Although analyzing five datasets with the F-UJI tool enabled a qualitative investigation, the reduced sample limits the findings. Based on the five most recent deposits up to February 2026, the sample provides a temporary diagnosis of representation practices in the Deposita Dados repository. The analysis also revealed how distinct disciplines, including Chemistry, Medicine and Computer Science, apply FAIR principles in describing and organizing digital objects.

4.0 Analysis and Discussion of the Results

Regarding the repository's adherence to the FAIR principles, the evaluation of the selected datasets revealed a compliance rate of 92%. The consistency of the results across all analyzed samples demonstrates the maturity and consistency of the curation, organization, and metadata policies implemented by the institution.

The diagnostic tool used categorizes each pillar—Findable, Accessible, Interoperable, and Reusable—into maturity levels: incomplete, beginner, moderate, and advanced, as shown in Figure 1 below:



	Fair level:
Findable:	advanced
Accessible:	advanced
Interoperable:	moderate
Reusable:	moderate

Figure 1. Summary of the evaluation conducted using the F-UJI tool

Deposita Dados has achieved advanced levels of discoverability and accessibility, demonstrating effective data indexing and access through standardized protocols. Interoperability and reusability were rated as moderate, indicating opportunities to improve structured semantics and usage licenses to support the integration and reuse of scientific knowledge. This asymmetry between the pillars is a common phenomenon in research data repositories. While Findability and Accessibility largely depend on technological

infrastructure and the use of persistent identifiers (such as DOI), Interoperability and Reusability require more granular curation efforts.

In line with the objective of this study, which focuses on representation practices and the quality of metadata in Deposita Dados, the analysis delves into the sub-principles that govern these activities. These elements are intrinsically linked to the Findable, Interoperable, and Reusable pillars, which form the fundamental framework for the effective retrieval and shared use of data.

With regard to the components of the Findable principle,, sub-principle F1—which requires the assignment of global, unique, and persistent identifiers to (meta)data—demonstrated full compliance in the Deposita Dados repository. The audit tool confirmed the systematic application of this requirement across all sampled datasets, attesting that the records operate under robust persistent identification mechanisms. Such adherence is fundamental to ensuring the unambiguous location of digital objects, guaranteeing their long-term traceability and semantic integration with global scientific information ecosystems.

Regarding sub-principle F2, which advocates for the description of data through rich and detailed metadata — a factor that directly impacts reusability — the Deposita Dados repository performed satisfactorily, with full compliance verified across all datasets in the sample. This result attests to the presence of dense and consistent informational descriptions, an essential element for reducing ambiguity, facilitating interpretation by third parties, and significantly expanding the potential for data reuse and scientific discovery.

Furthermore, sub-principle F3, which requires a clear and explicit link between metadata and the identifiers of the data they describe, has been established as a standard practice in Deposita Dados. Additionally, full compliance was observed with sub-principle F4, which advocates for the registration or indexing of (meta)data in search engines. The adherence to this set of criteria enhances the organic discovery of data, highlighting an informational structure that not only ensures visibility and efficient retrieval but also promotes the seamless integration of records into global digital ecosystems.

With regard to the Interoperability principle, it was observed that the Deposita Dados repository incorporates practices aligned with two of its fundamental sub-principles. I1—which advocates for the representation of (meta)data using a formal, accessible, shared, and widely applicable language—was found to be compliant in the datasets analyzed. This result indicates that the platform uses standardized metadata schemas, which facilitates the exchange of information between different systems and ensures that knowledge is represented in a structured manner that is readable by both humans and machines.

Likewise, the inclusion of qualified references to other (meta)data, as provided for by sub-principle I3, was also identified as a practice adopted by the repository. This compliance contributes to the creation of consistent informational links, enhancing the integration, contextualization, and interoperability potential of data within the digital environment.

In sub-principle I2, which stipulates that (meta)data must use vocabularies aligned with the FAIR principles, the Deposita Dados repository was not identified as meeting this criterion in the assessment conducted. However, it should be noted that this sub-principle still has limitations when verified using the automated F-UJI tool. Therefore, the result should be interpreted with caution, as non-compliance may be related to the limitations of automatic assessment and not necessarily to the absence of vocabularies aligned with the FAIR principles. This aspect indicates the need for complementary analyses capable of qualitatively assessing the semantic adequacy of the vocabulary used.

About the Reusable principle, the F-UJI tool implements the assessment by breaking down its sub-principles, transforming the four original criteria into five specific indicators. In this context, sub-principle R1 was divided into two assessment items: R1.01 and R1.1-01.

The first criterion, R1.01, which verifies whether the metadata adequately describes the data content, was met by the Deposita Dados repository, demonstrating that, although the metadata standard adopted is of a general nature, it includes sufficient information for the description and retrieval of the datasets. The second sub-criterion, R1.1-01, which evaluates whether the metadata includes information on the reuse licence, was also identified as a practice adopted by the repository. This result reinforces transparency regarding the conditions for data use and complements aspects observed in other evaluation criteria, thereby helping to expand the potential for reuse and the legal certainty associated with the reuse of the analyzed datasets.

The R2 sub-principle, which evaluates whether the metadata includes provenance information regarding the creation or generation of the data, was classified by the tool as partially satisfied across all analyzed datasets. Through the observation of the datasets, it was verified that the deposit date is provided; however, no information is available regarding possible modifications made subsequently.

Regarding sub-principle R3, R3.01-01 evaluates whether metadata follows standards recommended by a scientific community. Deposita Dados did not fully satisfy this requirement because, as a non-thematic repository, it adopts broader metadata standards rather than discipline-specific models. Although this affected the automated score, it does not compromise metadata quality. The R3.01-02 sub-principle, which verifies whether the data are available in file formats recommended by a specific scientific community, was fully satisfied across all evaluated datasets, according to the tool. This demonstrates that, even though the repository does not belong to a specific domain, it is committed to making data retrievable.

The analysis shows that Deposita Dados employs consistent practices for describing and organizing metadata, in line with the FAIR principles, particularly with regard to discoverability, accessibility, and potential for reuse. Full adherence to persistent identifiers, rich metadata, and search engine indexing demonstrates an information architecture that supports the retrieval, understanding, and circulation of data in digital environments. The moderate or partially assessed scores for semantic interoperability and vocabulary adoption reflect the limitations of automated evaluation and the multidisciplinary nature of the repository, rather than descriptive weaknesses. The use of

general-purpose metadata standards prioritizes flexibility and interoperability across domains, while maintaining the informational quality of the dataset descriptions.

Table 1 Summary of compliance with the FAIR principles in the Deposita Dados Repository.

PRINCIPLE	LEVEL OF MATURITY	SUBPRINCIPLES MET	NOTES
Findable	Advanced	F1, F2, F3 and F4	Full compliance. Robust use of persistent identifiers, rich metadata, and efficient search engine indexing.
Accessible	Advanced	(General)	High efficiency in data availability through standardized protocols and a robust technological infrastructure.
Interoperable	Moderate	I1 and I3	It uses formal language and precise references. The moderate rating is due to the absence of specific vocabulary (I2), partly because of limitations in the F-UJI tool.
Reusable	Moderate	R1.01, R1.1-01 and R3.01-02	Metadata accurately describes the content and includes usage licenses. It does not meet R3.01-01 because it is a multidisciplinary (non-thematic) repository.

Overall, the analysis suggests that the repository's metadata structure supports sound information representation practices, ensuring identifiability, contextualization, and clear conditions for reuse. This indicates a level of descriptive maturity consistent with contemporary scientific data environments, although there are opportunities for improvement regarding semantic standardization and enhanced interoperability.

The analysis shows that Deposita Dados performs efficiently in infrastructure and basic description, achieving 92% compliance in localization and accessibility. However, moderate scores in interoperability and reuse reflect the challenge between broad coverage and semantic specialization. While identifiers and access protocols are largely

automatable, semantic interoperability and community standards depend on human curation and controlled vocabularies, which are not always suitable for multidisciplinary repositories. Thus, these moderate scores reflect limitations of generalist platforms, not technical flaws.

From a critical perspective, the discrepancy observed between descriptive richness (F2) and compliance with FAIR vocabularies (I2) suggests that the repository possesses an "informational density" that the F-UJI tool is not yet capable of fully processing. This reality raises an important debate concerning the sovereignty of human versus automated evaluation: the fact that a metadata standard is general in nature (such as Dublin Core) ensures retrieval by humans but may penalize the repository in algorithmic audits that seek complex ontologies. Therefore, although the maturity of Deposita Dados is advanced, it reveals that the future of Open Science in Brazil will depend on the creation of metadata application profiles capable of "translating" the thematic diversity of Brazilian research (from Medicine to Computer Science) into schemes that machines are able not only to read, but also to interpret. The analysis reaffirms that information organization within the research data landscape is a dynamic process, in which technique must serve scientific utility rather than merely the cold metrics of auditing algorithms.

The analysis indicates that Deposita Dados' metadata adequately supports descriptive representation but faces limitations in structured semantic representation. While the richness identified in sub-principle F2 ensures resource identification and syntactic interoperability, full semantic interoperability requires machine-processable encoding schemes. The lack of compliance with sub-principle I2 suggests that, although the datasets are understandable to researchers, they still lack Linked Data mediation, limiting their use by AI agents and automated discovery systems that rely on URIs and formal ontologies.

By adopting general-purpose metadata standards, Deposita Dados ensures broad governance but shifts the responsibility for terminological disambiguation to users and external systems. Improving metadata quality in terms of reusability requires not only descriptive fields but also domain-specific metadata layers capable of supporting scientific communities. Thus, the representation of information in the repository must evolve through Semantic Web technologies and interconnected knowledge graphs.

5.0 Final considerations

The investigation evaluated the compliance of information representation practices and metadata quality of the Deposita Dados repository according to the FAIR Principles. The results indicate that the repository plays a strategic role in the Open Science ecosystem in Brazil, achieving 92% compliance in the analyzed datasets. The platform demonstrated advanced maturity in Findability and Accessibility, with the use of DOIs and rich descriptions that guarantee traceability and global location of scientific data.

The moderate ratings for Interoperability and Reuse reflect the multidisciplinary nature of Deposita Dados. By adopting general-purpose metadata standards rather than domain-specific ontologies, the repository prioritizes flexibility, although this limit automated

semantic interoperability. Nevertheless, the metadata provides transparency and legal certainty for scientific reuse. However, the small sample size limits generalizability, as outliers may disproportionately influence perceptions of the repository's maturity.

Additionally, the developmental stage of the F-UJI tool must also be considered, since, as a system undergoing continuous improvement, it may present limitations in the assessment of specific subprinciples (such as I2 and R3). The analysis of a limited number of objects reduces the possibility of correlating potential metric inconsistencies within the tool with a statistically significant body of data. Therefore, this investigation assumes an eminently descriptive character, positioning itself as an initial case study that indicates pathways for strengthening Open Science in Brazil.

This study supports managers and funding agencies by validating Deposita dados as a central infrastructure for preserving Brazilian scientific memory. Future studies should include large-scale quantitative and qualitative analysis focused on the semantic adequacy of adopted vocabularies. Despite sample limitations, the findings confirm that the repository promotes active practices that enhance the reuse of Brazilian scientific knowledge.

DECLARATION OF NON-USE OF ARTIFICIAL INTELLIGENCE: I hereby declare, for all intents and purposes, that this work was prepared by the authors, and that we bear full responsibility for the research, analysis, writing, and conclusions presented. I further declare that we did not use Generative Artificial Intelligence (GAI) tools.

References

Bonetti, Letícia Guarany, and Ana Carolina Simionato Arakaki. 2022. "A importância das boas práticas em repositórios de dados de pesquisa." In Anais do XXII Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação (XXII ENANCIB), 1–13. Porto Alegre: ANCIB.

Borgman, Christine L. 2010. Research Data: Who Will Share What, with Whom, When, and Why? RatSWD Working Paper no. 161. Berlin: German Data Forum (RatSWD). https://doi.org/10.2139/ssrn.1714427.

Deposita Dados. 2026. "About." Accessed March 13, 2026. https://depositadados.ibict.br/about.xhtml.

Diniz, Bárbara Carvalho, Guilherme Ataíde Dias, and Alzira Karla Araújo da Silva. 2018. "Os princípios FAIR como diretrizes para maximizar o uso e (re)uso dos dados de pesquisa: traçando um paralelo com a gestão de documentos arquivísticos digitais." In Anais WIDaT 2018: II Workshop de Informação, Dados e Tecnologia, edited by Ricardo César Gonçalves Sant'Ana, Moisés Lima Dutra, and Guilherme Ataíde Dias, 123–135. João Pessoa: Editora UFPB. https://doi.org/10.22477/ii.widat.142.

Gäal, Lígia Parreira Muniz, and Márcio Souza Martins. 2022. “Acesso aberto no contexto da pesquisa em Ciência da Informação.” Transinformação 34: e220016. https://doi.org/10.1590/2318-0889202234e220016.

Guizzardi, Giancarlo. 2020. “Ontology, Ontologies and the ‘I’ of FAIR.” Data Intelligence 2: 181–191. https://doi.org/10.1162/dint_a_00040.

Santos, Raimunda Fernanda, and Carla Beatriz Marques Felipe. 2025. “Repositórios de dados de pesquisa brasileiros à luz da Organização e Representação da informação.” ISKO Brasil 8. https://doi.org/10.22477/ISKO25.70.

Vogt, Lars, Philip Strömert, Nicolas Matentzoglu, Naouel Karam, Marcel Konrad, Manuel Prinz, and Roman Baum. 2025. “Suggestions for Extending the FAIR Principles Based on a Linguistic Perspective on Semantic Interoperability.” Scientific Data 12: 688. https://doi.org/10.1038/s41597-025-05011-x.