TransInformação

ORIGINAL

Editor
Luisa Angélica Paraguai Donati

Conflict of interest
The authors declare that they have no conflicts of interest.

Data Availability
The research data are available within the body of the document.

Received
May 7, 2025

Approved
October 16, 2025

Scientific production on data repositories and open science published in the Web of Science database: Methodi Ordinatio and content analysis

Produção científica sobre repositórios de dados e ciência aberta publicada na base Web of Science: Methodi Ordinatio e análise de conteúdo

Sinval Adalberto Rodrigues-Junior[1], Marcelo Votto Texeira[1]

[1] Universidade Comunitária da Região de Chapecó, Curso de Biblioteconomia, Departamento de Ensino à Distância. Chapecó, SC, Brasil. Correspondence to: S. A. RODRIGUES-JUNIOR. E-mail: <rodriguesjunior.sa@unochapeco.edu.br>.

How to cite this article: Rodrigues-Junior, S. A.; Texeira, M. V. Scientific production on data repositories and open science published in the Web of Science database: Methodi Ordinatio and content analysis. *Transinformação*, v. 37, e2513075, 2025. https://doi.org/10.1590/2318-0889202537e2513075

Abstract

The opening of scientific data proposed by the Open Science movement presupposes careful planning for data collection, organization, and treatment, aiming at their sharing, accessibility, and reuse. Data repositories have been conceived as structures necessary to enable open access to data. This study aimed to analyze the influence of data repositories on the disclosure and sharing of scientific data proposed by the Open Science movement. The Methodi Ordinatio, developed to organize a portfolio of scientific publications, was adopted to analyze the subject of 'Data Repositories' and 'Open Science'. The studies were ranked using the InOrdinatio index, and the 15 best ranked studies were included and analyzed through Bardin's content analysis. Most studies describe the structure involved in data repositories within the biological, chemical, and health areas. Other studies addressed data reuse, data organization and analysis processes and tools, as well as data selection and classification algorithms. The units of analysis selected for the content analysis were categorized as open access, information technologies, data processing, and information retrieval. Systems (processes and structures), metadata standards, ontologies, semantic web, data types, and their management were addressed by these studies. It is concluded that open data repositories are growing rapidly. Production with the greatest impact has occurred in the biological and biomedical/health areas, highlighting the structure involved in repositories within these fields. Data repositories provide systems for depositing, managing, searching, accessing, and reusing data based on processes and technologies – often developed as open-source software – in alignment with the proposed Open Science model.

Keywords: Content analysis. Data repository. Open access. Open science.

CC BY

Resumo

A abertura de dados proposta pelo movimento da Ciência Aberta pressupõe um planejamento para a coleta, organização e tratamento desses dados, visando ao compartilhamento, à acessibilidade e a reutilização deles. Os repositórios de dados foram planejados como estruturas que permitem o acesso aberto a esses conteúdos. Este estudo teve como objetivo analisar a influência dos repositórios de dados na abertura e compartilhamento de dados científicos propostos pelo movimento da Ciência Aberta. O Methodi Ordinatio, desenvolvido para organizar um portfólio de publicações, foi adotado para analisar o tema "Repositórios de Dados" e "Ciência Aberta". Os estudos foram classificados segundo o índice InOrdinatio, e os 15 estudos mais bem classificados foram analisados por meio da análise de conteúdo de Bardin. A maioria dos estudos descreve a estrutura de repositórios de dados nas áreas biológicas, químicas e de saúde. Os demais abordaram a reutilização de dados, processos e ferramentas de organização e análise de dados e algoritmos de seleção e classificação de informações. As unidades de análise selecionadas para a análise de conteúdo foram categorizadas em acesso aberto, tecnologias de informação, processamento de dados e recuperação de informação. Sistemas (processos e estruturas), padrões de metadados, ontologias, web semântica, tipos de dados e seu gerenciamento foram abordados nesses estudos. Conclui-se que os repositórios de dados abertos estão crescendo rapidamente. A produção de maior impacto ocorreu nas áreas biológica e biomédica/de saúde e destaca a estrutura envolvida nesses repositórios. Os repositórios de dados fornecem sistemas de depósito, gerenciamento, busca, acesso e reutilização de dados baseados em processos e tecnologias, em sua maioria desenvolvidos em software de código aberto, alinhando-se ao modelo de Ciência Aberta proposto.

Palavras-chave: *Análise de conteúdo. Repositório de dados. Acesso livre. Ciência aberta.*

Introduction

The scenario of scientific production has undergone significant changes over the last 20 years with the introduction of the Open Science paradigm. This movement is guided by the understanding that scientific knowledge should be widely and openly shared for the benefit of all people (United Nations Educational, Scientific and Cultural Organization, 2022). Notably, scientific knowledge is considered a driving force for economic growth and for solving complex problems in an interconnected world (Silva; Silveira, 2019; United Nations Educational, Scientific and Cultural Organization, 2022). Hence, the more open and global its production is, the faster these problems can be solved.

The Open Science model is a contemporary approach that aims to democratize access to scientific knowledge, strengthen the transparency of research processes, and expand the social impact of science. Among its main advantages are free access to publications and data, which reduce economic barriers and promotes greater equity in the dissemination of knowledge (United Nations Educational, Scientific and Cultural Organization, 2021). Furthermore, interdisciplinary collaboration and information sharing foster innovation and scientific efficiency, avoiding duplication of efforts and optimizing public resources (European Commission, 2016; Fecher; Friesike, 2014). Methodological transparency, supported by the openness of protocols, codes, and databases, contributes to the reproducibility and reliability of results (Nosek *et al.*, 2015). From a social perspective, the model encourages knowledge transfer and the practical application of discoveries, driving public policies and evidence-driven technological solutions (Vicente-Saez; Martinez-Fuentes, 2018). Finally, by favoring the participation of historically underrepresented countries and groups, Open Science promotes greater inclusion and cognitive justice in the global scientific system (Chan *et al.*, 2020). Thus, it has consolidated itself as a paradigm that combines scientific rigor, social responsibility, and equity in access to knowledge.

Despite the undeniable gains of opening the production of scientific knowledge, Open Science faces a series of challenges, including the need to prepare scientists for a necessary change in behavior regarding their activity. Incentives and technical training for data sharing are on the agenda (Walters, 2020). Additionally, an adequate information infrastructure is required and involves practices from library science, archival science, and data management (Sayão; Sales, 2023). There is also a demand for technological resources, namely servers, hardware, and software. In this sense, open data repositories are foundational elements of the Open Science ecosystem. Pampel *et al.* (2023) counted more than 3,000 data repositories registered in re3data, signaling a growing increase in the ecosystem that meets re3data's quality and openness requirements. Moreover, along with data repositories, data journals have been listed as possible sources of open data and are seen as a response to the need for recognition of data authors (Kindling; Strecker, 2022; Walters, 2020).

The open science paradigm

One of the first manifestations toward the opening of access to scientific information came from the Budapest Open Access Initiative (BOAI) in December 2001. This group advocated for the acceleration of efforts to make studies freely available to the global public (Budapest Open Access Initiative, 2002). Their initial approach recommended the practice of self-archiving journal articles in open electronic repositories by authors (green way) and the development of journals committed to open access (gold way) (Budapest Open Access Initiative, 2002). Therefore, the recommendation stands in opposition to the traditional publishing model established by commercial publishers who, in addition to charging researchers and institutions where the research is developed to publish, also profit from subscribers willing to access the scientific information (Silva; Silveira, 2019).

Ever since, Open Science has become a topic of discussion in the scientific community. In 2020, the United Nations Educational, Scientific and Cultural Organization (UNESCO) conducted a global consultation about Open Science, which presented the perceptions of the topic across countries from the five continents. Integration of Open Science through the entire research cycle, including open methodology, open source, open peer review, open software, open education, research evaluation metrics, and citizen science was emphasized (United Nations Educational, Scientific and Cultural Organization, 2020). It also gave birth to the UNESCO Recommendations on Open Science, which are based on four pillars: open scientific knowledge, open science infrastructures, open involvement of societal actors (scientific communication) and dialogue with other knowledge systems (Figure 1) (United Nations Educational, Scientific and Cultural Organization, 2021).

Open scientific knowledge involves scientific publications, open research data, open educational resources, free and open-source software, and open hardware. Open scientific infrastructures refer to virtual and physical infrastructures, equipment, instruments, information sources, files, scientific data, among others, which make scientific work more collaborative. The involvement of societal actors is associated with collaboration between scientists and members of society beyond the scientific community. Crowdfunding, third-party contributions, scientific volunteering, and citizen science are some related concepts. Finally, openness to other knowledge systems, such as local communities, marginalized academies and traditional populations constitutes the fourth foundational pillar of the Open Science movement (United Nations Educational, Scientific and Cultural Organization, 2021).

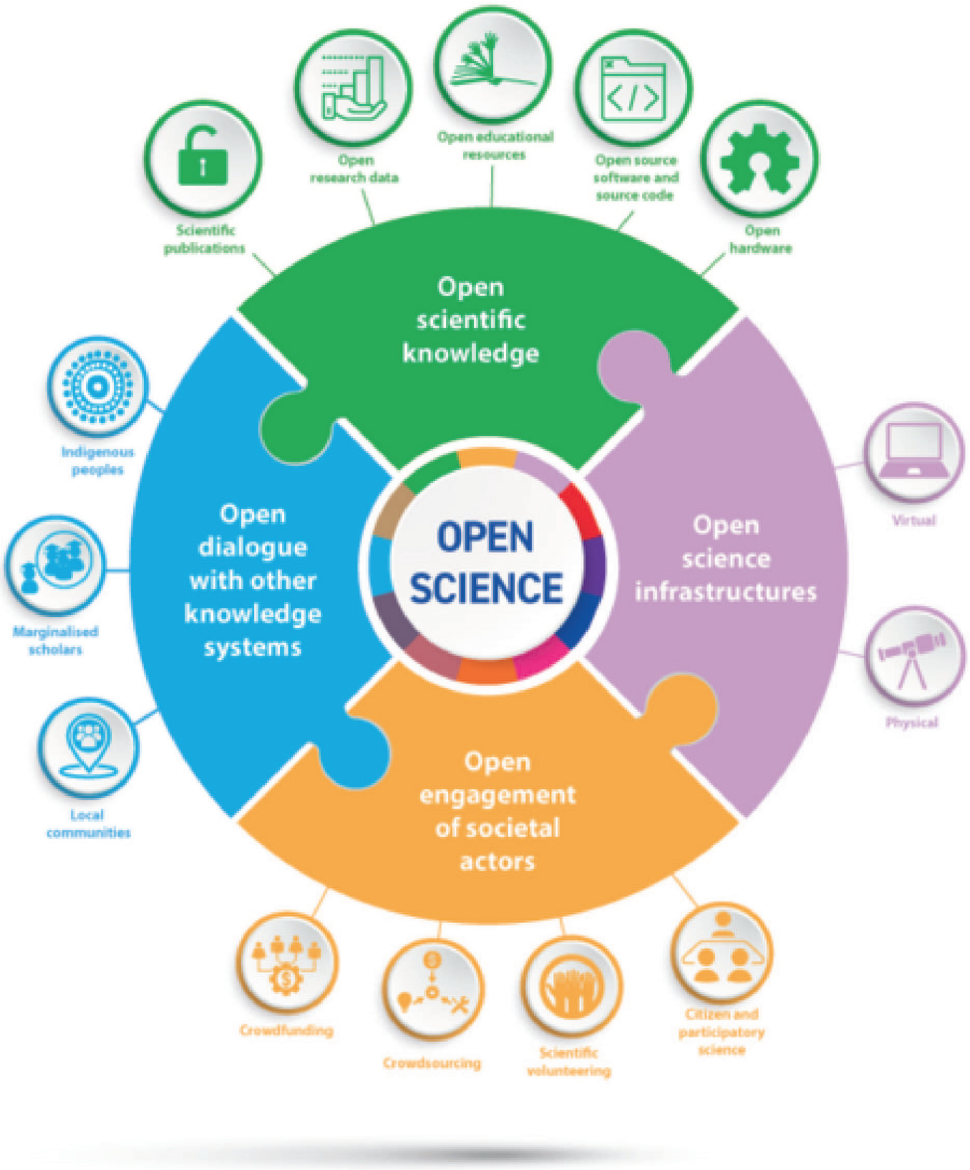

Figure 1 – Pillars of the UNESCO Recommendation on Open Science.
Source: United Nations Educational, Scientific and Cultural Organization (2021).

Open data within the open science paradigm

Open data is the second dimension of the Open Science Taxonomy described by Pontika *et al.* (2015) and expanded by Silveira *et al.* (2023). It addresses open big data, open data standards, the use and reuse of open data, requirements for opening data, open data visualization, government open data, open administrative data, FAIR principles, and public data (Silveira *et al.*, 2023). Open data, therefore, is a key dimension of the Open Science paradigm. Importantly, most of these sub-dimensions are interrelated. Open research data include digital and analogical data, raw and processed data, metadata, text records, images and sounds, protocols, analysis codes, and workflows (Silveira *et al.*, 2023; United Nations Educational, Scientific and Cultural Organization, 2020).

When it comes to Open Data in the revised Open Science taxonomy, the first subdivision of the nine refers to Open Big Data (Silveira *et al.*, 2023). Big data comprises large, complex sets of

data with interaction potential that cannot be processed with standard management techniques. Decision-making in several fields, including business, government and public administration, and health care, amongst others, has relied on properly analyzed big data (Acharjya; Kauser, 2016; Ularu *et al.*, 2012). Big data is based on four aspects, the so-called 4Vs: veracity, variety, volume, and velocity. Veracity stands for the truthfulness of the information generated with data, so decisions can be made based on it. Volume stands for data gathered, velocity stands for the processing time required to analyze big data, and variety stands for the variability of data types involved in big data (Ularu *et al.*, 2012). Open data standards are required so that open data may be correctly identified, cited, and reported (Sansone *et al.*, 2019). Such standards involve guiding principles or checklists that signal the information required to contextualize the object, dictionaries and ontologies to properly and unambiguously define the object, models and formats of structures that include transmission formats to ease the exchange of data between systems, and identifiers of digital objects that confer unique identities to the objects. These standards have been developed by communities under the FAIRsharing flag (https://fairsharing.org/) (Sansone *et al.*, 2019).

FAIRsharing derives from the FAIR principles, an acronym that stands for Findable, Accessible, Interoperable and Reusable open data (Wilkinson *et al.*, 2016). Findability, in this context is associated with the existence of a globally unique and persistent identifier, data described with rich metadata that includes the identifier of the object being described, and indexation in a searchable source. Accessibility relates to the retrieval of data by the identifier under a standard, open, and free communication protocol, and the maintenance of metadata even when data is no longer available. Interoperability is based on metadata composed by formal, accessible, shared, and widely applicable language, by vocabularies that follow the FAIR principles, and metadata that references other metadata. Finally, reusability relies on richly described metadata, released with a clear and accessible usage license, detailed data provenance, and metadata meeting domain-relevant community standards (Wilkinson *et al.*, 2016).

According to Wilkinson *et al.* (2016), open data is not an aim in itself; rather, it serves to provide open data as a valuable digital resource and an opportunity for the scientific community to pursue new discoveries and innovations. It is expected that data collection efforts will not be duplicated, and that data sharing and reuse will bring the scientific community closer together for collaboration.

As part of the eScience movement, the open data sharing has been extended beyond science, to government and administrative data. Advocates propose that governments can enhance transparency in stewardship by opening data in public repositories. The Open Government Partnership was launched in 2011 by a group of eight founding countries, including Brazil, and now extends to more than 80 countries (Sena; Segundo; Melo, 2022). Countries commit to initiatives that promote transparency, access to public information, and citizen participation. Each government elaborates two-year action plans focusing on state commitments to the open government principles – namely transparency, accountability, citizen participation, technology, and innovation. The execution of the proposed activities is monitored during the plan's implementation and, at the end of the two years, a new plan is implemented (Bertin *et al.*, 2019).

Data sharing in open data repositories

For one to deposit data in an open data repository, data management planning is required prior to the beginning of the research. Tools are available for these data management planning, such as DMPTool and DataUp (Strasser; Cruse, 2013, Strasser; Abrams; Cruse, 2014). A successful

data management plan would allow others to find, access, understand, and reuse data over a 20-year span (Strasser *et al.*, 2011). Figure 2 depicts the data management life cycle from the perspective of the researcher.

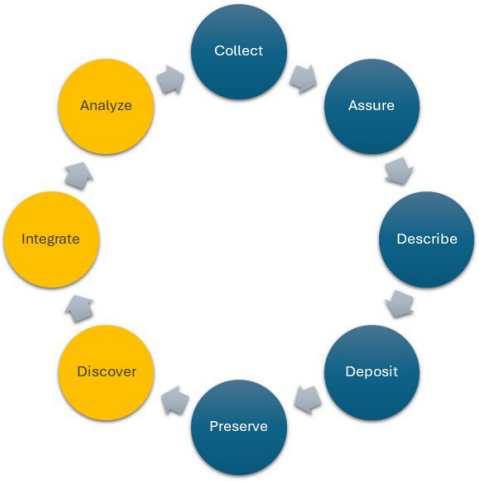

Figure 2 – Data management life cycle.
Source: Strasser *et al.* (2011).

Data collection also corresponds to the creation of the dataset and should explicitly include a sampling plan, the description of data collection and analysis, including the rationale for the methods chosen along with contextual information. The organization of data should consider the file formats used and the contents to be generated. Data management requires the definition of who is responsible for it, as well as the definition of the data version and how it will be handled, the necessary backups, and their frequency. Data description relies on the metadata. How metadata will be recorded, which metadata standard will be adopted, which tool will be used, and where it will be deposited are some of the questions to be asked when considering data description. Data sharing may occur even during the research process, within the research team. After the end of the research, sharing with a broader scientific community may also be planned. Similarly, data preservation may be required during the research execution and after its end. Preservation must consider archival requirements for data, documentation, and metadata (Strasser *et al.*, 2011).

Although technological issues are required to achieve the best experience in data sharing in science, digital technology keeps evolving to provide researchers with data to analyze (Carlson, 2006). The speed of development of the required digital infrastructure is high and supports the growing demand of data sharing under the eScience context (Greenberg *et al.*, 2009). It seems that the challenges to data sharing are more related to established practices adopted by researchers, motivated by different factors, including the perceived value of their time, concerns about misinterpretation and misuse of data, and lack of expertise in data sharing (Tenopir *et al.*, 2011). Moreover, recognition for data sharing has yet to be fully established, and it seems that the scientific community must be involved, including funders, research societies, universities, and scientific rating organizations. Issues that deserve attention include credit attribution for data sharing, support for education, and community standards for data sharing.

In light of the existing challenges related to data sharing in open data repositories, this study aimed to identify the most relevant studies on data repositories and open science and to analyze their content.

Methodological Procedures

The study was conceived as a quantitative-qualitative study, based on the context of scientific production on data repositories and open science indexed in the Web of Science (WoS) database. The Methodi Ordinatio, proposed by Pagani, Kovaleski and Resende (2015) was applied and adapted to achieve the proposed objectives. This method proposes a strategy for building a portfolio of relevant studies to answer the research question, following a systematized review proposal. It adopts a multi-criteria model for evaluating and ranking publications (InOrdinatio) based on the impact factor, year of publication, and number of citations to determine the relevance of studies to the field (Pagani; Kovaleski; Resende, 2015).

Methodi Ordinatio is composed of nine phases, as follows: Phase 1) – Establishment of research intention; Phase 2) – Preliminary exploratory research with keywords in the databases; Phase 3) - Definition and combination of keywords and databases; Phase 4) – Final search in databases; Phase 5) – Filtering procedures; Phase 6) – Identification of the impact factor, year of publication and number of citations; Phase 7) – Ranking of articles using InOrdinatio; Phase 8) - Search for full articles; Phase 9) – Final reading and systematic analysis of articles.

Next, the way in which the nine phases were applied in this study is described.

Phase 1) – Establishment of research intention: the initial research theme was defined as the requirements for building a data repository.

Phase 2) – Preliminary exploratory research with keywords in databases: the Web of Science (WoS) database was chosen because it provides the largest number of studies on the topic in an exploratory and comparative search with Scopus. Also, it was chosen because it provides one of the metrics for ranking studies, the impact factor, in addition to the number of citations. At this stage, a preliminary, broader search on the topic was carried out using the following search strategy:

'data repository OR data repositories OR open data repository (All fields) AND open science OR open data OR open data repository ecosystem OR open data repository eco-system (All fields) NOT Document types: Early access'

The search was carried out on November 8th 2023. Documents in press (with early access) were removed from the search strategy itself because the publication date was not established.

Phase 3) – Definition and combination of keywords and databases: Based on the analysis of indexed keywords and authors' keywords, a thematic focus was defined, considering the frequency of keywords and thematic growth trends. Therefore, it was decided to address the association between the theme "Open Science" and its derivations with data repositories. The question to be answered was: "What is the impact of the Open Science model on the construction of data repositories?"

Phase 4) – Final search in databases: The final search strategy was structured based on words representing 'data repositories' and words that represented the context of Open Science, according to the taxonomy defined in Pontika *et al.* (2015). Finally, early access studies were removed, as were conference papers, data articles, and editorial material. Therefore, only published original articles and review articles made up the portfolio.

'"data repository" OR "data repositories" (All fields) AND "open science" OR "open access" OR "open data" OR "open science policies" OR "open science tools" OR "open repositories" OR "open data repository" OR "open data repository ecosystem" OR "open data repository eco-system" (All fields) NOT Document types: Early access NOT Document types: Conference paper OR Data paper OR Editorial material'

Phase 5) – Filtering procedures: the complete WoS records database was downloaded in Excel format (.xlsx) and studies were selected considering whether the title represented the topic under study. A single researcher selected the studies, and, in case of doubt, the researcher kept the study among those eligible.

Phase 6) – Identification of the impact factor, year of publication and number of citations: the metric adopted was the WoS Impact Factor (IF). When the journal did not present an IF, the Journal Citation Indicator (JCI) – also provided in WoS – was considered. The year of publication and the number of citations across all databases were also collected from the database.

Phase 7) – Ranking of articles using InOrdinatio: the selected articles were ranked using the following formula:

$$InOrdinatio = (FI/1000) + \alpha * [10 - (Year\ of\ research - Year\ of\ publication)] + (\Sigma Ci)$$

Where FI is the impact factor, α is a weighting factor for the weight of the year of publication, which can vary from 1 to 10; the year of research is the year in which the study is being carried out, the year of publication is the year in which the article was published, and ΣCi is the sum of citations. Regarding the weighting factor of the year of publication, it is considered that topics in which more recent studies are more relevant can adopt weighting factors closer to or equal to 10, while topics in which the date of publication is irrelevant or which tends to consider older studies as more relevant, brings the weighting factor closer to 1. In this study, the weighting factor adopted was 10, considering the current relevance of the studies in the context analyzed.

Phase 8) – Search for full articles: the 15 studies with the highest InOrdinato index were included in the study.

Phase 9) – Final reading and systematic analysis of the articles: at this stage, the reading, interpretation, and synthesis of the texts were conducted following the content analysis method proposed by Bardin (2020). First, an exploratory reading of the studies was carried out. Then, excerpts (units of analysis) relevant to answering the research question were selected and represented by thematic categories and subcategories.

Results

The results of the final search (Phase 4) indicated a total of 545 documents. Filtering these documents by title led to the exclusion of 358 records, resulting in 152 remaining documents. The InOrdinatio index was applied to all 152 documents and Table 1 presents the 15 studies with the highest InOrdinatio index included in the study.

Table 1 – Included studies based on InOrdinatio ranking.

1 of 2

Ranking of Phase 7	Studies (first author, year, title, and journal)	Country	Impact factor (Phase 6)	Citations (Phase 6)	Year (Phase 6)	InOrdinatio (Phase 7)
1	Smedley (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*	Italy	14.9	475	2015	495.0149
2	Bhattacharya (2018). ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific Data*	USA	9.8	372	2018	422.0098
3	Sud (2016). Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*	USA	14.9	376	2016	406.0149

Table 1 – Included studies based on InOrdinatio ranking.

Ranking of Phase 7	Studies (first author, year, title, and journal)	Country	Impact factor (Phase 6)	Citations (Phase 6)	Year (Phase 6)	InOrdinatio (Phase 7)
4	Okuda (2017). jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Research*	Japan	14.9	302	2017	342.0149
5	Taylor (2017). The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage*	United Kingdom	5.7	272	2017	312.0057
6	Guo (2020). CNSA: a data repository for archiving omics data. *Database*	China	5.8	147	2020	217.0058
7	Piwowar (2013). Data reuse and the open data citation advantage. *PeerJ*	USA	2.7	176	2013	176.0027
8	Shi (2022). DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides. *Nucleic Acids Research*	China	14.9	53	2022	143.0149
9	Michener (2015). Ecological data sharing. *Ecological Informatics*	USA	5.1	123	2015	143.0051
10	Kearnes (2021). The Open Reaction Database. *Journal of the American Chemical Society*	USA	15	61	2021	141.015
11	Moriya (2019). The jPOST environment: an integrated proteomics data repository and database. *Nucleic Acids Research*	Japan	14.9	56	2019	116.0149
12	Nargesian (2018). Table Union Search on Open Data. *Proceedings of the VLDB Endowment*	Canada	2.5	61	2018	111.0025
13	Bishop (2017). Revisiting Qualitative Data Reuse: A Decade On. *SAGE Open*	United Kingdom and Finland	2	68	2017	108.002
14	Zheng (2021). DrugComb update: a more comprehensive drug sensitivity data repository and analysis portal. *Nucleic Acids Research*	Finland	14.9	27	2021	107.0149
15	Gad (2022). An improved binary sparrow search algorithm for feature selection in data classification. *Neural Computing and Applications*	Egipt and Australia	6	15	2022	105.006

Source: Prepared by the authors (2023).

Ten out of the 15 studies included (66.7%) described open data repositories, eight of which (53.3%) were from the biological field. The United States was the country with the highest number of studies (*n*=5), followed by Japan and China (*n*=2). Eight of the repositories belonged to the biological field, one of chemistry and one of health. The other five studies addressed data reuse (Piwowar; Vision, 2013), sharing data on ecology (Michener, 2015), merging tables of open data from different sources (Nargesian *et al.*, 2018), reuse of qualitative data (Bishop; Kuula-Luumi, 2017) and data selection and classification algorithms (Gad *et al.*, 2022).

The content analysis of the 15 studies resulted in 57 units of analysis representing four distinct categories: "*open access, information technologies, data processing, and information retrieval*" (Figure 3).

Open data repositories enable open access to data, promoting transparency and accessibility. Twelve units of analysis highlighted this relationship:

> Study 2 (p. 2) – The goal of ImmPort is to ensure that basic and clinical research data are accessible to researchers in ways that allow effective sharing of data and knowledge. In order to facilitate this process, the ImmPort ecosystem includes four major applications: Private Data, Shared Data, Data Analysis, and Resources.
>
> Study 10 (p. 18824) – Commitment to Open Access. As the name implies, the Open Reaction Database is designed for open access and community contributions.

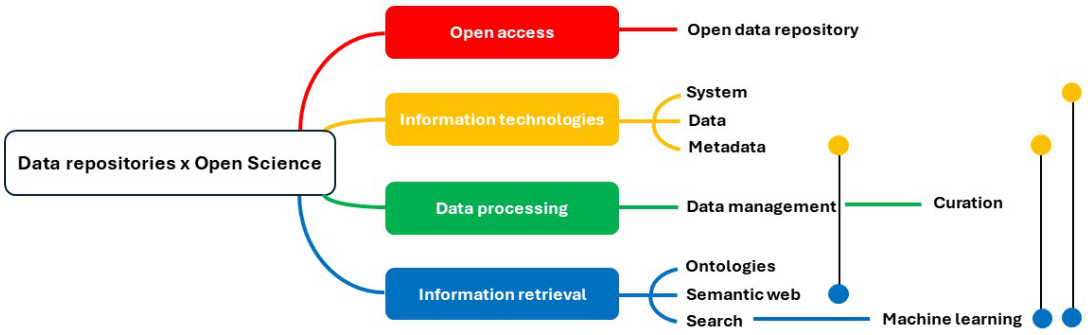

Figure 3 – Structure of categories presented by the assessed studies.
Source: Prepared by the authors (2023).

Furthermore, repositories provide *information technologies* such as systems, metadata and the data itself. Twelve units of analysis referred to systems including relational data management systems written in open source, graphical and programmatic interfaces (data sets, filters and attributes), databases and data analysis systems written in different open codes and machine learning algorithms aimed at combining data:

> Study 2 (p. 2) – The ImmPort database architecture is designed to support and maintain a variety of multi-modal immunological data such as study method documentation, metadata and standardized data formats and terminologies. Together, these efforts facilitate accurate and efficient secondary analysis of large-scale immunology data.
>
> Study 6 (p. 2) – Based on the Django framework which is a high-level Python Web framework for web development and maintenance (https://www.djangoproject.com/), CNSA is developed in Python.

Another four units of analysis referred to metadata standards, as follows:

> Study 9 (p. 41) – Data products cannot be re-used unless the context, structure, collection and processing methods, and quality of the data are sufficiently documented. Ideally, all aspects of the data are documented throughout the entire project from planning and hypothesis formulation through QA/QC and metadata creation through analysis and dissemination. Creating comprehensive metadata is most effective when researchers are routinely documenting data collection, processing and analysis activities. New tools such as open lab notebooks allow research notes and data to be published online as they are created. Metadata management is facilitated when standards such as ISO 19115 and EML are adopted and comprehensive, user-friendly tools like Morpho are employed to create, manage, and disseminate the project's metadata.

Besides, another nineteen units of analysis identified the data itself as a constituent of information technologies. These units present elements related to data analysis, such as open-source technologies and programs, data management tools, incentives and barriers to data sharing, citations and metrics related to data:

> Study 7 (p. 9) – Citations were 9% higher for papers with available data, independent of other variables ($p < 0.01$, 95% confidence intervals [5% to 13%]).
>
> Study 9 (p. 36) – Funders, journals and professional societies can each drive sociological change with respect to data sharing. Establishing and enforcing mandates for data archiving greatly increase the likelihood that data will be

> available for the long-term (Vines *et al.*, 2013). Moreover, funders, publishers and professional societies can all contribute to incentivizing and reducing barriers to data sharing by providing credit, supporting education, establishing community standards for data and data sharing, and streamlining approaches to data submission.
>
> Study 9 (p. 41) – The DMPTool and DMPonline are tools that make it easy for researchers to create an initial data management plan that meets the requirements for a particular research sponsor in the USA and UK, respectively (see https://dmptool.org/ and https://dmponline.dcc.ac.uk/). The DMPTool also allows one to share a plan with the project team and publish it openly for broader viewing and attribution.

The *data processing* category was represented by four units of analysis, and addressed data generation and data management, including its curation. The unit below expresses this context:

> Study 9 (p. 41) – Before a project gets underway, researchers should have a plan for how the data will be managed. Plans should cover: (1) data collection and processing methods, organization in tables or databases, and relevant access and use policies; (2) quality assurance and quality control procedures; (3) metadata creation and management; (4) data preservation; (5) integration, analysis, synthesis and dissemination; (6) relevant policies including data sharing plans; and (7) a budget that explicitly details costs (i.e., time and money) for preparing, documenting, and archiving data.

Finally, six units of analysis expressed the information retrieval category. The category involved the concepts search, semantic web and ontologies, all provided by open repositories to provide accessibility to research data:

> Study 11 (p. D1223) – jPOSTdb is constructed based on the Semantic Web technology that facilitates integration of various databases containing big data (29). The protein sequence and annotation databases, such as UniProt and neXtProt, have released data formatted by the Resource Description Framework (RDF) data model, which supports the Semantic Web technology (21,30). In addition, not only the proteome data, but also other omics and life science-related data have been converted to RDF data and released in the NBDC RDF portal (https://integbio.jp/rdf/) and EBI RDF platform (31) and have been incorporated into various life science databases such as TogoGenome, a genomic database based on RefSeq data (http://togogenome.org/), TogoVar, a database of human genome variants/variations (https://togovar.biosciencedbc.jp/) and GlyTouCan, a glycan structure repository (32).

Discussion

The quantitative-qualitative study method was based on Pagani, Kovaleski, and Resende's (2015) Methodi Ordinatio, adapted to answer the proposed question. The adaptations involved the application of the bibliometric method to analyze the scenario of scientific production regarding data repositories in Phases 2 and 4. The Methodi Ordinatio generated a portfolio of studies, among which the 15 best ranked were analyzed. The analysis of the studies attributed to data repositories the collaboration with the openness of scientific data and its transparency (Bhattacharya *et al.*, 2018; Kearnes *et al.*, 2021; Zheng *et al.*, 2021), as proposed by United Nations Educational, Scientific and Cultural Organization (2020). Furthermore, our results indicated repositories proposing strategies for processing data, helping to organizing them for sharing and reuse since their generation by

researchers, in addition to their curation and management (Bhattacharya *et al.*, 2018; Michener, 2015; Zheng *et al.*, 2021). In addition, repositories provide the necessary information technology structure, such as systems, metadata and the data itself (Bhattacharya *et al.*, 2018; Guo *et al.*, 2020; Smedley *et al.*, 2015; Zheng *et al.*, 2021), as well as the structure for information retrieval that involves ontologies, semantic web, and the search processes themselves (Bhattacharya *et al.*, 2018; Gad *et al.*, 2022; Moriya *et al.*, 2019; Okuda *et al.*, 2017). Most studies were not necessarily related to the area of information science but rather described active data repositories in the biological, chemical, and health areas.

The method adopted in the study, involving InOrdinatio to highlight the most relevant studies, considered the criteria involved in calculating the index (Pagani; Kovaleski; Resende, 2015). The decision to adopt a weighting factor (α) equal to 10 together with the year of publication was based on the understanding of the relevance of the recent aspect of the study, in contrast to the total number of citations, known to be lower in more recent studies. On the other hand, the adoption of InOrdinatio may have presented studies that did not necessarily associate open data repositories with information science, which may have generated a selection bias due to the impact factor of the journals.

The studies analyzed revealed the commitment of data repositories to open access to data, in line with what is recommended in the UNESCO Recommendations (United Nations Educational, Scientific and Cultural Organization, 2021). The Recommendations advocate for "open access to scientific publications, research data, metadata, open educational resources, software, source codes and hardware available in the public domain or under copyright and open licenses". In this sense, in addition to the openly expressed commitment to open access in some studies (Bhattacharya *et al.*, 2018; Kearnes *et al.*, 2021; Zheng *et al.*, 2021), others observed the construction of repository systems in open source (Bhattacharya *et al.*, 2018; Smedley *et al.*, 2015); and the use of open data analysis resources (Sud *et al.*, 2016; Zheng *et al.*, 2021).

Importantly, the concept "system", in the context of information retrieval presupposes – beyond the hard aspect of software or technology – "the set of processes, human skills, techniques and technologies, sources of information united in a systemic way with the purpose of effectively providing Information Retrieval" (Lancaster, 1979). Thus, it includes all processes and technologies required to guarantee the findability of information and its use or reuse. Regarding data repositories, the FAIR principles (Findability, Accessibility, Interoperability and Reusability) guide this particular information system, including metadata standards (United Nations Educational, Scientific and Cultural Organization, 2021; Wilkinson *et al.*, 2016). Based on this, models of good metadata practices for data repositories have focused on components of dataset description in order to align with the semantic web proposal and the linked data perspective, enhancing the automatic synthesis of data and its reuse (Greenberg *et al.*, 2009). Interestingly, none of the studies analyzed mentioned FAIR principles. Despite this, some of the studies that described the repositories mentioned the use of systems built on open-source software (Sud *et al.*, 2016), metadata standards based on machine-readable XML and RDF schemas (Okuda *et al.*, 2017) and open-source data analysis tools (Kearnes *et al.*, 2021).

Conclusion

Based on the proposed method, and considering the limitations of the study, it was possible to conclude that the most impactful production has occurred in the biological, biomedical or interdisciplinary fields and highlights the structure involved in existing repositories. Data

repositories provide systems for depositing, managing, searching, accessing and reusing data based on processes and technologies, most of which are developed using open-source software, in alignment with the proposed Open Science model.

References

Acharjya, D. P.; Kauser, A. P. A survey on big data analytics: challenges, open research issues and tools. *International Journal of Advanced Computer Science and Applications*, v. 7, n. 2, p. 511-518, 2016. Available from: https://thesai.org/Downloads/Volume7No2/Paper_67-A_Survey_on_Big_Data_Analytics_Challenges.pdf. Cited: Oct. 1st 2023.

Bardin, L. *Análise de conteúdo*. Lisboa: Edições 70, 2020.

Bertin, P. R. B. *et al.* A parceria para Governo Aberto como plataforma para o avanço da Ciência Aberta no Brasil. *Transinformação*, v. 31, e190020, 2019. Doi: http://dx.doi.org/10.1590/2318-0889201931e190020.

Bhattacharya, S. *et al*. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific Data*, v. 5, e180015, 2018. Doi: https://.doi.org/10.1038/sdata.2018.15.

Bishop, L.; Kuula-Luumi, A. Revisiting qualitative data reuse: a decade on. *SAGE Open*, p. 1-15, 2017. Doi: https://.doi.org/10.1177/2158244016685136.

Budapest Open Access Initiative. *Budapest Open Access Initiative*. Budapeste: BOAI, 2002. Available from: http://www.opensocietyfoundations.org/openaccess/read. Cited: Oct. 1st 2023.

Carlson, S. Lost in a sea of science data. *Chronicle of Higher Education*, v. 52, n. 42, p. A35, 2006.

Chan, L. *et al. Open Science Beyond Open Access*: for and with communities. A step towards the decolonization of knowledge. Ottawa: The Canadian Commission for UNESCO's IdeaLab., 2020.

European Commission. *Open innovation, open science, open to the world: a vision for Europe*. Publications Office. 2015. Doi: https://doi.org/doi/10.2777/061652.

Fecher, B.; Friesike, S. Open Science: One term, five schools of thought. *In*: Bartling, S.; Friesike, S. (ed.). *Opening Science*. New York: Springer, 2014. Doi: https://doi.org/10.1007/978-3-319-00026-8_2.

Gad, A. G. *et al.* An improved binary sparrow search algorithm for feature selection in data classification. *Neural Computing and Applications*, v. 34, p. 15705-15752, 2022. Doi: https://doi.org/10.1007/s00521-022-07203-7.

Greenberg, J. *et al.* A metadata best practice for a scientific data repository. *Journal of Library Metadata*, v. 9, n. 3-4, p. 194-212, 2009. Doi: https://doi.org/10.1080/19386380903405090.

Guo, X. *et al.* CNSA: a data repository for archiving omics data. *Database*, v. 2020, p. 1-6, 2020. Doi: https://doi.org/10.1093/database/baaa055.

Kearnes, S. M. *et al.* The open reaction database. *Journal of the American Chemical Society*, v. 143, p. 18820-18826, 2021. Doi: https://doi.org/10.1021/jacs.1c09820.

Kindling, M.; Strecker, D. List of Data Journals. *Zenodo*, Version 1.0., 2022. Doi: https://doi.org/10.5281/zenodo.7082125.

Lancaster, F. W. *Information Retrieval Systems*: Characteristics, Testing and Evaluation. 2. ed. Los Angeles: John Wiley & Sons, 1979. (Information Sciences Series).

Michener, W. K. Ecological data sharing. *Ecological Informatics*, v. 29, n. 1, p. 33-44, 2015. Doi: https://doi.org/10.1016/j.ecoinf.2015.06.010.

Moriya, Y. *et al.* The jPOST environment: an integrated proteomics data repository and database. *Nucleic Acids Research*, v. 47, n. 1, p. D1218-D1224, 2019. Doi: https://doi.org/10.1093/nar/gky899.

Nargesian, F. *et al.* Table union search on open data. *Proceedings of the VLDB Endowment*, v. 11, n. 7, p. 813-825, 2018. Doi: https://doi.org/10.14778/3192965.3192973.

Nosek, B. A. *et al.* Promoting an open research culture. *Science*, v. 348, n. 6242, p. 1422. 2015. Doi: https://doi.org/10.1126/science.aab2374.

Okuda, S. *et al*. jPOSTrepo: an international standard data repository for proteomes. *Nucleic Acids Research*, v. 45, n. D1, p. D1107-D1111, 2017. Doi: https://doi.org/10.1093/nar/gkw1080.

Pagani, R. N.; Kovaleski, J. L.; Resende, L. M. Methodi Ordinatio: a proposed methodology to select and rank relevant scientific papers encompassing the impact factor, number of citation, and year of publication. *Scientometrics*, v. 105, p. 2109-2135, 2015. Doi: https://doi.org/10.1007/s11192-015-1744-x.

Pampel, H. *et al.* Re3data: Indexing the global research data repository landscape since 2012. *Scientific Data*, v. 10, n. 1, p. 571. 2023. Doi: https://doi.org/10.1038/s41597-023-02462-y.

Piwowar, H. A.; Vision, T. J. Data reuse and the open data citation advantage. *PeerJ*, v. 1, e175, 2013. Doi: https://doi.org/10.7717/peerj.175.

Pontika, N. *et al.* Fostering Open Science to research using a taxonomy and an eLearning Portal. *In*: iKnow: 15th International Conference on Knowledge Technologies and Data Driven Business, 21-22., 2015, Graz, Austria. *Proceedings* [...]. New York: Association for Computing Machinery, 2015. p. 1-8. Doi: http://dx.doi.org/doi:10.1145/2809563.2809571.

Sansone, S. A. *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, v. 37, n. 4, p. 350-369, 2019. Doi: https://doi.org/10.1038.s41587-019-0080-8.

Sayão, L. F.; Sales, L. F. Plataformas de gestão de dados de pesquisa: expandindo o conceito de repositórios de dados. *Palabra Clave*, v. 12, n. 1, e171, 2023. Doi: https://doi.org/10.24215/18539912e171.

Sena, P. M. B.; Segundo, W. L. R. C.; Melo, B. A. Ciência Aberta como parceria para Governo Aberto: compromisso por um novo modelo de avaliação. *Informação & Informação*, v. 27, n. 3, p. 14-33, 2022. Doi: https://doi.org/10.5433/1981-8920.2022v27n3p14.

Shi, G. *et al.* DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides. *Nucleic Acids Research*, v. 50, n. D1, p. D488-D496, 2022. Doi: https://doi.org/10.1093/nar/gkab651.

Silva, F. C. C.; Silveira, L. O ecossistema da Ciência Aberta. *Transinformação*, v. 31, e190001, 2019. http://dx.doi.org/10.1590/2318-0889201931e190001

Silveira, L. *et al.* Taxonomia da Ciência Aberta: revisada e ampliada. *Encontros Bibli*, v. 28, e91712, 2023. Doi: https://doi.org/10.5007/1518-2924.2023.e91712.

Smedley, D. *et al.* The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, v. 43, n. W1, p. W589-W598, 2015. Doi: https://doi.org/10.1093/nar/gkv350.

Strasser, C.; Abrams, S.; Cruse, P. DMPTool 2: expanding functionality for better data management planning. *International Journal of Digital Curation*, v. 9, n. 1, 2014. Doi: https://doi.org//10.2218/ijdc.v9i1.319.

Strasser, C. *et al.* Promoting data stewardship through best practices. *In*: Jones, M. B.; Gries, C. (ed.). *Proceedings of the environmental information management conference*. Santa Barbara: University of California, 2011. p. 126-131.

Strasser, C.; Cruse, P. The DMPTool and DataUp: helping researchers manage, archive, and share their data. *Research Data Management Implementations Workshop*, v. 13-14, 2013.

Sud, M. *et al.* Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Research*, v. 44, n. D1, p. D463-D470, 2016. Doi: https://doi.org/10.1093/nar/gkv1042.

Taylor, J. R. *et al.* The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage*, v. 144, p. 262-269, 2017. Doi: https://doi.org/10.1016/j.neuroimage.2015.09.018.

Tenopir, C. *et al.* Data sharing by scientists: practices and perceptions. *Plos One*, v. 6, e21101, 2011. Doi: http://dx.doi.org/10.1371/journal.pone.0021101.

Ularu, E. G. *et al.* Perspectives on big data and big data analytics. *Database Systems Journal*, v. 3, n. 4, p. 3-14, 2012.

United Nations Educational, Scientific and Cultural Organization. *An introduction to the UNESCO recommendation on Open Science*. Paris: UNESCO, 2022. Available in: https://www.unesco.org/en/open-science. Cited: October 1st 2023.

United Nations Educational, Scientific and Cultural Organization. *Towards a global consensus on open science*: report on UNESCO´s global online consultation on open science. Paris: UNESCO, 2020. Available in: https://www.unesco.org/en/open-science. Cited: October 1st 2023.

United Nations Educational, Scientific and Cultural Organization. *Recommendation on Open Science*. Paris: 2021. Available in: https://www.unesco.org/en/open-science. Cited: October 1st 2023.

Vicente-Saez, R.; Martinez-Fuentes, C. Open Science now: a systematic literature review for an integrated definition. *Journal of Business Research*, v. 88, p. 428, 2018. Doi: https://doi.org/10.1016/j.jbusres.2017.12.043

Walters, W. H. Data journals: incentivizing data access and documentation within the scholarly communication system. *Insights*, v. 33, p. 18, 2020. Doi: https://doi.org/10.1629/uksg.510.

Wilkinson, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, v. 15, n. 3, p. 160018, 2016. Doi: https://doi.org/10.1038/sdata.2016.18.

Zheng, S. *et al.* DrugComb update: a more comprehensive drug sensitivity data repository and analysis portal. *Nucleic Acids Research*, v. 49, n. 1, p. 174-184, 2021. Doi: https://doi.org/10.1093/nar/gkab438.

Contributors

Conceptualization: S. A. RODRIGUES-JUNIOR and M. V. TEXEIRA. Data curation: S. A. RODRIGUES-JUNIOR. Formal Analysis: S. A. RODRIGUES-JUNIOR. Investigation: S. A. RODRIGUES-JUNIOR. Methodology: S. A. RODRIGUES-JUNIOR. Supervision: M. V. TEXEIRA. Validation: M. V. TEXEIRA. Writing – original draft: S. A. RODRIGUES-JUNIOR. Writing – review and editing: M. V. TEXEIRA.