Biblios
ISSN 1562-4730 (online)

Defining geosciences research data through metadata reuse: a case study of PANGEA data repository

Alexandre Ribas Semeler
Universidade Federal do Rio Grande do Sul – UFRGS, Porto Alegre, Brasil

Luana Farias Sales
Instituto Brasileiro de Informação em Ciência e Tecnologia – Ibict, Rio de Janeiro, Brasil.

Adilson Luiz Pinto
Universidade Federal de Santa Catarina – UFSC, Florianópolis, Brasil

Roberta Pereira da Silva de Paula
Instituto Brasileiro de Informação em Ciência e Tecnologia – Ibict, Rio de Janeiro, Brasil.

Valquer Cleyton Paes Gandra
Instituto Brasileiro de Informação em Ciência e Tecnologia – Ibict, Rio de Janeiro, Brasil.

Heloisa Costa
Universidade Federal de Santa Catarina – UFSC, Florianópolis, Brasil

ORIGINAL

Abstract

Objective. Research data refers to factual records used as primary scientific research resources. Reusing research data metadata provides a new perspective, allowing the presentation of new tests, hypotheses, and new research developments. This study aims to identify the nature of the types of Geosciences research data based on the reuse of metadata from the PANGEA Data Publisher for Earth and Environmental Science available at (https://www.pangaea.de/). The research question to be analyzed is “Can the processes of analyzing and manipulating PANGEA research data metadata be used to define a concept of Geosciences research data?” To address this question, we considered data specification attributes used by data journals to describe the nature of research data: domain of specialization, accessibility, language, data type, acquisition, source location, specific subject area, and related publications.

Method. The methodology in question involved collecting, analyzing, and visualizing PANGEA research data metadata. In total, (426,272) records were downloaded from the data repository and compared to the data specifications used by data journals to describe the nature of research data in data papers. The methodology required the application of techniques and technologies used for descriptive analysis, information retrieval, data manipulation, and visualization of Dublin Core metadata. These techniques were implemented using the Python programming language and other data manipulation software, including OpenRefine and VOSviewer.

Results. The results of our analysis suggest a detailed examination of the metadata for (137,218) research data records from (6) six Geosciences collections. The number of records in the Geochemistry collection is (73,992), in the Atmospheric Sciences collection it is (32,314), in the Paleontology collection it is (25,903), in the Oceanography collection it is (22,287), in the Geophysics collection it is (4,175), and in the Hydrology collection, it is (834). PANGEA's (6) six research data metadata collections allow for the discussion of a concept of Geosciences research data as a type of data on studies related to the Earth, atmosphere, and oceans, across different geo-disciplines. The data come from a range of disciplines, including geochemistry, atmospheric science, paleontology, oceanography, geophysics, and hydrology, using technologies such as satellites, electronics microscopes, climate sensors, ships, computer modeling, and others. In addition, the data are augmented by other sources related to the study of the Earth and its processes.

Conclusions. In conclusion, research data metadata are domain-specific objects that serve as valuable research resources, regardless of their usage timing, purpose, data characteristics, or user. Geosciences research data combine laboratory and fieldwork techniques, utilizing technologies like satellites and climate sensors to study Earth’s processes. PANGEA metadata defines Geosciences research data as including observations, experiments, and modeling. Geosciences research data support replication, reinterpretation, and new research across disciplines, showcasing various facets of data reuse in scientific research.

DOI 10.5195/biblios.2024.1233

Keywords: research data, research data reuse, metadata, research data repository,Data Web Scraping,Geosciences

Definiendo los datos de investigación en geociencias a través de la reutilización demetadatos: un estudio de caso del repositorio de datos PANGEA

Resumen

Objetivo. Los datos de investigación se refieren a registros factuales utilizados como recursos primarios de investigación científica. La reutilización de metadatos de datos de investigación proporciona una nueva perspectiva, permitiendo la presentación de nuevas pruebas, hipótesis y nuevos desarrollos de investigación. Este estudio pretende identificar la naturaleza de los tipos de datos de investigación en Geociencias a partir de la reutilización de metadatos del PANGEA Data Publisher for Earth and Environmental Science disponible en (https://www.pangaea.de/). La pregunta de investigación a analizar es ¿Pueden los procesos de análisis y manipulación de metadatos de datos de investigación PANGEA utilizarse para definir un concepto de datos de investigación en Geociencias? Para abordar esta pregunta, se consideraron los atributos de especificación de datos utilizados por las revistas de datos para describir la naturaleza de los datos de investigación: dominio de especialización, accesibilidad, idioma, tipo de datos, adquisición, ubicación de la fuente, área temática específica y publicaciones relacionadas.
Método. La metodología involucró la recolección, análisis y visualización de metadatos de datos de investigación de PANGEA. En total, se descargaron (426,272) registros del repositorio de datos y se compararon con las especificaciones de datos utilizadas por las revistas para describir los datos en los artículos.La metodología en cuestión consistió en recopilar, analizar y visualizar los metadatos de datos de investigación de PANGEA. En total, se descargaron (426.272) registros del repositorio de datos y se compararon con las especificaciones de datos utilizadas por las revistas de datos para describir la naturaleza de los datos de investigación en los documentos de datos. La metodología requirió la aplicación de técnicas y tecnologías utilizadas para el análisis descriptivo, la recuperación de información, la manipulación de datos y la visualización de metadatos Dublin Core. Estas técnicas se implementaron utilizando el lenguaje de programación Python y otros software de manipulación de datos, incluyendo OpenRefine y VOSviewer.
Resultados. Los resultados de nuestro análisis sugieren un examen detallado de los metadatos de (137.218) registros de datos de investigación de (6) seis colecciones de Geociencias. El número de registros en la colección de Geoquímica es de (73.992), en la colección de Ciencias Atmosféricas es de (32.314), en la colección de Paleontología es de (25.903), en la colección de Oceanografía es de (22.287), en la colección de Geofísica es de (4.175), y en la colección de Hidrología, es de (834). Las (6) seis colecciones de metadatos de datos de investigación de PANGEA permiten discutir un concepto de datos de investigación en Geociencias como un tipo de datos sobre estudios relacionados con la Tierra, la atmósfera y los océanos, a través de diferentes geo-disciplinas. Los datos proceden de una serie de disciplinas, como la geoquímica, la ciencia atmosférica, la paleontología, la oceanografía, la geofísica y la hidrología, y utilizan tecnologías como los satélites, los microscopios electrónicos, sensores climáticos, barcos de investigaciones, modelos informáticos y otros. Además, los datos se complementan con otras fuentes relacionadas con el estudio de la Tierra y sus procesos.
Conclusiones. En conclusión, los metadatos de datos de investigación son objetos específicos de un dominio que sirven como valiosos recursos de investigación, independientemente de su momento de uso, finalidad, características de los datos o usuario. Los datos de investigación en geociencias combinan técnicas de laboratorio y de campo, utilizando tecnologías como los satélites y sensores climáticos para estudiar los procesos de la Tierra. Los metadatos PANGEA definen los datos de investigación en geociencias como observaciones, experimentos y modelización. Los datos de investigación en geociencias apoyan la réplica, la reinterpretación y la nueva investigación entre disciplinas, mostrando varias facetas de la reutilización de datos en la investigación científica.

Palabras clave: datos de investigación, reutilización de datos de investigación, metadatos, repositorio de datos de investigación, Raspado Web de Datos, Geociencias

Definindo dados de pesquisa em geociências por meio da reutilização de metadados: um estudo de casosobre o repositório de dados PANGEA

Resumo

Objetivo. Dados de pesquisa referem-se a registros factuais usados como recursos primários para pesquisas científicas. A reutilização de metadados de dados de pesquisa fornece uma nova perspectiva, permitindo a apresentação de novas evidências, hipóteses e novos desenvolvimentos de pesquisa.Este estudo pretende identificar a natureza dos tipos de dados de pesquisa em Geociências a partir da reutilização de metadados do PANGEA Data Publisher for Earth and Environmental Science, disponível em (https://www.pangaea.de/). A questão de pesquisa busca compreender se os processos de análise e manipulação de metadados de dados de pesquisa do PANGEA podem ser usados para definir um conceito de dados de pesquisa em Geociências? Para responder a essa pergunta, consideramos os atributos de

especificação de dados usados pelos periódicos de dados para descrever a natureza dos dados de pesquisa, como: domínio de especialização, acessibilidade, idioma, tipo de dados, aquisição, local de origem, área de assunto específica e publicações relacionadas.
Método. No total, (426.272) registros foram baixados do PANGEA e comparados com as especificações de dados usadas pelos periódicos de dados para descrever a natureza dos dados de pesquisa nos artigos de dados. A metodologia exigiu a aplicação de técnicas e tecnologias usadas para análise descritiva, recuperação de informações, manipulação de dados e visualização dos metadados do Dublin Core. Essas técnicas foram implementadas usando a linguagem de programação Python e outros softwares de manipulação de dados, incluindo o OpenRefine e o VOSviewer.
Resultados. Os resultados de nossa análise sugerem um exame detalhado de 137.218 registros de metadados de dados de pesquisa em 6 coleções de dados sobre Geociências. O número de registros na coleção de Geoquímica é de 73.992, na coleção Ciências Atmosféricas é de 32.314, na coleção Palaeontologia é de 25.903, na coleção Oceanografia é de 22.287, na coleção Geofisica é de 4.175 e na coleção Hidrologia é de 834. As 6 coleções de metadados de dados de pesquisa do PANGEA permitem discutir um conceito de dados de pesquisa em Geociências como um tipo de dados sobre estudos relacionados à Terra, à atmosfera e aos oceanos, em diferentes geodisciplinas. Os dados vêm de diversas disciplinas, como geoquímica, ciência atmosférica, paleontologia, oceanografia, geofísica e hidrologia, e são o resultado do uso de diversas tecnologias como como satélites, microscópios eletrônicos, sensores climáticos, navios de pesquisa, modelos de computador entre outros. Além disso, os dados são complementados por outras fontes relacionadas aos estudos da Terra e de seus processos.
Conclusões. Concluindo, os metadados de dados de pesquisa são objetos específicos de domínios e servem como valiosos recursos de pesquisa, independentemente do tempo de uso, da finalidade, das características dos dados ou dos usuários. Os dados de pesquisa em geociências combinam técnicas de laboratório e de campo, usando tecnologias como os satélites e sensores climáticos para estudar os processos da Terra. Os metadados do PANGEA definem os dados de pesquisa em geociências como observações, experimentos e modelagem. Os dados de pesquisa em geociências apoiam a replicação, a reinterpretação e novas pesquisas entre disciplinas, mostrando várias facetas da reutilização de dados na pesquisa científica.

Palavras-chave: dados de pesquisa, reutilização de dados de pesquisa, metadados, repositório de dados de pesquisa, extração de dados da Web, Geociências

1 Introduction

The term research data was officially defined at the international level in 2007 by the Organization for Economic Cooperation and Development (OECD) through its declaration on the accessibility of scientific data with funding. Other researchers and institutions around the world dealing with the topic consider research data to be factual objects that are used as primary resources in scientific investigation. In this paper, research data is defined as records, files, or other evidence, regardless of their content or form (e.g., printed or digital material), that comprise research observations, findings, or results, including primary materials and data that have already been analyzed. Any information concerning research (Rice & Southall, 2016; Shutsko & Stock, 2023).

This paper explores the nature of research data presented in data papers, focusing on the data specification attributes utilized by data journals. Data papers represent a growing academic genre dedicated to describing research data objects. Unlike traditional academic papers, which address methodological questions, data papers aim to provide precise descriptions of data collections, emphasizing their reuse potential. These papers serve as scientific information sources specifically designed for publishing datasets in their raw form, facilitating their application in future research (Li & Jiao, 2022; Jiao et al., 2024; Jiao & Darch, 2020; Walters, 2020; Felden et al., 2023).

Research data metadata are descriptions of that data created for inclusion as records in databases, following the principle of being data about research data. Research data metadata is usually available in data repositories or data papers. According to Uzwyshyn (2016) and Kim (2020) research data repositories, often referred to as data archives or libraries, are collections of research data in various formats that can be reused, and represent an advancement over document repositories.

In studies about research data metadata, another key point to highlight is the diversity of domains in which the data is generated. In the case of this paper, we will investigate the generation of research data metadata in the Geosciences domain. We chose the Geosciences area because it has excellent potential for generating research data.

We will conduct our research using Dublin Core metadata records, collected via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) interoperability protocol, from PANGEA Data Publisher for Earth & Environmental Science collections. PANGEA is a highly specialized source of research data in the Geosciences. A relevant platform for the reuse of research data metadata is the PANGEA, a research data repository that functions as an open access library for archiving, publishing, and distributing georeferenced earth and environmental science data. The PANGEA Data Publisher it was developed by the Alfred Wegener Institute at the Helmholtz Center for Polar and Marine Research (AWI) and the Center for Marine Environmental Sciences (MARUM) at the University of Bremen in Germany. PANGEA emerged as a database in 1987, designed to structure earth sciences and environmental science data (Felden et al., 2023).

Geosciences are a synonym for Earth sciences. Earth sciences are usually used in a more general sense; for example, when they include climatology and hydrology. In this paper, we prefer the term Geosciences, which is an interdisciplinary field that focuses on the study of the Earth's geological, chemical, climatic, paleontological, physical, oceanographic, and water processes. Geosciences seek to understand the Earth's natural processes and phenomena that, in general, have shaped and continue to shape the planet, from its formation to current events, it examines the history of the Earth, its composition, structure, and current and historical processes (Lyell, 1853; Tarbuck, Lutgens & Tasa, 2015; Keller & DeVecchio, 2019).

The research method is based on the reuse of metadata from research data published in PANGEA. The development of the method required the application of knowledge inherent in the techniques and technologies used for descriptive data analysis, information retrieval, data manipulation, metadata analysis, and visualization using the Python programming language.

The aim is to carry out an exploratory and descriptive investigation of PANGEA's research data metadata collections based on the specifications adopted for describing data in data papers. The research question to be analyzed is “Can the processes of analyzing and manipulating PANGEA research data metadata be used to define a concept of Geosciences research data?” To address this question, we considered data specification attributes used by data journals to describe the nature of research data: domain of specialization, accessibility, language, data type, acquisition, source location, specific subject area, and related publications. With this in mind, the aim is to map the multidisciplinarity of research data in the Geosciences domain, based on specific fields related to the geoscientific disciplines, such as geochemistry, atmospheric sciences, paleontology, oceanography, geophysics, and hydrology. These disciplines are massive producers of quantities of research data in various formats and typologies, both analog and digital.

Therefore, the following text presents the challenges of reusing research data, and the methodology of the research problem, and reflects on the relevance and addresses six (6) types of Geosciences research data based on metadata collections from the PANGEA data repository.

2 Background: Reuse of Research Data

The reuse of research data is a growing concern in open science. In several disciplines, data reuse has become a common goal in recognition of the potential for new combinations and analyses of data to address new research questions. According to Ferder et al. (2015), reuse is using a collection of data to solve a new problem. Research data should be reused to enhance research reproducibility and enable scientific discovery (Van de Sandt et al., 2019). Another context pertains to research data reuse, which involves combining diverse datasets with a distinct research question by utilizing a novel approach to identical data.

Thus, Tenopir et al. (2015) define reuse as the application of any research resource, irrespective of its time of use, its purpose, or the characteristics of the information and its user. An essential practice of open science is the reuse of research data. Each type of research data reuse has its benefits and challenges. Disciplinary reuse facilitates replicability and validation of results, while interdisciplinary reuse can generate significant innovation by applying perspectives and methods from different fields. Reuse by the researchers themselves provides an efficient way to maximize the value of the data collected, allowing for a more in-depth exploration of a dataset. Promoting a culture of research data sharing and providing adequate infrastructure for data storage and accessibility are crucial steps to support these data reuse practices (Borgman, 2012; Pampel et al. 2013).

Therefore, the following text presents the methodology of the research problem.

3 Materials and Methods

The proposed method of our paper is to interpret the PANGEA metadata according to some data specifications applied to describe data papers. As a strategy for identifying the nature and typology of Geosciences research data, we adopted the following methodology: (a) collect all metadata from the PANGEA data repository; (b) separate collections from the Geosciences domain; (c) describe the nature of research data under the Dublin Core metadata standard and the data specification attributes used by data journals.

PANGEA offers a wide range of web services (SOAP/REST). The OAI-PMH API allows retrieving any set of numeric and textual data. All PANGEA datasets are also provided in Dublin Core compliant metadata format (https://schema.org/Dataset). It provides collection API and Dublin Core metadata interoperability services via the OAI-PMH protocol (https://ws.pangaea.de/oai/provider/). The PANGEA data repository provides targeted support for research data management as well as long-term data archiving and publication.

The data files used to analyze these studies have been deposited in the Open Science Framework. They can be viewed in Table 1 and can be downloaded and viewed at the following link (https://doi.org/10.17605/OSF.IO/3BSX2).

Table1

DataSets and codes

File Name	Link	Data category	Description
Supplementary File S1	https://osf.io/baz4e	Script python	Webscraping metadata pangea
Supplementary File S2	https://osf.io/2d9bs	Script python	Convert xml to csv
Supplementary File S3	https://osf.io/uywm9	Script python	Convert tsv to ris
Supplementary File S4	https://osf.io/6qg4s	General data base	All data pangea in xml format
Supplementary File S5	https://osf.io/5sw9x	General data base	Main coordinates research data
Supplementary File S6	https://osf.io/4cdkf	Geochemical research data	All data in topic chemistry
Supplementary File S7	https://osf.io/3czer	Geochemical research data	20 largest occurrences
Supplementary File S8	https://osf.io/xtkeu	Geochemical research data	Frequency dc subject
Supplementary File S9	https://osf.io/sprvq	Atmospheric research data	All data in topic atmosphere
Supplementary File S10	https://osf.io/gryqk	Atmospheric research data	20 largest occurrences
Supplementary File S11	https://osf.io/5yve3	Atmospheric research data	Frequency dc subject
Supplementary File S12	https://osf.io/b824t	Paleontological research	All data in topic paleontology
Supplementary File S13	https://osf.io/qp7rw	Paleontological research data	20 largest occurrences
Supplementary File S14	https://osf.io/x5qga	Paleontological research data	Frequency dc subject
Supplementary File S15	https://osf.io/5h67b	Oceanographic research data	All data in topic ocean
Supplementary File S16	https://osf.io/89m2c	Oceanographic research data	20 largest occurrences
Supplementary File S17	https://osf.io/9zqb6	Oceanographic research data	Frequency dc subject
Supplementary File S18	https://osf.io/ewjrt	Geophysical research data	All data in topic geophysical
Supplementary File S19	https://osf.io/jr54c	Geophysical research data	20 largest occurrences
Supplementary File S20	https://osf.io/kd6m9	Geophysical research data	Frequency dc subject
Supplementary File S21	https://osf.io/27u3r	Hydrological research data	All data in topic geophysical
Supplementary File S22	https://osf.io/xbe5y	Hydrological research data	20 largest occurrences
Supplementary File S23	https://osf.io/u7gpz	Hydrological research data	Frequency dc subject

Note. Source: Elaborated by the Authors (2024).

For data collection, Python scripts developed by Phillips (2013) were employed to download an XML file containing all Dublin Core metadata via the OAI-PMH interoperability protocol. The scripts are accessible via the following link: (https://osf.io/baz4e) (see Supplementary File S1). Phillips' project at the University of North Texas

Libraries Digital Projects Unit developed a script to collect Dublin Core records via the OAI-PMH protocol and transform them into an XML file that can be analyzed with Unix/Linux-based command-line tools.

The (2) two Python scripts in Table 1, developed by Semeler (2024) were designed to convert and separate the collections from the downloaded XML file. One to convert XML to CSV (see Supplementary File S2) available at (https://osf.io/2d9bs) and the other to convert TSV to RIS (see Supplementary File S3) available at (https://osf.io/uywm9). The conversions are performed to facilitate the data handling by transforming it into smaller collections, counting the frequencies, and creating keyword visualization networks using the VOSviewer software. Finally, the general frequencies of the attributes of the Dublin Core tags were generated using the OpenRefine software, by analyzing the textual facets in the converted files.

For the description of the research data attributes, we followed the model for preparing data papers from Elsevier's journal Data in Brief (https://www.sciencedirect.com/journal/data-in-brief). These specifications are used in the table specification to describe the nature of data types in data papers. Our method involves comparing the description of Dublin Core metadata records published in PANGEA with the data specifications used to describe research data in data papers.

To accomplish this, we used the OAI-PMH protocol to extract metadata and describe the datasets according to several specifications: field of knowledge, data accessibility, language, type of data, how the data were acquired and their current state, location of the data, specific subject area, and related publications. Based on the potential for data reuse, we adopted a research data specification table that highlights the typology of research data. We then compared the Dublin Core attributes of PANGEA with the specification attributes of a data paper, following guidelines from Elsevier's journal Data in Brief.

Table 2

Presents the specifications for the data paper, as outlined in the journal Data in Brief table. These specifications are under the Dublin Core metadata.

Data specification attribute	Dublin Core Metadata Field at PANGEA	Key in text
Field of knowledge	header – setSpec	A1
Data accessibility	header – identifier	A2
Language	dc: language	A3
Type of data	dc: format	A4
How the data were acquired and the state of the data	dc: description; dc:title; dc: relation	A5
Location of the data	dc: coverage	A6
Specific subject area	dc: subject	A7
Related publication source	dc: source	A8
Number of records in PANGEA dataset	header – setSpec	A9

*Note.*Source: Elaborated by the Authors (2024).

We have selected several research datasets from PANGEA to illustrate the types of data available in the repositories, focusing on those related to Geosciences research data metadata. These datasets were chosen for their exemplary representation of the data available. What follows is an analysis of the data collected, presented as our research findings.

4 Results

The findings of this research were based on a complete copy of PANGEA with approximately (426,272) records downloaded in April 2024. The download process took approximately 6 hours and resulted in an XML file size of 1.08 GB (see Supplementary File S4) available at (https://osf.io/6qg4s). This enabled the mapping of six (6) collections of research data directly related to the field of Geosciences, as shown in Table 3.

Table 3

PANGEA on-line header SetSpec collections (A1) and number of records in DataSets (A9).

Header - SetSpec (A1)	Link to online SetSpec header at PANGEA	DataSets (A9)
topicChemistry	https://ws.pangaea.de/oai/provider?verb=ListRecords&metadataPrefix=oai_dc&set=topicChemistry	73992
topicAtmosphere	https://ws.pangaea.de/oai/provider?verb=ListRecords&metadataPrefix=oai_dc&set=topicAtmosphere	32314
topicPaleontology	https://ws.pangaea.de/oai/provider?verb=ListRecords&metadataPrefix=oai_dc&set=topicPaleontology	25903
topicOceans	https://ws.pangaea.de/oai/provider?verb=ListRecords&metadataPrefix=oai_dc&set=topicOceans	22287
topicGeophysics	https://ws.pangaea.de/oai/provider?verb=ListRecords&metadataPrefix=oai_dc&set=topicGeophysics	4175
topicLakesRivers	https://ws.pangaea.de/oai/provider?verb=ListRecords&metadataPrefix=oai_dc&set=topicLakesRivers	834

Note. Source: Elaborated by the Authors (2024).

The geographical distribution (see Supplementary File S5) available at (https://osf.io/5sw9x) of the dataset analyzed can be seen according to the main coordinates (dc: Coverage (A6)). The dataset was most frequently collected at the following locations: atmospheric data (0°01'18.7"N 86°27'45.5"W), paleontological data (0°01'18.7"N 86°27'45.5"W), and geochemical data (22°45'00.0"N 86°27'45.5"W), which were found in the North Pacific Ocean. The most frequent oceanographic data were collected in the Bering Sea (65°00'00.0"N 170°00'00.0"W). The geophysical data originated from Aguilar de Campo. The data were collected in Valencia, Spain (latitude 42.816949, longitude -4.260681) and the hydrological data (46°52'22.5"N 10°42'52.0"E) were gathered in the Kaunertal region, Austria at Lake Weißsee.

To identify the nature of some types of research data in the Geosciences, we present the (6) six research data collections described in Table 3, based on (137,218) research data metadata records deposited in the English (dc: language (A3)) across key research domains available on PANGEA.

4.1 Geochemical Research Data

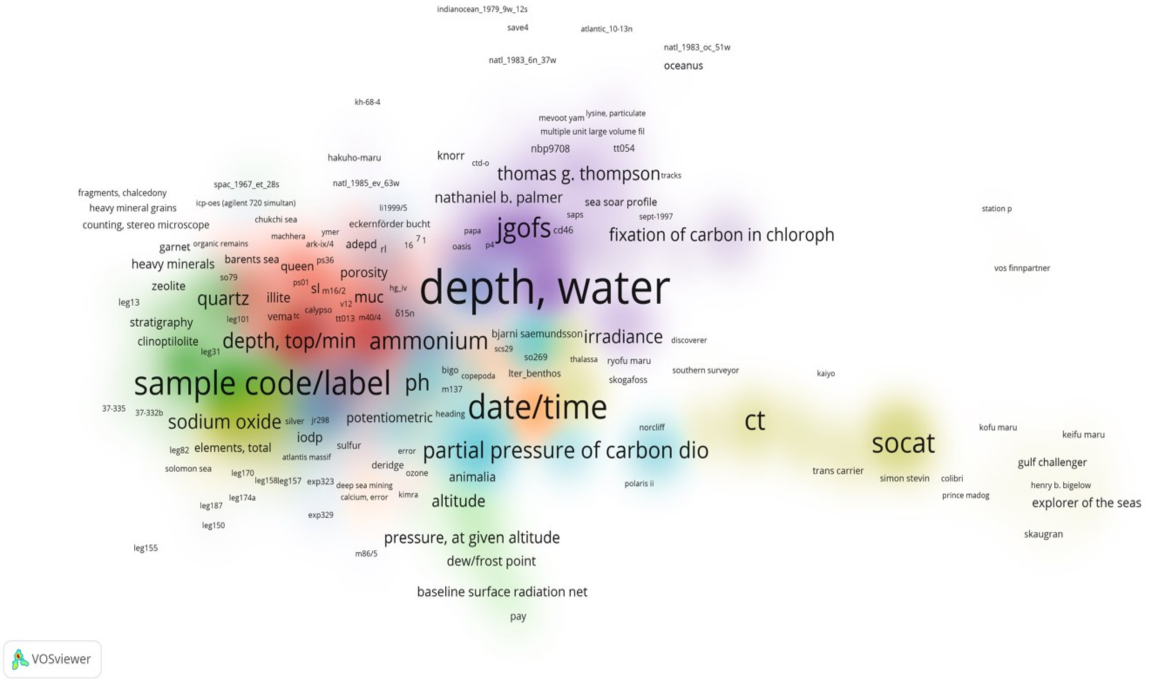
As defined by White, (2013) and Clarke (1924), geochemistry is the application of chemical tools to address geological issues. Geochemistry is a specialized field of chemistry that is utilized in the context of Earth studies, encompassing various aspects of chemical principles, such as thermodynamics, isotope chemistry, rock traces, and the study of soil in general. Geochemistry is a field of Geosciences that studies the chemical elements in terms of their composition, distribution, and transformation processes on Earth. It examines how these elements are found in the crust, mantle, atmosphere, and oceans, how they move and interact via chemical processes, and employs current technologies.

The data available in Table 1 (see Supplementary File S6) available at (https://osf.io/4cdkf), can be used to describe this collection. A total of (73,992) records are available in PANGEA, of which the attribute (header - setSpec (A1)) represents the geochemistry collection.

To analyze the metadata collection, the (20) largest occurrences, represent the core of the collection in PANGEA. The largest thematic occurrences (dc: Subject (A7)), (see Supplementary File S7) available at (https://osf.io/3czer), this file shows information about baseline surface radiation network, monitoring station, ny-ålesund, ny-ålesund, Spitsbergen, ozone, partial pressure, pressure, at given altitude, radiosonde, temperature, air, wind direction, wind speed, and carbon, organic. The total per volume, chlorophyll a, chlorophyll a standard deviation, and comment are also included. The depth of the water, the Hawaiian Islands, North Central Pacific, Hawaii Ocean Time-Series, Joint Global Ocean Flux Study, Moana Wave, and open ocean station are among the other themes included in Fig 1.

Figure 1

Geochemical research data topics (dc: Subject (A7)). This is a network map of the thematic density of Geochemistry research data.



Note. Source: Elaborated by the Authors (2024).

Based on (4,839) keywords with a minimum frequency (20) in the dataset, the core of the collection (see Supplementary File S8) is available at (https://osf.io/xtkeu). The data illustrates the types of data collected, the number of occurrences of each type, and the cluster to which they belong. The thematic focus of PANGEA's geochemistry collection is on studies of the sedimentology of rocks, silicates, nitrites, oxygen, and carbon gas chromatography; the timing and dating of gases and minerals; algae and water nutrients such as chlorophyll; temperature; and chemical element analysis from the depths of oceans such as the South Pacific and North Atlantic and the Arctic Ocean. The data are organized into (23) clusters, which suggest the existence of common patterns or characteristics. For instance, clusters 5 and 13 are particularly prevalent in depth and water temperature data, respectively, while cluster 2 is associated with sample codes and drilling. This reflects the thematic breadth of the collection in terms of ocean depth studies, chemical elements and nutrients, and sediment and rock dating and drilling processes.

A review of the data, (see Supplementary File S7) available at (https://osf.io/3czer) reveals that the majority of data types (dc: Format (A4)) are text/tab-separated-values and application/zip. The source data in question (dc: Source (A8)) is derived from a multitude of research projects conducted by various marines and Geoscientific research institutions, the most noteworthy: Institut français de recherche pour l'exploitation de la mer; Scripps Institution of Oceanography, UC San Diego: Second largest; Hellenic Center of Marine Research, Institut of Oceanography, Greece. Other notable institutions include the Alfred Wegener Institute in various locations, the Universidad Arturo Prat, and various meteorological and oceanographic organizations. These contributions reflect a broad international collaboration in marine geochemistry research.

According to Bienhold and Boetius (2015), an illustrative example of geochemical research data is https://doi.pangaea.de/10.1594/PANGAEA.849054, with (dc: title (A5)), Porosity in sediment cores from the central Arctic Ocean during POLARSTERN cruise ARK-XXVII/3 from August-September 2012. The (dc: description (A5)) represents a set of research data related to the publication Dissolved Organic Matter in Pore Water of Arctic Ocean Sediments: Environmental Influence on Molecular Composition. This paper was published in the journal Organic Geochemistry, utilizing the (DOI) of https://doi.org/10.1016/j.orggeochem.2016.04.003. This dataset reports on the RV Polarstern ARK-XXVII/3 cruise in the Arctic Ocean in the summer of 2012. During this period, sea ice decreased to a record low, and sediment cores were collected in the Nansen and Amundsen basins for porosity measurements.

The next set of data pertains to atmospheric measurements, a type of climate data that is of particular interest.

4.2 Atmospheric Research Data

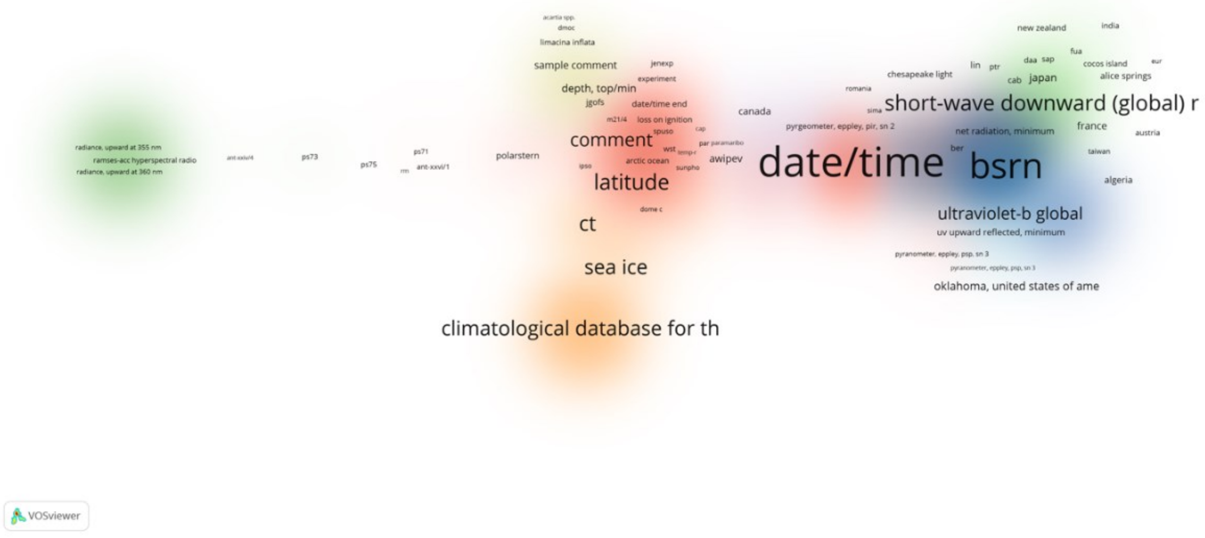
Climatology is the branch of Geosciences that studies the atmosphere, its variability, and long-term meteorological phenomena. According to Rohli and Viega (2008) and Köppen (1931), climatology is a holistic science that incorporates theories and data on all parts of the earth-ocean-atmosphere system, including human influence, to integrate the whole to explain atmospheric properties. The field studies the average atmospheric conditions in a given region over some time. Climatology also studies the causes and effects of climate change and searches for patterns and trends that may affect the environment and human life.

The data available in Table 1 can be used to describe this collection. Based on the (32,314) records available in PANGEA, attribute (header - setSpec (A1)), the collection of atmospheric research data, (see Supplementary File S9) is available at (https://osf.io/sprvq). An example of this type of data is the (20) the largest occurrences, (see Supplementary File S10) available at (https://osf.io/grygk), the core of the PANGEA collection.

The largest thematic occurrences (dc: Subject (A7)) represent studies on air temperature; barometer; dronningmaud land, Antarctica; humidity, relative; hygrometer; meteorological long-term observations; pyranometer; short-wave downward radiation; station pressure; sunshine duration; thermometer; total ultraviolet radiometer, Eppley; ultraviolet radiation; wind direction and speed; baseline surface radiation network; Estonia; monitoring station; ultraviolet-b global; uv-radiometer. See all topics in Fig. 2.

Figure 2

Atmospheric research data topics (dc: Subject (A7)). This is a network map of the thematic density of Atmospheric research data.



Note. Source: Elaborated by the Authors (2024).

A total of (4,227) keywords with a minimum frequency of (10) in the dataset were used to identify the thematic core of the collection. This collection belongs to the PANGEA atmospheric dataset and is organized into 53 clusters. The data (see Supplementary File S11) is available at (https://osf.io/5yve3). The data indicate that the thematic center of PANGEA's atmospheric data is primarily concerned with measurements and observations related to radiation (both shortwave and long wave), atmospheric pressure, temperature, and humidity. It can be observed that the PANGEA collection is designed to provide precise and comprehensive measurements of atmospheric components. This is achieved through the use of a combination of monitoring networks and various instruments, which ensures the collection of high-quality and reliable data.

The following data, (see Supplementary File S9) available at (https://osf.io/sprvq), indicates that the majority of data types (dc: Format (A4)) are text/tab-separated-values and application/zip. The sources of this data (dc: Source (A8)) are collections of research data produced by research institutions. The NOAA - Air Resources Laboratory in Boulder, the National Maritime Museum in Greenwich, United Kingdom, the Climate Monitoring & Diagnostics Laboratory in Boulder, the Swiss Meteorological Agency in Payerne, the National Archief of the Netherlands in Den Haag, the Netherlands, and the Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research, Bremerhaven, Germany, are among the institutions that have contributed to this research.

The variety of sources ranges from atmospheric and climate research laboratories to historical archives and national meteorological centers. This diversity is crucial for providing a wide range of atmospheric, climatic, and historical data, which allows for a comprehensive and detailed analysis. The data from these institutions is fundamental to the study of climate, atmospheric monitoring, and environmental research, contributing significantly to the global understanding of climate change and its implications. PANGEA's data collection is enhanced by the contributions of numerous research institutions worldwide, each providing specific and valuable data that collectively constitute a robust foundation for the study and monitoring of atmospheric and climatic conditions.

Under research data metadata from Jones et al. (2007), as an illustration, the dataset (dc_identifie (A2)) on atmospheric research accessible in https://doi.pangaea.de/10.1594/PANGAEA.611088, with (dc: title (A5)) Climatological observations from ship logbooks between 1750 and 1854. The (dc: description (A5)) represents a set of research data linked to the publication’s improved understanding of past climatic variability from early daily European instrumental sources. The article was published in the journal Climatic Change and is available online via DOI: https://doi.org/10.1023/A:1014902904197.This dataset concerns the Climatological Database for the World's Oceans: 1750-1854 (CLIWOC), the dataset comprises observations from ship logbooks between 1750 and 1854, which is research data on daily meteorological observations from the logbooks of British, Dutch, French, and Spanish ships engaged in imperial trade during the 18th and 19th centuries. This database contains information on ocean wind field patterns, which provide evidence of phenomena such as the El Niño-Southern Oscillation and the North Atlantic Oscillation.

The next set of data is paleontological, a relevant type of Geosciences research data.

4.3 Paleontological Research Data

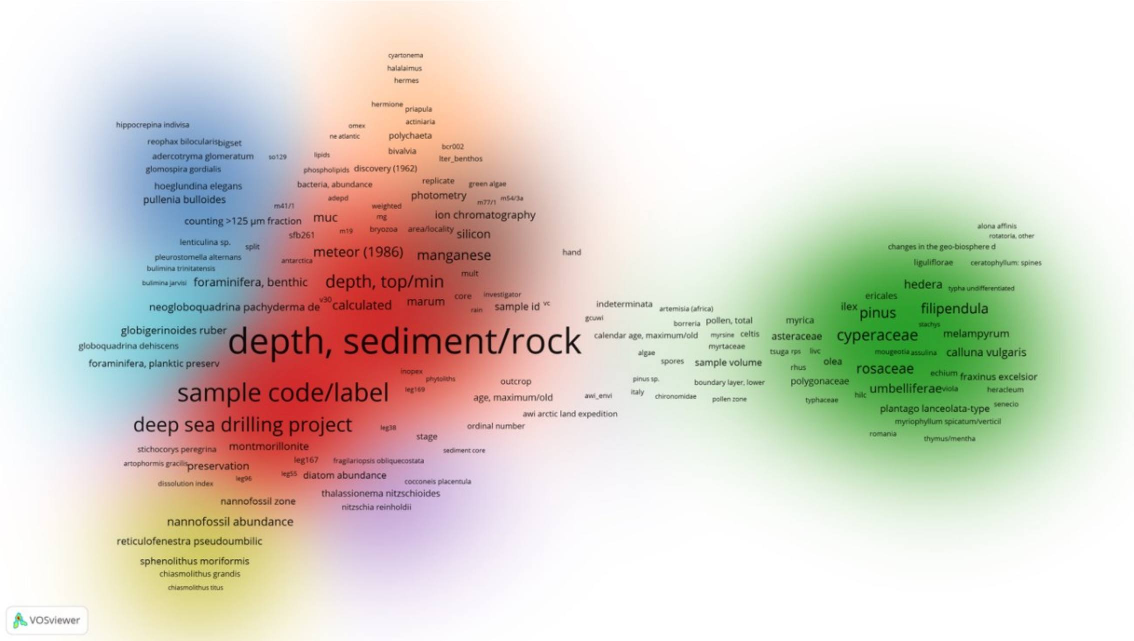
By definition Foote and Miller (2007); Gould, (2002) and Jones (2011), paleontology is the scientific study of fossils and encompasses questions on the history, current state, and future of life on Earth. These inquiries are approached from a multidisciplinary perspective, integrating insights from geology and biology. Fossils are the remains or traces of ancient organisms that have been preserved mechanically or chemically within rocks. This discipline endeavors to reconstruct the history of life on Earth by investigating and analyzing the remains of animals, plants, and other organisms that inhabited the planet in the distant past. A fundamental concept in paleontology is the study of fossils as a means of reconstructing the history of life on Earth.

A total of (25,903) records are available in PANGEA for the paleontology collection. Based on these records, the attribute (header setSpec (A1)) can be described according to (see Supplementary File S12) available at (https://osf.io/b824t).

The largest thematic occurrences in paleontology (dc: Subject (A7)), (see Supplementary File S13) available at (https://osf.io/qp7rw), represent data on marine and geological research institutes such as the Center for Marine Environmental Sciences; MARUM and the Ocean Drilling Program. Despite representing subject matter, these institutions are also research institutions. All the themes can be seen in Fig 3.

Figure 3

Paleontological research data topics (dc: Subject (A7)). This is a network map of the thematic density of Paleontological research data.



Note. Source: Elaborated by the Authors (2024).

A total of (4,656) keywords were identified with a minimum frequency of (30), (see Supplementary File S14) available at (https://osf.io/x5qga). The dataset is a set of data related to paleontology, with a focus on rock sedimentology and ocean research. The analysis reveals that the thematic core of the collection is in cluster 1, as indicated by the prevalence of terms related to drilling and sampling in ocean drilling projects such as DSDP (Deep Sea Drilling Project), ODP (Ocean Drilling Program), and IODP (Integrated Ocean Drilling Program). The specific mention of research vessels such as the JOIDES Resolution and the Glomar Challenger reinforces the focus on ocean drilling expeditions. Thus, the dataset involves research into sedimentology and stratigraphy; ocean drilling; rock age and lithology, fossils, and chemical composition. These aspects form a central part of the PANGEA collection, reflecting a multidisciplinary approach to understanding paleontological history through advanced sampling and analysis techniques.

The data (see Supplementary File S12) is available at (https://osf.io/b824t) and indicates that the majority of data types (dc: Format (A4)) are text/tab-separated-values, application/zip, and video/mpeg. The data sources, which are publications related to the research data metadata (dc: Source (A8)), are also data produced by research institutions.

The data sources and publications presented here demonstrate collaboration between various research institutions in the fields of paleontology and Geosciences. These include the European Pollen Database (EPD), the Alfred Wegener Institute, the Shirshov Institute of Oceanology, the IFM-GEOMAR Leibniz-Institute, the GEOMAR - Helmholtz Centre, the MARUM - Center for Marine Environmental Sciences, and the Department of Geosciences, Bremen University. The presence of numerous institutions from diverse countries, including Russia (Shirshov Institute of Oceanology, All-Russian Research Institute), the United States (United States Geological Survey), Belgium (University of Ghent), and the United Kingdom (National Oceanography Centre), exemplifies the global and collaborative nature of research in paleontology and Geosciences. The diversity of the sources ensures breadth and depth to the studies, encompassing a spectrum of methodologies, from detailed analysis of microfossils and sediments to major ocean drilling projects and paleoclimate reconstruction.

According to Gastaldello et al. (2024), an illustration, the paleontological research dataset (dc_identifie (A2)) can be accessed via the following digital (DOI): https://doi.org/10.1594/PANGAEA.962075. The dataset includes the following metadata: (dc: title (A5)) Age model, carbonate mass accumulation rates, and benthic foraminifera from ODP Site 175-1085, which represents a set of research data linked to the publication A benthic foraminifera perspective of the Late Miocene-Early Pliocene Biogenic Bloom at ODP Site 1085 (Southeast

Atlantic Ocean). The article was published in the journal Palaeogeography, Palaeoclimatology, Palaeoecology under the following (DOI): https://doi.org/10.1016/j.palaeo.2024.112040.This dataset is particularly well-suited to the observed increase in marine biological productivity worldwide at various ocean sites, which has been attributed to an increase in nutrient input or a significant reorganization of nutrients in the ocean. The work is related to the International Ocean Drilling Program (ODP) in the Atlantic Ocean, which investigated the ostracod data used to develop an age model for the benthic foraminifera. The data describe linear sedimentation rates, carbonate mass accumulation rates, benthic foraminiferal assemblage data, and associated indices such as benthic foraminiferal accumulation rate and diversity.

The subsequent dataset pertains to the oceans, a category of oceanographic data that is of particular relevance.

4.4 Oceanographic Research Data

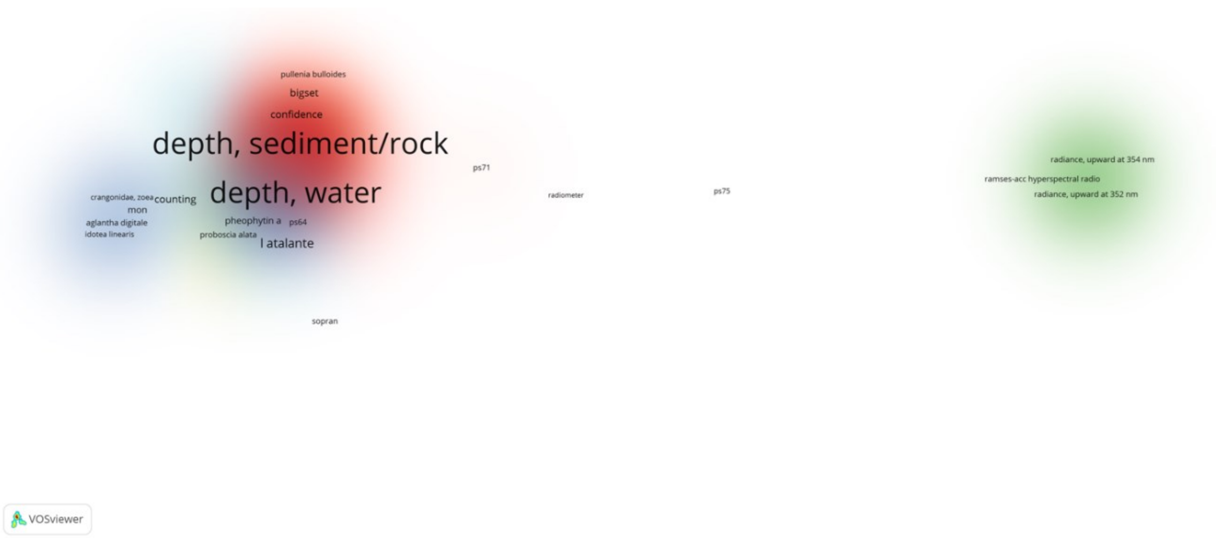
Oceanographic studies seek to describe the characteristics of aquatic environments from various perspectives. According to Garrison, (2017), ocean studies involve disciplines such as geology, physics, biology, chemistry, and engineering. These oceanographic studies can be linked to the interior composition of the planet Earth and connected to the analysis of ocean floor sediments, in the field of marine geology. Others are related to the observation of wave dynamics, and physical oceanography. Others are linked to research on marine organisms the impact of atmospheric pollutants, and marine biology. Others are related to the treatment of gases and dissolved solids in the ocean, and chemical oceanography. And finally, naval engineering which designs oil platforms, ships, ports, and other structures.

The oceanography collection can be described based on data from (22,287) records available in PANGEA. This is according to the attribute (header setSpec (A1)), as indicated by (see Supplementary File S15) available at (https://osf.io/5h67b).

Consequently, the most prevalent thematic occurrences (dc: Subject (A7)), (see Supplementary File S16) available at (https://osf.io/qp7rw) represent data collected by research institutions such as the Center for Marine Environmental Sciences; MARUM, and the Ocean Drilling Program. Despite representing subject matter, these institutions are also international oceanography research institutions. An overview of all the themes can be seen in Fig 4.

Figure 4

Oceanographic research data topics (dc: Subject (A7)). This is a network map of the thematic density of Oceanographic research data.



Note. Source: Elaborated by the Authors (2024).

The (3,602) keywords, which have a minimum frequency of (20) in the dataset, reveal that the data, (see Supplementary File S17) available at (https://osf.io/9zqb6) constitutes the nucleus of the oceanography data collection. The thematic center of PANGEA's oceanography collection is data on water temperature and pressure, primary carbon and oxygen production, depth and sedimentology, gas and mineral dating, temporal and compositional studies of gases and minerals, algae and water nutrients, and orfiis irradiance. Studies of irradiance and its influence on marine biogeochemistry, with a focus on the Indian and tropical Atlantic oceans, as well as the Peruvian basin. Finally, terms such as "CTD-RO" and "CTD/rosette" demonstrate the utilization of specific instruments for the measurement of the physical properties of water, which are fundamental for the collection of oceanographic data.

The data (see Supplementary File S15) available at (https://osf.io/5h67b) indicates that the majority of data types (dc: Format (A4)) are text/tab-separated-values and application/zip. The sources of this data (dc: Source (A8)) are research data publications produced by research institutions and universities. The principal contributing institutions are Universidad Arturo Prat, Iquique; Marine Research Institute, Reykjavik/Iceland; Shirshov Institute of Oceanology; Alfred Wegener Institute; GEOMAR - Helmholtz Center for Ocean Research Kiel and IFM-GEOMAR Leibniz-Institute of Marine Sciences; German Center for Marine Biodiversity Research and Leibniz Institute for Baltic Sea Research; European Pollen Database (EPD). This indicates Germany's strong presence in marine research. Other institutions from different parts of the world, including the University of Pretoria (South Africa), Stazione Zoologica Anton Dohrn (Italy), and Instituto Antártico Argentino (Argentina); Department of Biology, University of Ghent (United States Geological Survey), demonstrate the global scope of research. The variety of institutional sources ensures the quality and diversity of the data, allowing for a deeper understanding of oceanographic processes and their global implications.

In this sense, Kaleschke and Müller (2022) describe oceanographic research data that may be found at https://doi.org/10.1594/PANGAEA.941334, with (dc: title (A5)) Sea Ice Drift from Autonomous Measurements from 15 Buoys, Deployed During the IRO2/SMOSIce Field Campaign in the Barents Sea, March 2014. The (dc: description (A5)) represents a set of research data linked to the publication Tidal dissipation from free drift sea ice in the Barents Sea assessed using GNSS beacon observations. The latter was published in the journal Ocean Dynamics, under the following (DOI): https://doi.org/10.1007/s10236-022-01516-w. This dataset is particularly suited to investigating the interaction between sea ice and tides due to its high temporal resolution. The surface temperature and air pressure measurements were not subjected to quality checks.

The subsequent set of data on physical phenomena related to Geosciences represents a pertinent type of geophysical research data.

4.5 Geophysical Research Data

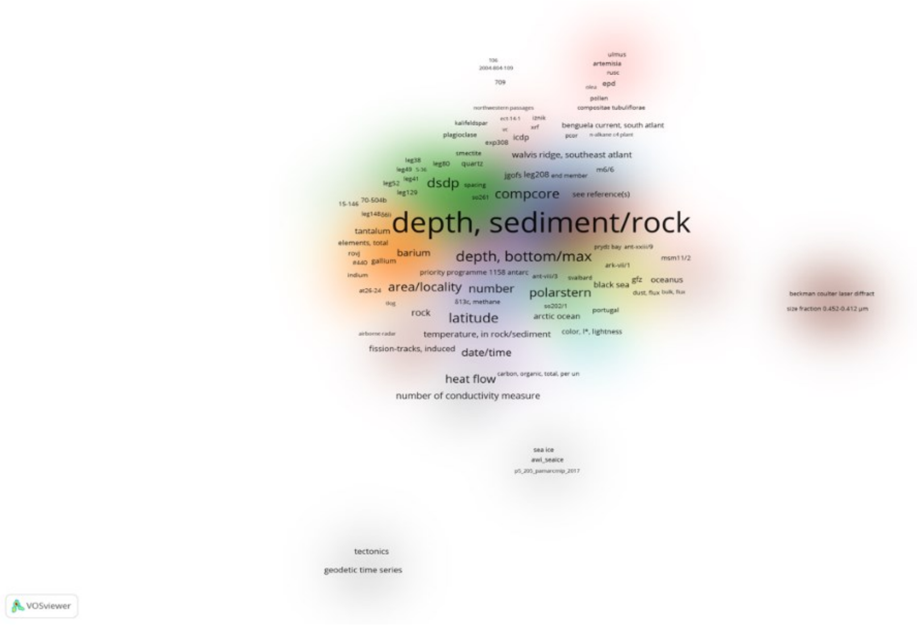
Geophysical represents studies of the structure, dynamics, and evolution of the Earth using the principles and methods of physics, involving the investigation of natural phenomena such as the magnetic field, gravity, earthquakes, and the internal structure of the planet through measurements and quantitative analysis (Backus, 1996).

The geophysical collection can be described according to the data, (see Supplementary File S18) available at (https://osf.io/ewjrt), based on the (4,175) records available in PANGEA, attribute (header - setSpec (A1)).

The classified research data on geophysics, represented in the (20) the largest occurrences, the core of the collection in PANGEA (dc: Subject (A7)), (see Supplementary File S19) available at (https://osf.io/jr54c), represent work carried out in projects such as the Ocean Drilling Program, which, although represented as a subject, represents an interdisciplinary and international program of research in Geosciences linked to ship expeditions in the oceans. The main topics are shown in Fig 5.

Figure 5

Geophysical research data topics (dc: Subject (A7)). This is a network map of the thematic density of Geophysical research data.



*Note.*Source: Elaborated by the Authors (2024).

Based on (2,647) keywords with a minimum frequency (5) in the dataset, the core of the collection can be seen in the dataset (see Supplementary File S20) available at (https://osf.io/kd6m9). This collection covers a wide range of topics in marine geophysics, providing a solid foundation for research in sedimentology, tectonics, and advanced chemical analysis techniques. Terms such as "MARUM" and "Center for Marine Environmental Sciences" indicate the participation of institutions specialized in marine sciences and reflect institutional cooperation in geophysical projects. The diversity of keywords and their frequency indicate a well-structured and comprehensive collection capable of supporting advanced and multidisciplinary studies in the Geosciences. PANGEA's geophysics collection is thematically centered around sedimentological studies of marine rocks at high depth; tectonics of Tierra del Fuego in Argentina; fission track and chemical elements; chromatography and mass spectrometry techniques; applied to the precise physical analysis of chemical compounds in samples, and georeferencing and data cataloging used geographic coordinates and event labels to systematically organize and analyze data.

According to the data (see Supplementary File S18) available at (https://osf.io/ewjrt), most of the data types (dc: Format (A4)) are text/tab-separated-values and application/zip. The sources of these data and the publications to which they are related (dc: Source (A8)) are also data produced by research institutions and supplementary material published in scientific papers. The sources of the data in the PANGEA geophysical collection include contributions from large databases, marine research institutions, and scientific paper supplements, reflecting some of these institutions and databases are the European Pollen Database (EPD); Research Institutions such as Alfred Wegener Institute, University of Cologne, and GeoForschungsZentrum Potsdam, are renowned institutions that regularly contribute to the collection. Beyond institutions, many supplements to scientific papers are published, such as in Earth and Planetary Science Letters, demonstrating the integration of high-quality peer-reviewed data.

In this sense, O'Nions, Hamilton and Evensen (1977) describe a sample (dc_identified (A2)) of geophysics research data is https://doi.pangaea.de/10.1594/PANGAEA.721776, where (dc: title (A5)) is Nd and Sr isotope ratios of oceanic basalt. The (dc: description (A5)) represents a set of research data metadata associated with the publication Variations in 143Nd/144Nd and 87Sr/86Sr ratios in oceanic basalts. Published in the journal Earth and Planetary Science Letters, (DOI): https://doi.org/10.1016/0012-821X(77)90100-5. This dataset is particularly suitable for studying the geophysical responses of oceanic basalts dredged and excavated from oceanic islands in Iceland and the Reykjanes mountain range.

The next set of research data is hydrological, a relevant type of data on rivers and lakes.

4.6 Hydrological Research Data

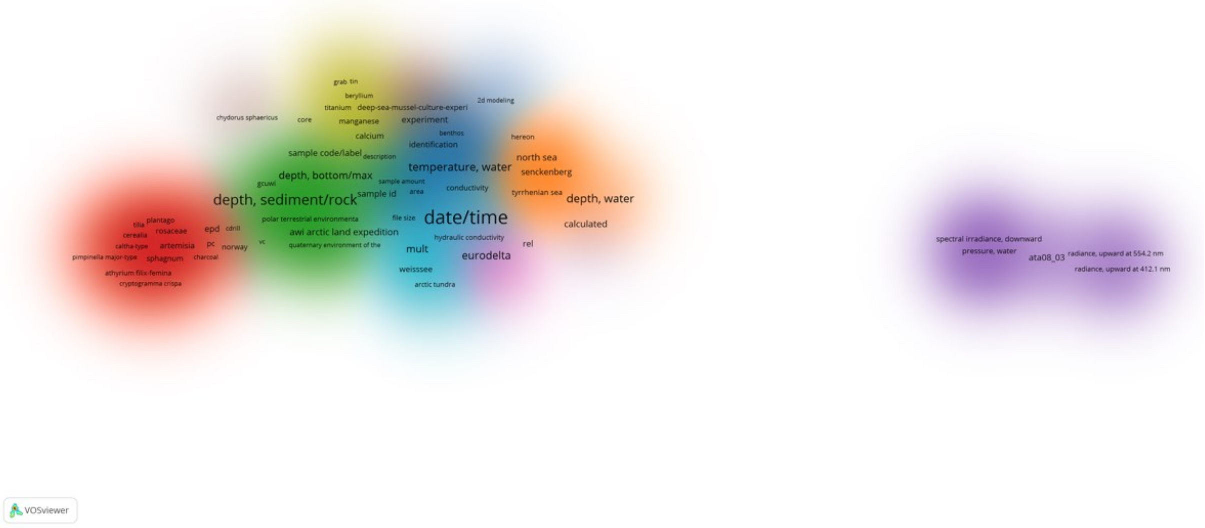
According to Fetter (1994), hydrology is the science that studies the distribution, circulation, and properties of water on Earth and in the atmosphere. This includes the hydrologic cycle, the processes of evaporation, precipitation, infiltration, and storage of water in various compartments of the environment such as rivers, lakes, aquifers, and oceans. Hydrology also studies how water interacts with soil, vegetation, and the atmosphere, as well as how it affects ecosystems and human activities.

Based on the (834) records available at PANGEA, the collection of hydrological research data metadata on rivers and lakes can be defined as the data (see Supplementary File S21) available at (https://osf.io/27u3r). The (20) the largest occurrences, (see Supplementary File S22) available at (https://osf.io/xbe5y), the core of the PANGEA collection, are an example of this type of research data.

The largest thematic occurrences (dc: Subject (A7)) represent acoustic Doppler current profiling; current velocity, horizontal; current velocity, vertical; date/time; depth, bathymetric; depth, water; MARUM; MATLAB date; North Sea; senckenberg; suspended particles; Alaska, USA; permafrost research (periglacial dynamics); remote sensing (Landsat); Western-Alaska-LCC; which, while representing themes, also represent some research institutions such as the Center for Marine Environmental Sciences, MARUM in Bremen, Germany. It is also possible to observe all the themes in Fig 6.

Figure 6

Hydrological research data topics (dc: Subject (A7)). This is a network map of the thematic density of Hydrological research data.



Note. Source: Elaborated by the Authors (2024).

The core of the research data collection analyzed focuses on four main areas temporal and multi-paradigmatic documentation, depth studies and sedimentology, physicochemical water parameters, and European coordination of regional studies, the core of the collection can be seen in the dataset, (see Supplementary File S23) available at (https://osf.io/u7gpz). Examples and highlights in the data are the coordination of European efforts in specific regional studies, such as the Mediterranean and Black Sea deltas, as well as the study of the Têt Basin. Concepts related to water depth and regional hydrological studies, such as those carried out in the North Sea, indicate an emphasis on hydrodynamic and oceanographic characteristics; time records, essential for documenting time series and analyzing trends over time; physicochemical water parameters, essential for characterizing environmental conditions; mainly geological and depth studies, including the characterization of sediments and rocks at different depths, as well as specific events occurring in these water layers.

Most of the data types (dc: Format (A4)) are text/tab-separated-values and application/zip according to the data (see Supplementary File S21) available at (https://osf.io/27u3r). The sources of these data (dc: Source (A8)) as well as data produced by research institutions and supplementary materials published in scientific papers. Data provided by the European Pollen Database (EPD), the Leibniz Institute for Baltic Sea Research, Warnemünde, and other institutions such as the ForschungszentrumJülich GmbH and the GeoForschungsZentrum Potsdam, the Oceanographic Institute of the University of São Paulo Brazil and the Max Planck Institute for Marine Microbiology are described as sources. These data indicate various supplementary sources used in scientific publications, with some institutions and databases appearing repeatedly.

In this sense Giertz and Diekkrüger (2003) describe an example, (dc_identifie (A2)), of research data in hydrology https://doi.pangaea.de/10.1594/PANGAEA.831196, with (dc: title (A5)) Discharge data derived from five water level gauges and discharge measurements in the Aguima and Niaou catchment, Benin, West Africa, which represents - (dc: description (A5)), a set of research data associated with the publication Analysis of hydrological processes in the sub-humid tropics of West Africa under special consideration of land use on the example of the Aguima catchment in Benin. Published in the journal Physics and Chemistry of The Earth, (DOI): https://doi.org/10.1016/j.pce.2003.09.009. This dataset is particularly suitable for the study of hydrological processes related to soil properties in the Aguima catchment in Benin (West Africa). It presents an overview of the concept of hydrological measurement and highlights the impact of land use in cultivated areas compared to natural land cover close to the watershed, which shows significant differences in runoff behavior.

The following section presents a discussion of the empirical data presented in the research study.

5 Discussion

The results of our study suggest the need for two distinct areas of discussion. The first of these concerns the proposal of a definition for Geosciences research data. The second, of a methodological nature, concerns the comparison of Dublin Core metadata standards with the specifications adopted by data journals to describe the nature of research data in data papers.

PANGEA's (6) six research data metadata collections (geochemistry, atmospheric research, paleontology, oceanography, geophysics, and hydrology) allow us to discuss the concept of the Geosciences research data. Thus, based on Table 2 and Table 3, which describe the specific attributes (knowledge domain, accessibility of the data, language, type of data, how the data was acquired and the state of the data, location of the data, specific subject area and related publication) compared to the Dublin Core metadata fields of PANGEA, the following concept is proposed.

Geosciences research data is data on studies that relate to the Earth, the Atmosphere, and the Oceans, based on different disciplines (such as geochemistry, atmospheric research, paleontology, oceanography, geophysics, and hydrology) and technologies (such as satellites, electron microscopes, climate sensors, ships, computer modeling), and other sources related to the study of the Earth and its processes. Each of these domains generates data for global understanding of physical and chemical processes, including the composition, structure, and evolution of the planet, as well as environmental and climate phenomena.

Geosciences research data is collected by various research institutions, through international collaborations and marine and terrestrial exploration projects. This research data covers a wide range of topics, such as sedimentology, chemical analysis, physical analysis, temperature and pressure measurements, fossil studies, deep seas, and climatic phenomena. The predominant data formats vary between text, tables, images, zip files, MPEG, kml, and other less frequent formats. They are the result of the use of analytical instrumentation, sampling, laboratory techniques, acoustic characterization, wave detection, depth sensors, and remote sensing technologies. The comprehension of this data facilitates the visualization of environmental and climate modifications in the surface, atmosphere, and oceans of the Earth.

Another point of discussion, based on Table 2, is the comparison of Dublin Core metadata standards with specifications adopted by data journals to describe the nature of research data in data papers. This method served to visualize and describe large amounts of research data metadata according to their similarities and a common standard (A1-10). Our research indicates that the reuse of research data metadata has contributed to significant discoveries about the nature of the thousands of data items deposited in PANGEA. Furthermore, the clustering of metadata records has allowed for a deeper reflection on the content of the data repositories, facilitating the advancement of scientific knowledge and interdisciplinary research.

The reuse of Geosciences research data metadata, which includes observations, findings, and results related to data reuse, can enhance the reproducibility of research, facilitate scientific discovery, and encourage interdisciplinary collaboration within the Geosciences, Information Science, and Librarianship. One limitation of our study is the sample size. The results demonstrated that in April 2024, PANGEA had a total of (426,272) research data metadata records, of which only (137,218) were analyzed to define the nature of (6) six types of Geosciences research data. Although PANGEA is a disciplinary data repository for earth and environmental sciences, not all of its collections are related to Geosciences. Indeed, PANGEA contains a large part of its data collection in other fields, such as biology, which is outside the scope of our investigation. This was the limitation of our study.

Future research could focus on analyzing different data repositories and in different contexts to verify the robustness of our concept for Geosciences research data. Furthermore, an analysis of data journals could provide a more comprehensive understanding of this type of research data.

The final considerations of the study are addressed in the following section.

6 Conclusion

Geosciences is a field of study that employs both laboratory and fieldwork techniques. It makes use of cutting-edge technologies such as satellites, electron microscopes, climate sensors, and ships to investigate chemical, physical, and geological phenomena related to the planet Earth as a system. At PANGEA metadata, Geosciences research data is defined as the collection, analysis, and interpretation of data from direct observations, field experiments, laboratory measurements, satellite imagery, computer modeling, and other sources related to the study of the Earth and its processes.

In this sense, the organization of these research data metadata is a heterogeneous task and allows the creation of models and the reuse of research already carried out, making it possible to qualify collections and increase the visibility of researchers who share their data. Accessing and preserving various types of research data metadata allows materials used before, during, and after a scientific investigation to be used and reused to create other investigations and to better teach methodologies to students.

Finally, research data metadata are disciplinary objects that depend on a specific domain of knowledge. However, their reuse is linked to the use of these objects as a research resource, regardless of the time of use, the purpose, the characteristics of the data, and the user. It is important to keep in mind the research question, the methodology, and the research method used. The same research question, methods, and data would lead to a replication of the research, while a methodological change would lead to a replicability of the sciences. Different data would lead to new research or replication of the research. Research that considers a different research question and uses the same data and methods leads to reinterpretation. Reuse is the application of different methods and different data to a different research question, and reuse also occurs when the same methodology and data are applied.

Scientific research provides these facets of the reuse of research data metadata. It can be concluded that Geosciences research data are interdisciplinary objects that can be reused across disciplinary boundaries, regardless of when and how they were used. When applied to relevant research questions, reuse in the Geosciences can include replication, reanalysis, new research, reproduction, reinterpretation, or even the application of different methods and data to new research questions. This illustrates the many facets of reuse in scientific research.

References

Backus, G. E. (1996). *Foundations of Geophysics.* Cambridge University Press.

Bienhold, C.; &Boetius, A. (2015). *Porosity in sediment sores from the Central Arctic Ocean during POLARSTERN cruise ARK-XXVII/3 from August-September 2012* [Dataset]. PANGAEA. https://doi.org/10.1594/PANGAEA.849054

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078; 2012. https://doi.org/10.1002/asi.22634.

Clarkke, F. W. (1924). *The data of Geochemistry* (5th ed.). United States Geological Survey, Washington Government Printing Office. https://pubs.usgs.gov/bul/0770/report.pdf

Daniels, M. G. (2014). *Data reuse in museum contexts: Experiences of archaeologists and botanists* [Dissertation]. University of Michigan. http://hdl.handle.net/2027.42/108953

Federer, L.; Lu, Y.; Joubert, D.; Welsh, J. &Brandys, B. (2015, june). Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff. *PLOS One.*https://10.1371/journal.pone.0129506

Felden, J.; Möller, L.; Schindler, U. et al. (2023). PANGAEA - Data Publisher for Earth & Environmental Science. *Sci Data, 10*(347). https://doi.org/10.1038/s41597-023-02269-x

Fetter, C. W. (1994). *Applied hydrogeology* (3rd ed.). Prentice Hall.

Foote, M.; & Miller, A. (2007). *Principles of paleontology* (3rd ed.). Freeman and Company.

Garrison, T. (2017). *Fundamentos de oceanografia.* Cengage.

Gastaldello, M.; Agnini, C., Westerhold, T.; Drury, A.; &Alegret, L. (2024). *Age model, carbonate mass accumulation rates and benthic foraminifera from ODP Site 175-1085* [Dataset bundled publication]. PANGAEA. https://doi.org/10.1594/PANGAEA.962075

Giertz, S.; &Diekkrüger, B. (2003). *Discharge data derived from five water level gauges and discharge measurements in the Aguima and Niaou catchment* [Dataset publication series]. PANGAEA. https://doi.org/10.1594/PANGAEA.831196

Gould, S. J. (2002). *The structure of evolutionary theory*. Belknap Press. https://archive.org/details/TheStructureOfEvolutionaryTheory

Jiao, C.; &Darch, P. T. (2020). The role of the data paper in scholarly communication. *Proc AssocInfSciTechnol, 57,* e316. https://doi.org/10.1002/pra2.316

Jiao, H.; Qiu, Y.; Ma, X.; & Yang, B. (2024). Dissemination effect of data papers on scientific datasets. *Journal of the Association for Information Science and Technology, 75*(2), 115-131. https://doi.org/10.1002/asi.24843

Jones, P.; Wheeler, D.; Können, G.; Koek, F.; Prieto, M.; &García-Herrera, R. (2007). *Climatological observations from ship logbooks between 1750 and 1854 (release 2.1)* [Dataset publication series]. PANGAEA. https://doi.org/10.1594/PANGAEA.611088

Jones, R. W. (2011). *Applications of paleontology: Techniques and case studies.* Cambridge University Press.

Kaleschke, L.; & Müller, G. (2022). *Sea ice drift from autonomous measurements from 15 buoys, deployed during the IRO2/SMOSIce field campaign in the Barents Sea March 2014* [Dataset publication series]. PANGAEA. https://doi.org/10.1594/PANGAEA.941334

Keller, E. A.; &Devecchio, D. (2019). *Introduction to Environmental Geology.* Pearson.

Kim, J. (2020). An analysis of data paper templates and guidelines: Types of contextual information described by data journals. *Science Editing, 7*(1), 16-23.

Köppen, W. (1931). *Grundriss der Klimakunde: Outline of climate science*. Walter de Gruyter& Co. https://api.pageplace.de/preview/DT0400.9783111667751_A40793869/preview-9783111667751_A40793869.pdf

Li, K.; & Jiao, C. (2022). The data paper as a sociolinguistic epistemic object: A content analysis on the rhetorical moves used in data paper abstracts. *Journal of the Association for Information Science and Technology, 73*(6), 834-846. https://doi.org/10.1002/asi.24585

Lyell, C. (1853). *Principles of Geology: The modern changes of the earth and its inhabitants* (9th ed.). Little, Brown and Company. https://archive.org/details/principlesgeolo00lyelgoog/page/n5/mode/2up

O'Nions, R. K.; Hamilton, P. J.; &Evensen, N. M. (1977). *Nd- and Sr- isotope ratios of oceanic basalts* [Dataset publication series]. PANGAEA. https://doi.org/10.1594/PANGAEA.721776

Pampel, H. et al. (2013, november 4). Making research data repositories visible: The re3data.org registry. *PLOS One*.https://doi.org/10.1371/journal.pone.0078080

Phillips, M. (2013). Metadata Analysis at the Command-Line. *Code4Lib, 19*. https://journal.code4lib.org/articles/7818

Rice, R.;&Southall, S. (2016). *The data librarian's handbook*. Facet Publishing.

Rohli, R.; &Viega, A. (2008). *Climatology*. Jones and Bartlett.

Semeler, A. R. (2024). Reuse of metadata Pangea Data Publisher for Earth & Environmental Science Repository [Dataset]. OSF. osf.io/3bsx2

Shutsko, A.; & Stock, W. (2023). Information scientists' motivations for research data sharing and reuse. *Libri, 73*(4), 307-320. https://doi.org/10.1515/libri-2023-0052

Tarbuck, E. J.; Lutgens, F. K.; &Tasa, D. (2015). *Earth Science*. Pearson.

Tenopir, C. et al. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLOS One, 10*(8), e0134826. https://doi.org/10.1371/journal.pone.0134826

Uzwyshyn, R. (2016, april). Research data repositories: The what, when, why, and how. *Computers In Libraries, 36*(3), 18-21. https://www.researchgate.net/publication/304780954_Online_Research_Data_Repositories_the_What_When_Why_and_How

Van de Sandt, S.; Dallmeier-Tiessen, S.; Lavasa, A.; &Petras, V. (2019). The definition of reuse. *Data Science Journal, 18*(1), Article 22, 1-19. https://doi.org/10.5334/dsj-2019-022

Walters, W. H. (2020). Data journals: incentivizing data access and documentation within the scholarly communication system. *Insights: the UKSG journal, 33*, Article 18, 1-20. https://doi.org/10.1629/uksg.510

White, W. M. (2013). *Geochemistry*. Wiley-Blackwell.

Publishing data

Alexandre Ribas Semeler

I hold a PhD and a postdoctoral degree in Information Science from the Federal University of Santa Catarina. I currently work as a data librarian at the Institute of Geosciences of the Federal University of Rio Grande do Sul in Brazil. As an independent researcher and data librarian, I have an interdisciplinary interest in data Librarianship. I believe in the fourth paradigm of sciences (e-science and digital humanities) and see the current digital data technologies as great transformation drivers in academic libraries.

alexandre.semeler@ufrgs.br

https://orcid.org/0000-0002-8036-4271

Luana Farias Sales

PhD in Information Science from the Graduate Program at IBICT/UFRJ (2011-2014). Master's in Information Science from the UFF/IBICT agreement (2004-2006), Degree in Library Science and Documentation from the Fluminense Federal University (2003). Productivity scholarship holder Pq-B. Young Scientist of the State of Rio de Janeiro. She is a C&T Analyst at MCTI/IBICT, teaching in the Postgraduate Program in Information Science under the IBICT-UFRJ agreement and at DIECI - Scientific Publishing Division. She is the General Coordinator of the GO FAIR Brazil office.

luanasales@ibict.br

http://orcid.org/0000-0002-3614-2356

Adilson Luiz Pinto

Graduated in Library Science from PUC-Campinas (2000), Master in Information Science from PUC-Campinas (2004) and in Audiovisual Documentation from Universidad Carlos III de Madrid (2006); PhD in Documentation from Universidad Carlos III de Madrid (2007). Member of LEMME Lab and Leader of Metric Studies in Data Librarianship and Geosciences; Editor of the Iberoamerican Journal of Science Measurement and Communication.

adilson.pinto@ufsc.br

https://orcid.org/0000-0002-4142-2061

Roberta Pereira da Silva de Paula

PhD student in Information Science at PPGCI - IBICT/UFRJ (Start 2020). Master's in Information Science from the IBICT/UFF Agreement (2007). Graduated in Librarianship (2004) and Specialist in Knowledge Organization for Information Retrieval (2005) from UNIRIO. She is currently Head of Library at the Geological Survey of Brazil.

beta.depaula@gmail.com

https://orcid.org/0000-0002-4546-2239

Valquer Cleyton Paes Gandra

Master's student in Information Science at PPGCI IBICT-UFRJ. Postgraduate in UI and UX Digital Product Design from UNOPAR. Bachelor in Library Science from UNIRIO. Postgraduate student in Data Science at UNOPAR. Qualification in Access to Scientific and Technological Health Information from ICICT-FIOCRUZ.

cleytonvalquer@gmail.com

https://orcid.org/0000-0003-0476-1651

Heloisa Costa

She has a degree in Library Science from the Federal University of Santa Catarina (UFSC), a specialization in Information Unit Management from the State University of Santa Catarina (UDESC), a PhD and a Master's degree in Information Science from UFSC, in the Postgraduate Program in Information Science (PGCIN-UFSC). She is a substitute lecturer in the Department of Information Science at the Federal University of Santa Catarina. She has experience as a consultant in the management of documentary and bibliographic collections and in the field of Information Science, with an emphasis on the management of information units and document management. She works as a proofreader of documents and academic papers, including ABNT standardization.

helocosta7@hotmail.com

http://orcid.org/0000-0003-2380-1831

Originality

I declare that the text is original and is not under evaluation in any other publication.

Preprints

The manuscript has not been submitted to the Preprints platform.

Author contribution

Conception and preparation of the manuscript: Semeler, A. R.; Sales, L. F.;
Data collection: Semeler, A. R.; Sales, L. F.; Pinto, A. L.; Paula, R. P. S.; Gandra, V. C. P.; & Costa, H.
Data analysis: Semeler, A. R.; Sales, L. F.; Pinto, A. L.; Paula, R. P. S.; Gandra, V. C. P.; & Costa, H.
Discussion of results: Semeler, A. R.; Sales, L. F.; Pinto, A. L.; Paula, R. P. S.; Gandra, V. C. P.; & Costa, H.

Review and approval: Semeler, A. R.; Sales, L. F.; & Pinto, A. L.;

Use of artificial intelligence

Not applicable.

Funding

This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Funding Code 001; the National Council for Scientific and Technological Development (CNPq), through the call for productivity grants Process No 312627/2023-8 and the Carlos Chagas Filho Foundation for Research Support in the State of Rio de Janeiro (FAPERJ) Process No 201.380/2022.

Consent to use image

Not applicable.

Research ethics committee approval

Not applicable.

Conflict of interest

Not applicable.

Data availability and dataset

Semeler, A. R. (2024, June 18). Reuse of metadata PANGEA Data Publisher for Earth and Environmental Science Repository [Conjunto de dados]. OSF. https://doi.org/10.17605/OSF.IO/3BSX2

License of use

Publisher

Published by the University Library System of the University of Pittsburgh. Responsibility shared with partner universities. The ideas expressed in this article are those of the authors and do not necessarily represent the views of the editors or the university.

Editors

Fabiano Couto Corrêa da Silva, Lúcia da Silveira.

Background

Received: 02-08-2024- Approved: 23-11-2024 - Published: 07-02-2025.