ON THE DEVELOPMENT OF A DATASET PUBLICATION GUIDELINE: DATA REPOSITORIES AND KEYWORD ANALYSIS IN ISPRS DOMAIN

L. E. Budde, T. Kullmann, D. Iwaszczuk

Dept. of Civil and Environmental Engineering, Remote Sensing and Image Analysis, Technical University of Darmstadt
Germany - (lina.budde, dorota.iwaszczuk)@tu-darmstadt.de, timo.kullmann@stud.tu-darmstadt.de

KEY WORDS: Data repository, FAIR principle, data life cycle, ISPRS Annals.

ABSTRACT:

The FAIR principle (find, access, interoperability, reuse) forms a sustainable resource for scientific exchange between researchers. Currently, the implementation of this principle is an important process for future research projects. To support this process in the ISPRS community, the usage of data repositories for dataset publication has the potential to bring closer the achievement of the FAIR principle. Therefore, we (1) analysed available data repositories, (2) identified common keywords in ISPRS publications and (3) developed a tool for searching appropriate repositories. Thus, infrastructures from the field of geosciences, that can already be used, become more accessible.

1. MOTIVATION

With the funding of the European open science cloud (EOSC Future, 2019) and Gaia-X (Gaia-X European Association for Data and Cloud AISBL, 2019) at the European level and national research data infrastructures such as NFDI in Germany (Nationale Forschungsdateninfrastruktur (NFDI) e.V., 2021), (research) data management has come into focus. For guiding the processes in data management, the FAIR principle (find, access, interoperability and reuse) is used as a key component (Kinkade and Shepherd, 2022). Lack of interoperability, for example, restricts the use of data and constrains research innovations (Atkinson et al., 2022). In addition, the increasing amount of data poses a challenge in terms of reusability and reproducibility (Trisovic et al., 2021; Crystal-Ornelas et al., 2022).
To overcome such limitations, the FAIR principles must be taken into account throughout the whole data life cycle (Kinkade and Shepherd, 2022). Various tools can support the researcher in making data and metadata FAIR. BeMeDa (Budde et al., 2022), for example, improves the findability of photogrammetry and remote sensing datasets. Prior to the publication of new data sets, efforts should be made to take the FAIR principle into account. However, there are a variety of data publication options and workflows and these are also often very domain-specific (Austin et al., 2017). Currently, the ISPRS only provides guidance for articles and websites, but no (domain) specific information regarding the publishing of data (ISPRS, 2022). Few contributions point to lack of practices (Ivánová et al., 2019; Atkinson et al., 2022). Thus, the development of a guidance for data publishing in the ISPRS domain has never been as important as it is now.
With our ISPRS initative we want to develop such a guideline. This will assist in the publication of new datasets and provide a basis for a standardized approach. In addition, this guide is integrated into a possible workflow for the dissemination of datasets in the ISPRS community using existing data repositories (Figure 1).

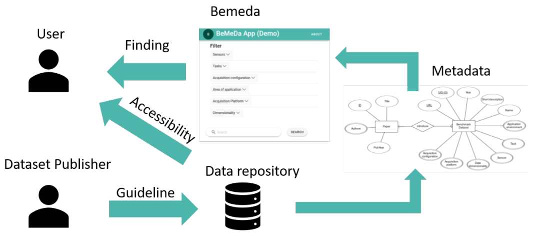
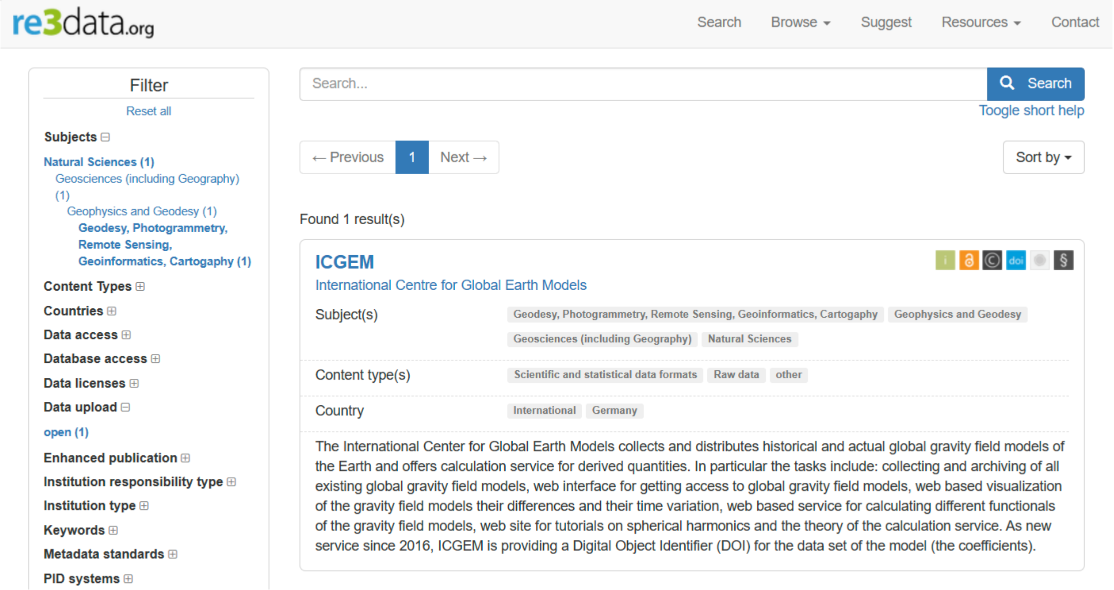


Figure 1. Relation of BeMeDa, data repositories and a guideline for dataset publisher.

In this work, we first present a short overview about the topic of already available data repositories (section 2). The next section presents the specification of our domain to better evaluate the requirements of the selected

https://doi.org/10.5194/isprs-archives-XLVIII-1-W3-2023-17-2023 | © Author(s) 2023. CC BY 4.0 License.



Found 1 result(s)

ICGEM
International Centre for Global Earth Models

Subject(s)	Geodesy, Photogrammetry, Remote Sensing, Geoinformatics, Cartography Geophysics and Geodesy Geosciences (including Geography) Natural Sciences
Content type(s)	Scientific and statistical data formats Raw data other
Country	International Germany

The International Center for Global Earth Models collects and distributes historical and actual global gravity field models of the Earth and offers calculation service for derived quantities. In particular the tasks include: collecting and archiving of all existing global gravity field models, web interface for getting access to global gravity field models, web based visualization of the gravity field models their differences and their time variation, web based service for calculating different functionals of the gravity field models, web site for tutorials on spherical harmonics and the theory of the calculation service. As new service since 2016, ICGEM is providing a Digital Object Identifier (DOI) for the data set of the model (the coefficients).

Figure 2. Screenshot of repositories found in re3data (GFZ German Research Centre For Geosciences et al., 2013) filtered by the subject “Geodesy, Photogrammetry, Remote Sensing, Geoinformatics, Cartogaphy” and “open” data upload.

data repositories. In section 4, finally, the current state of our developed repository finder is presented.
This finder is an important component in order to reach our goal of a publication guide for datasets.

2. DATA REPOSITORIES

To publish data, different options are available. In practice, data is commonly shared via request or local websites (Austin et al., 2017). Data repositories, i.e. institutional, domain-specific or general repositories already contain a strategy for publishing data so that this data can be reused (Kinkade and Shepherd, 2022). Data repositories make a significant contribution for this reusability (Trisovic et al., 2021). Although data repositories can be found using the re3data (GFZ German Research Centre For Geosciences et al., 2013) repository registry, the application does not target the specific needs of the ISPRS community. Re3data provides only a single results for the subject “Geodesy, Photogrammetry, Remote Sensing, Geoinformatics, Cartogaphy”, which is open for data upload (Figure 2). This shows the lack of existing repositories in this subject domain. In contrast, the parent category “geosciences (including geography)” lists at least 15 proposed repositories.
One example for a geoscience-specific repository is PANGAEA (Felden et al., 2023). The PANGAEA repository uses the FAIR principle and DOI for permanent identification. Further benefits are the 2 TB free data space, different open licence options and the standardized metadata concept. In contrast, general repositories such as figshare (figshare, 2011; Singh, 2011) contain significantly more datasets compared to PANGAEA, but they also include datasets without geodata relation. Thus, the comparison of data repositories is difficult and depends on how domain specific they are.
Furthermore, ensuring data quality is an important feature of data repositories (Kindling and Strecker, 2022). In particular, the quality of metadata has a significant impact on the reusability of data (Elouataoui et al., 2022; Kindling and Strecker, 2022). However, the requirements of high quality metadata can be guided by repositories (Trisovic et al., 2021). Additionally, when dealing with repositories with such high amounts of datasets, like figshare, the quality of the data can vary and therefore not be in line with the requirements for a given project. Some repositories put more emphasis on a good and standardized quality for all uploads while others are more liberal. Figshare, for example, has a dedicated review process to ensure adequate data quality before publishing (figshare, 2011).

3. DOMAIN SPECIFICATION

For the specification of our domain and to evaluate data repositories, the keywords of papers from the IS-

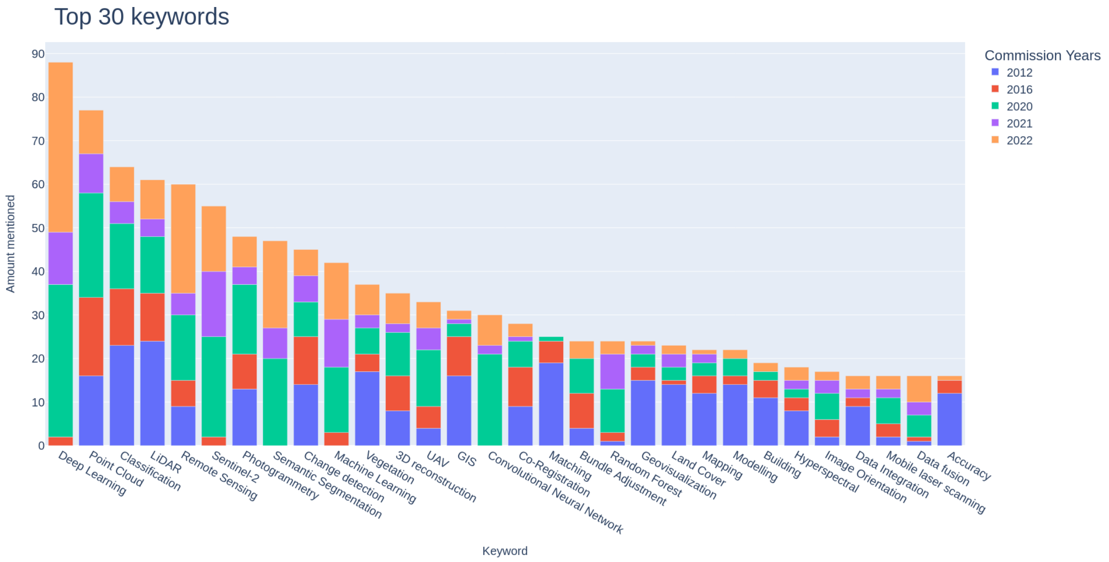


Figure 3. The 30 top most frequent annals paper keywords from the past ten year ISPRS congress. Similarly spelled keywords are grouped together e.g. “point cloud”, “3d point clouds”, “3d point cloud”, “point clouds” and “3d-point cloud”.

PRS congresses held over the last ten years and listed in the annals publications are analysed.

3.1 Keyword extraction and connection

For this analysis, each paper keyword is extracted similar to Budde et al. (2022). After collecting the keywords with the help of a web crawler and the number of corresponding paper from the ISPRS website, the keywords are automatically grouped. For this grouping, keywords with similar spelling are fused, e.g. “LiDAR” and “LIDAR” grouped to the keyword “LiDAR”. The word similarity values are measured using the Levenshtein distance. Thus, the keywords must have a similarity of at least 80 %. Due to this threshold, a small number of shorter keywords are prone to be grouped incorrectly, for example, “GIS” and “GIDS”. To have a fully automated process, we avoid manual corrections for such abbreviation keywords. An increased similarity threshold would lead to missing fusion of keywords.

Based on the extracted and grouped keywords, the connections to each other can be determined. Several keywords are mentioned for each publication. So, the assignment of the keywords to their respective publication is used. A keyword network is created from the number of publications in which keywords are mentioned together. A keyword combination must occur at least three times to be considered in the keyword network.

3.2 Results

According to Figure 3, with a difference of 11 mentions, deep learning is the most frequently studied subject area. Followed by the second placed “Point Cloud” and third placed “Classification”. “Point cloud” is a keyword directly related to data and in combination with “LiDAR” on the fourth place, 3d reconstruction and mobile laser scanning, which also appear in the top 30, indicates a high degree of use of 3D data. In particular, the mention of data acquisition by means of laser scanning or drone allows to assume that the corresponding data is acquired itself or belongs to an existing accessible dataset. This keywords show a high potential to represent part of the ISPRS domain in terms of dataset publication.

Figure 4 shows the connections between keywords that are mentioned in the same publication. The size of the link represents the frequency of this specific connection. The size of the node on the other hand shows the overall frequency of the respective keyword. As result, LiDAR and point cloud are mapped close to another. In addition, deep learning, semantic segmentation and remote sensing are often related. This is reinforced by the fact that the connection between deep learning and semantic segmentation is the largest overall.

Both the top keywords and keyword connections provide information about relevant topics in the ISPRS in the past years. Deep learning first appeared

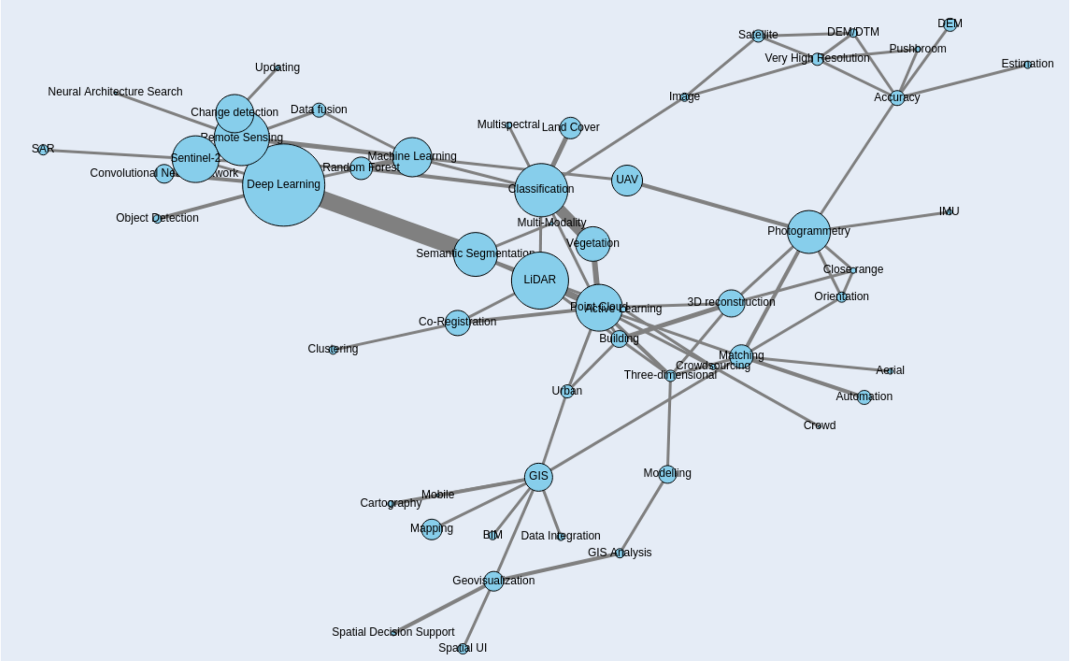


Figure 4. Visualization of the connections between keywords as force-directed graph. The size of the circles depends on the keyword frequency (Figure 3), the frequency of the link is symbolized with the thickness of the graph line.

as a keyword in 2016 and then became very prevalent in 2020 and 2022. The use of the keyword “matching”, on the other hand, decreased since 2016 and is no longer mentioned in 2021 and 2022.

In addition to the keywords photogrammetry and remote sensing, which are already included in the name ISPRS, the following keywords, for example, can be used to describe some main aspects of the domain, caused by their relevance for publications:

- LiDAR,
- 3D reconstruction,
- GIS,
- Co-Registration, and
- land cover.

4. REPOSITORY FINDER

The goal described in the following section is helping to find a fitting repository to publish data based on specific requirements of the data and the publishers. Before repositories can be suggested, a list of relevant repositories had to be collected.

4.1 Repository Selection and Evaluation

For a first overview of the most relevant repositories, a list of 13 better known data repositories was created. This list contains general and geospatial specific repositories. The user conditions of these selected repositories differs in some cases considerably. In addition, information about the repositories have a different level of detail, depending on the maintainer. While PANGAEA, for example, has a detailed wiki section that transparently describes the platform (Felden et al., 2023), it is much more difficult to find the relevant information on other platforms. Thus, quality descriptive repository attributes are specified for the comparison and evaluation of each collected repository. A short selection of the list of repositories and their attributes can be seen in Table 1. The collection of information about repositories and their properties allows a first assessment of common offerings. For example PANGAEA contains 547 LiDAR, 5112 remote sensing and 7323 point cloud related datasets. In contrast, figshare provides 883 LiDAR, 19,473 remote sensing and 58,500 point cloud datasets (September 2023). We used descriptive attributes with relation to user requirements, FAIR principle and domain support (Table 1). While all of the repositories listed provide a

DOI for datasets, the repositories differ in their use of metadata schemes. On the one hand there is partly no further information, on the other hand known schemes like DublinCore are used. Based on these differences, the relevance of the selected repositories is evaluated. For that, the available attributes are weighted equally, since their importance depends on the use-case and therefore is hard to determine beforehand.

4.2 Repository Selection Tool

The aformentioned selection of suitable repositories was used as a baseline selection for a newly created repository selection tool called RepoFinder. This tool allows the user to input their requirements for their data publishing and returns a list of repositories sorted by how well they fit the input parameters. This process simplifies the search for a repository and gives a good overview of the available options and their attributes. Figure 5 shows a screenshot of the selection tool as explained before. On the left is the current selection of preferences by the user. An empty field means that the user has no preference and therefore this does not influence the repositories that are shown as suitable. On the right are the filtered repositories sorted by their suitability. The suitability is visualized by the blue bar. If more information about a repository is needed, additional information can be found in the expandable section.

5. DISCUSSION AND FUTURE WORK

When developing a guideline for publishing research data, the specifics of data in the ISPRS domain must be considered. The keyword analysis shows, that both 3D and 2D data are used. In addition, there is usually a geo-reference. Existing data repositories offer are standardized workflow to improve the publication process. However, besides general repositories, only repositories in the superordinate geoscience context are available. Existing registries for finding repositories, such as re3data, allow only limited search for ISPRS matching repositories. By developing our own repository finder, we pre-select eligible repositories and evaluate them with the ISPRS keywords in addition to the FAIR criteria. Even though this is still being implemented at this stage, the tool is already offered in combination with BeMeDa and allows first selection options. BeMeDa supports the search of existing datasets in a meta search. The inclusion of a repository finder now also standardizes the publishing process and helps with a better overview of available options. In the next steps, we want to extend the analysis of the repositories with regard to ISPRS-specific requirements such as common used data types. Furthermore, the development of a standardized metadata concept which is suitable for all ISPRS dataset could be an important simplification. Nevertheless, to improve a keyword based search for papers, notes on the spelling of keywords, such as lowercase only, or suggestions for common abbreviations used as keywords, such as LiDAR, could be easily adapted for future paper publications.

ACKNOWLEDGMENT

We thank the ISPRS for funding the scientific initiative ”Publishing dataset guideline: gaps and trends in research data management in the ISPRS community”.

References

Atkinson, R. A., Zaborowski, P., Noardo, F., Simonis, I., 2022. SMART CITIES – SYSTEMS OF SYSTEMS INTEROPERABILITY AND OGC ENABLERS. X-4/W3-2022, 19–26.

Austin, C. C., Bloom, T., Dallmeier-Tiessen, S., Khodiyar, V. K., Murphy, F., Nurnberger, A., Raymond, L., Stockhause, M., Tedds, J., Vardigan, M., Whyte, A., 2017. Key components of data publishing: using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries*, 18(2), 77–92.

Budde, L. E., Schmidt, J., Javanmard-Ghareshiran, A., Hunger, S., Iwaszczuk, D., 2022. DEVELOPMENT OF A DATABASE FOR BENCHMARK DATASETS IN PHOTOGRAMMETRY AND REMOTE SENSING. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-1-2022, 187–193.

Crystal-Ornelas, R., Varadharajan, C., O’Ryan, D., Beilsmith, K., Bond-Lamberty, B., Boye, K., Burrus, M., Cholia, S., Christianson, D. S., Crow, M., Damerow, J., Ely, K. S., Goldman, A. E., Heinz, S. L., Hendrix, V. C., Kakalia, Z., Mathes, K., O’Brien, F., Pennington, S. C., Robles, E., Rogers, A., Simmonds, M., Velliquette, T., Weisenhorn, P.,

[2] `https://figshare.com/` (last access: 11.9.23)

[4] `https://data.mendeley.com/` (last access: 11.9.23)

[7] `https://sedac.ciesin.columbia.edu/` (last access: 11.9.23)

[1] `https://datadryad.org/stash` (last access: 11.9.23)

[3] `https://dataverse.harvard.edu/` (last access: 11.9.23)

[5] `https://osf.io/` (last access: 11.9.23)

[8] `https://zenodo.org/` (last access: 11.9.23)

[6] `https://pangaea.de/` (last access: 11.9.23)

https://doi.org/10.5194/isprs-archives-XLVIII-1-W3-2023-17-2023 | © Author(s) 2023. CC BY 4.0 License.

Name	Repository Type	Metadata Schema	Free Limit (GB)	Paid Limit (GB)
Dryad (D)[1]	General	Dublin Core, DataCite	-	300
Figshare (F)[2]	General	Variety	20	5000
Harvard Dataverse (H)[3]	General	Unknown	1000	-
Mendeley Data (M)[4]	General	Dublin Core, Schema.org	10	-
Open Science Framework (O)[5]	General	Unknown	50	-
PANGAEA (P)[6]	Geospatial Specific	Schema.org	2000	-
SEDAC (S)[7]	Geospatial Specific	CSDGM	Unrestricted	-
Zenodo (Z)[8]	General	Unknown	-	50

Name	FAIR	DOI	Special Features	Access Restriction	Licensing	LiDAR datasets
D	✓	✓	-	-	CC0, CC-BY	93
F	✓	✓	-	Group Restriction	CC0, CC-BY, Others	883
H	✓	✓	-	-	CC0, Others	85
M	✓	✓	Long-term Preservation	Access On Request	CC0, CC-BY, Others	4791
O	✓	✓	Version Control	Access On Request	CC0, CC-BY, Others	1
P	✓	✓	Long-term Preservation	Password Protection	CCO, CC-BY, Others	547
S	✓	✓	-	-	CC-BY	0
Z	✓	✓	-	Embargo	CC0, CC-BY, Others	615

Table 1. Overview of exemplary selected repositories with the used comparison criteria. This selection is already implemented in our new repository selection tool. LiDAR is used as an example for ISPRS relation. The number of found datasets is from September 2023.

Welch, J. N., Whitenack, K., Agarwal, D. A., 2022. Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats. 9(1), 700.

Elouataoui, W., El Alaoui, I., Gahi, Y., 2022. Metadata quality in the era of big data and unstructured content. Y. Maleh, M. Alazab, N. Gherabi, L. Tawalbeh, A. A. Abd El-Latif (eds), *Advances in Information, Communication and Cybersecurity*, 1st ed. 2022 edn, Lecture Notes in Networks and Systems, 357, Springer International Publishing and Imprint Springer, 110–121.

EOSC Future, 2019. Your unified access to the european hub of research data,tools and services for innovation and education. https://eosc-portal.eu/ (12 June 2023).

Felden, J., Möller, L., Schindler, U., Huber, R., Schumacher, S., Koppe, R., Diepenbroek, M., Glöckner, F. O., 2023. PANGAEA - Data Publisher for Earth & Environmental Science. *Scientific data*, 10(1), 347.

figshare, 2011. Store, share, discover research. https://figshare.com/ (21 July 2023).

Gaia-X European Association for Data and Cloud AISBL, 2019. Gaia-x: A federated secure data infrastructure. https://gaia-x.eu/ (12 June 2023).

GFZ German Research Centre For Geosciences, Humboldt-Universität Zu Berlin, Germany Karlsruhe Institute Of Technology, Purdue University Libraries, Bertelmann, R., Buys, M., Cousijn, H., Dierolf, U., Elger, K., Fenner, M., Ferguson, L. M., Fritze, F., Fuchs, C., Goebelbecker, H.-J., Gundlach, J., Kindling, M., Kloska, G., Klump, J., Kramer, C., Manova, S., Pampel, H., Petras, V., Reuter, E., Rücknagel, J., van de Sandt, S., 2013. Registry of research data repositories.

ISPRS, 2022. Appendix 4: Isprs publication policy. https://www.isprs.org/documents/orangebook/app4.aspx (12 June 2023).

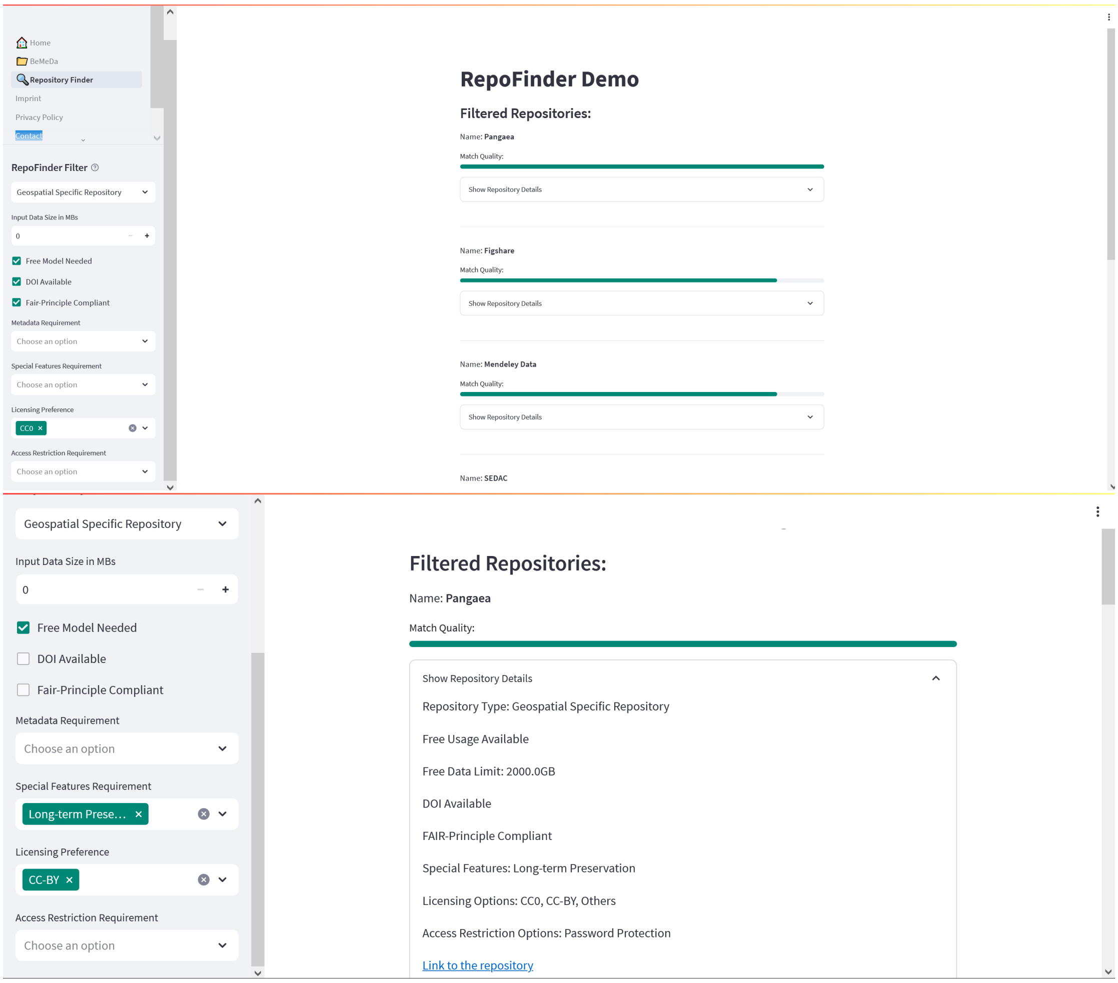
Ivánová, I., Brown, N., Fraser, R., Tengku, N., Rubinov, E., 2019. FAIR AND STANDARD ACCESS TO SPATIAL DATA AS THE MEANS FOR ACHIEVING SUSTAINABLE DEVELOPMENT GOALS. XLII-4/W20, 33–39.

Kindling, M., Strecker, D., 2022. Data Quality Assurance at Research Data Repositories. 21.

Kinkade, D., Shepherd, A., 2022. Geoscience data publication: Practices and perspectives on enabling the FAIR guiding principles. *Geoscience Data Journal*, 9(1), 177–186.

Nationale Forschungsdateninfrastruktur (NFDI) e.V., 2021. German national research data infrastructure. https://www.nfdi.de/ (12 June 2023).

Singh, J., 2011. FigShare. 2, 138–139. Journal Article Conflict of Interest: None declared Journal Article Conflict of Interest: None declared.



RepoFinder Filter

Geospatial Specific Repository

Input Data Size in MBs: 0

☑ Free Model Needed

☑ DOI Available

☑ Fair-Principle Compliant

Metadata Requirement: Choose an option

Special Features Requirement: Choose an option

Licensing Preference: CC0

Access Restriction Requirement: Choose an option

RepoFinder Demo

Filtered Repositories:

Name: **Pangaea**

Match Quality:

Show Repository Details

Name: **Figshare**

Match Quality:

Show Repository Details

Name: **Mendeley Data**

Match Quality:

Show Repository Details

Name: **SEDAC**

Geospatial Specific Repository

Input Data Size in MBs: 0

☑ Free Model Needed

□ DOI Available

□ Fair-Principle Compliant

Metadata Requirement: Choose an option

Special Features Requirement: Long-term Prese...

Licensing Preference: CC-BY

Access Restriction Requirement: Choose an option

Filtered Repositories:

Name: **Pangaea**

Match Quality:

Show Repository Details

Repository Type: Geospatial Specific Repository

Free Usage Available

Free Data Limit: 2000.0GB

DOI Available

FAIR-Principle Compliant

Special Features: Long-term Preservation

Licensing Options: CC0, CC-BY, Others

Access Restriction Options: Password Protection

Link to the repository

Figure 5. Screenshot of the Repository Finder tool.

Trisovic, A., Mika, K., Boyd, C., Feger, S., Crosas, M., 2021. Repository Approaches to Improving the Quality of Shared Data and Code. 6(2), 15. PII: data6020015.