Technical Report on ISPRS Scientific Initiative 2023

Publishing Dataset Guideline: gaps and trends in research data management in the ISPRS community (BenchPub)

Investigators

Dorota Iwaszczuk (Principal Investigator); Technical University of Darmstadt, Germany

Lina Budde (Co-Investigator); Technical University of Darmstadt, Germany

Timo Kullmann (Co-Investigator); Technical University of Darmstadt, Germany

Abstract

Data is one of the most important building blocks of modern research. Huge amounts of new data are generated every day. However, making these data treasures usable is a challenge. Well-structured data should be used for one's own research as well as made available to other researchers and interested parties in the spirit of open science and sustainability. However, the jungle of sometimes still unfamiliar terms from the field of open science and the wide variety of information and tools present a hurdle. As a result, the existing potential of data in the field of photogrammetry and remote sensing cannot yet be fully exploited.

Our scientific initiative has identified a number of needs to promote open science in ISPRS research areas. We analyzed exemplary the change and relationship of the past research topics by a keyword analysis. We compared data storage options and developed a web tool to rank them according to the individual requirements. In combination with our enhanced database about datasets, we are able to present an introduction and guideline for the publication and dissemination of datasets with regard to ISPRS topics.

1 Introduction

Knowledge about open science and its processes is not yet as widespread as it should be. As mentioned by Austin et al. (2017), the availability of guidelines for data publishing can improve research significantly. However, to find such guidelines suitable for the own research field can be challenging (Austin et al. 2017). For optimal dissemination of data sets, the topic of data publication is linked to data repositories, data articles and data journals (Austin et al. 2017). However, in our project we focus on the data repository aspect. Data repositories offer key functionalities such as persistent identifier, standardized metadata and a certain amount of curation (Austin et al. 2017). Even if it is not part of our instructions, the publication of datasets in combination with an explicit data article should be aimed for. In the following sections we present our results of our ISPRS initiative. Based on a keyword analysis (Section 2.1), we compare different data repositories (Section 2.2) and present a recommendation tool based on this comparison (Section 2.3). Finally, we offer the new tool in combination with our existing but now updated BeMeDa tool, which can be used to find existing datasets that are not necessarily in data repositories (Section 2.4).

2 Results

2.1 Keyword Analysis

To gain a comprehensive understanding of the ISPRS domain and the papers published within it, our project commenced with a thorough keyword analysis. This analysis encompassed all papers published in the context of the ISPRS congresses held since 2012, specifically covering the years 2012, 2016, 2020, 2021, and 2022 at the time of writing. In total, we scrutinized over 1100 papers, thoroughly collecting their keywords.

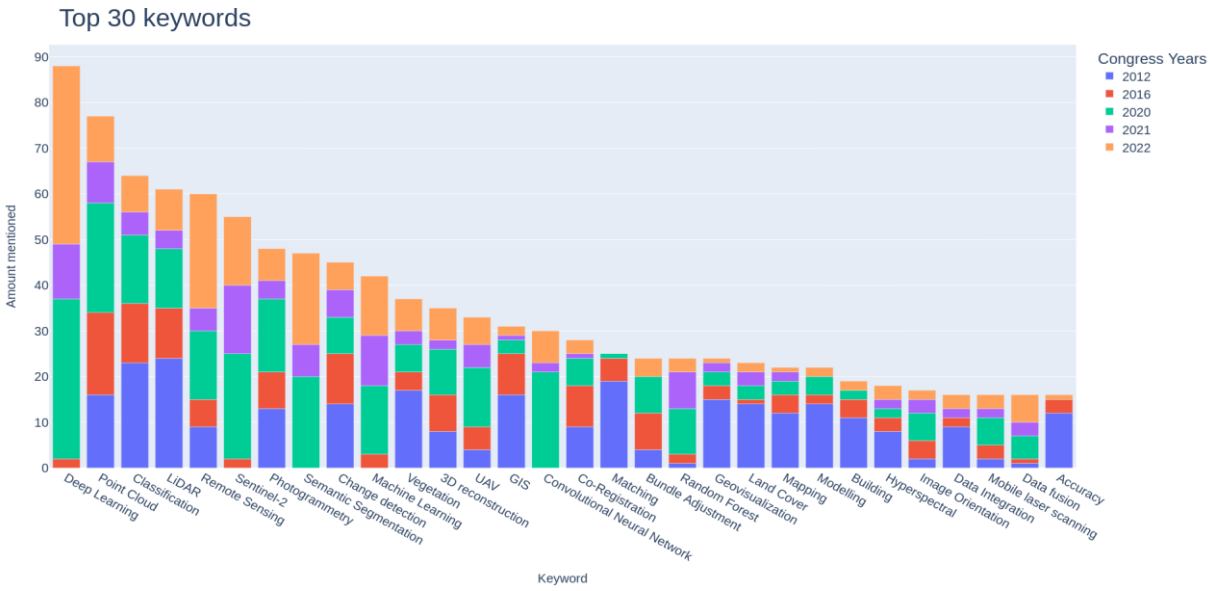

Figure 1: The top 30 keywords extracted from the paper published in the ISPRS Annals. Sorted by total count in descending order.

To visualize the gathered data in a way that is humanly interpretable, we have chosen two different methods. First, the frequency of all collected keywords in visualized in a bar graph for a user-defined number of top keywords. Figure 1 shows a bar graph with the top 30 keywords by overall mentions. Figure 2 shows the same keywords, but sorted by their mentions in the year 2022, highlighting which keywords were most important in recent publications. Overall, this visualization provides a reliable estimate of the most significant keywords and, consequently, the prevailing topics within the community. Moreover, it offers insights into how the importance of specific topics has evolved over time. Figure 2 especially shows the increasing importance and usage of machine learning. The third visualization aims to provide a better understanding of the interconnections between keywords (Figure 3). For this, all keywords that appear together in a paper are registered as pairs. Subsequently, we constructed a force-directed graph using these pairs. In this graph, each node represents a single keyword and the size of the node shows the overall amount of mentions for this specific keyword. The connections between nodes are created from the pair connections. The larger a connection is, the more often this keyword combination was found.

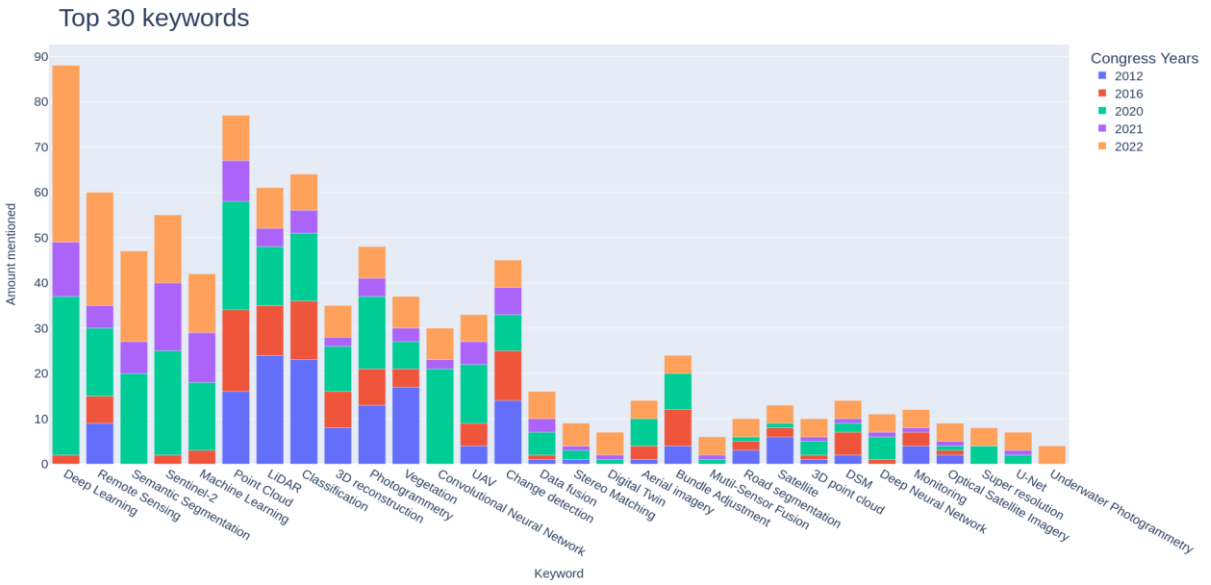

Figure 2: In contrast to Figure 1, the keywords are sorted by the total count of the congress 2022 in descending order.

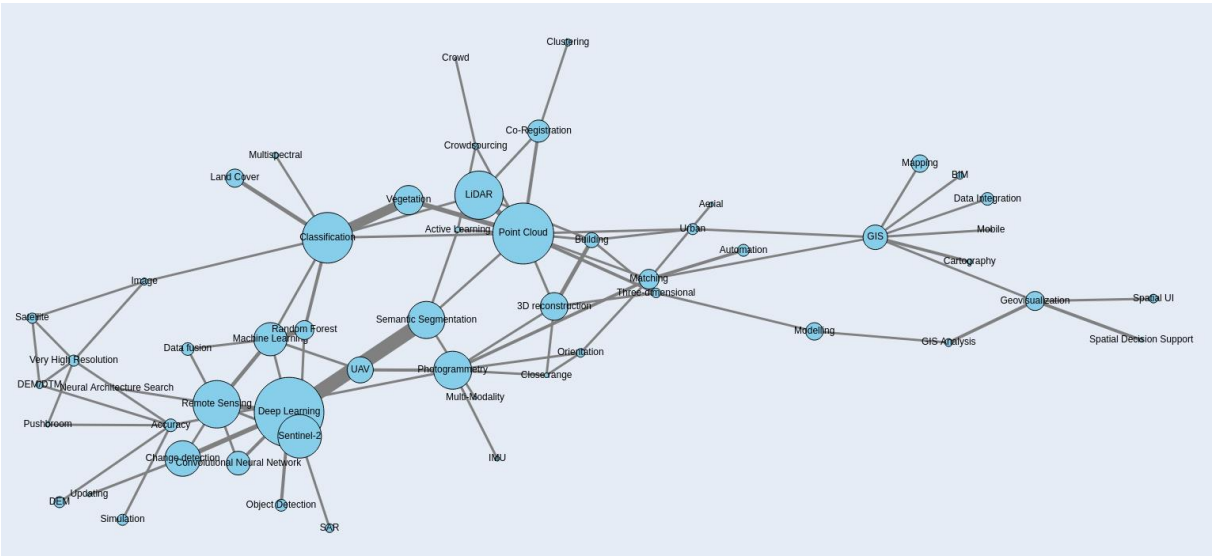

Figure 3: Screenshot of the relationships between the extracted keywords in a force-directed graph. The size of the circle represents the frequency of the single keyword, whereas the line thickness represents the frequency of the keyword pair. This graph visualization is also available on our website where it can be zoomed in (https://benchmedata.org/).

Through this analysis, it becomes evident that the current research focus is shifting towards machine learning-related topics. This shift implies an increasing demand for well-documented and easily accessible data, given that machine learning is reliant on data. For more information, you can look in Budde et al. (2023b).

2.2 Evaluation of Repositories

The goal of creating a repository recommendation guideline requires the search for adequate repositories that are suitable as recommendations. This entails establishing a list of requirements and prioritizing them. We evaluate both well-known and geoscience-specific data repositories against these criteria, either selecting or discarding them accordingly. Using this approach, we initially compiled a selection of ten repositories that meet our requirements, for example Mendeley Data (Swab 2016), Figshare (Thelwall and Kousha 2016) and PANGAEA (Felden et al. 2023).

In regards to our requirements, we focused on the following attributes. Initially, we categorized repositories based on their primary data domain. Subsequently, we documented all supported metadata schemes to ensure compatibility with common standards. Key considerations also include the availability of a digital object identifier (DOI), adherence to the FAIR principle (findable, accessible, interoperable, reusable) and the provision of free usage. Regarding the cost, we mainly focus on the available data limit for the free tier, as well as the paid tier, if there is one and not on the specific pricing. In addition, we gathered information on access restrictions, such as embargoes, supported licensing options, and other unique features, which can vary greatly between repositories. Finally, we use the APIs of the individual repositories to query for specific keywords related to the ISPRS domain. Compared to re3data which offers a database about existing repositories (re3data.org 2024), we enable a strong link to domain-specific knowledge. A list of the chosen repositories and their attributes can be found in our "Dataset Publication Guideline" (see Appendix) as well as on our website https://benchmedata.org/ (Budde et al. 2024).

2.3 Recommendations for dataset publications

Based on the evaluation of the suitable data repositories, the "Repository Finder" tool (Budde et al. 2024) developed as part of the project provides a ranking of possible publication options using data repositories. The web application consists of two main elements: user requirements and the evaluated data repositories. An overview of the components is visualized in Figure 4. With a selection interface, the user can specify requirements of the repository for their specific dataset. Besides basic requirements, the user can also select the most fitting keywords from a given list, that describe the dataset best. For the selected keywords, the number of similar datasets is listed for each repository. Since each keyword is handled separately, some results may count multiple times. This is caused by the static, and therefore responsive, implementation, which could be replaced with a dynamic one in the future. For the purpose of giving a recommendation, this approach is sufficient for now. The static keyword data was parsed at an earlier date and includes all keywords found in the ISPRS domain with their respective search results for some of the given repositories. Only repositories, where an automation of this process was not possible, are not included. Each time a requirement is changed, the properties of the individual repositories are compared with the selection and a corresponding ranking is generated. The ranked data repositories can be seen as recommendation list where to publish the dataset. The top listed repository fits the user selection best. However, equal ranked repositories can occur. The quality of the match is also visualized by a bar, the fuller it is, the better the fit. Within a dropdown menu for each repository, further information as well as the link to the repository is shown.

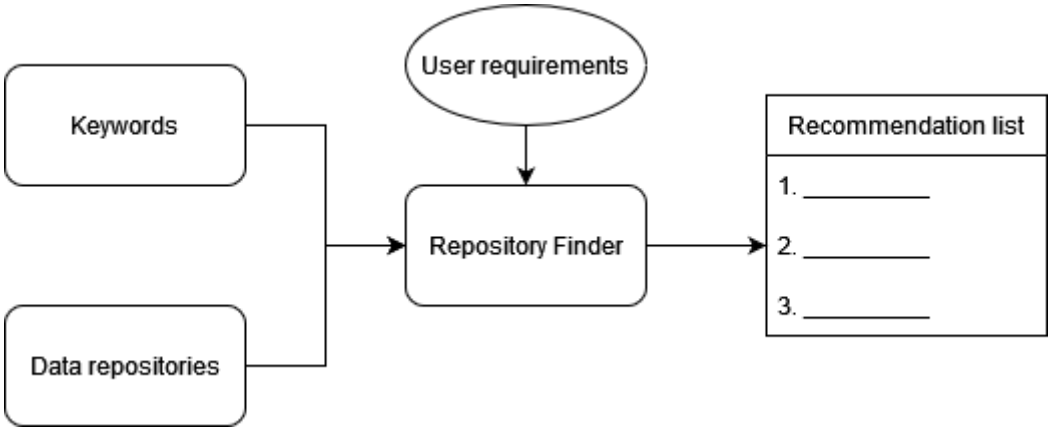

Figure 4: The repository finder compares the user requirements with the data repositories and generates a recommendation list of possible data repositories for the data publication. In addition, keywords can be used to check how many datasets are already published in the repositories.

In addition to the Repository Finder tool, the separate "Dataset Publication Guideline" (see Appendix) contains some basic information with regard to research data management as well as the information on which the Repository Finder is based.

2.4 Improvements of BeMeDa Database

The database we had already developed for finding benchmark data sets in the field of remote sensing and photogrammetry was also updated as part of the initiative (Budde et al., 2022). Thanks to a new structure of the BeMeDa website, the newly created recommendations for the dataset publication and the BeMeDa database itself can be accessed under one URL.

The updates to BeMeDa include the expansion of metadata adding further geographical information and more datasets. From an implementation perspective, the search results are now automatically adjusted based on the filter results. It is also possible to reset the filters and a slider makes it easier to select a time period from which the desired data should originate. Finally, the list of the found datasets can be downloaded. Thus, the metadata can be further used by the searcher. (Budde et al. 2023c)

3 Conclusion

With our website (Budde et al. 2024) we offer two tools specialized for the photogrammetry and remote sensing domain. First, we created a search tool for finding already published data with and without data repositories, based on a multitude of user specific requirements. Secondly, we also introduced a guideline tool for the selection of data repositories for dataset publication that meet certain requirements, which are selected by the user and fit their needs. Additionally, we offer a form for users to contribute to our BeMeDa database and to the selection of data repositories. Moreover, the "Dataset Publication Guideline" contains some basic background information with regard to open science and research data management as well as an overview about the data repositories included in our web application and finally an instruction how the web application can be used. Overall this project aims to strengthen the toolkit and knowledge regarding data publication in the photogrammetry and remote sensing domain, which ultimately improves data availability and quality for all researchers and other projects.

We already successfully presented parts of our contributions at the national research data infrastructures for engineers conference (Budde et al. 2023a) and at the Geobench Workshop in Krakow (Budde et al. 2023b; Budde et al. 2023c).

4 Literature

Austin, C. C.; Bloom, T.; Dallmeier-Tiessen, S.; Khodiyar, V. K.; Murphy, F.; Nurnberger, A. et al. (2017): Key components of data publishing: using current best practices to develop a reference model for data publishing. In Int J Digit Libr 18 (2), pp. 77–92. DOI: 10.1007/s00799-016-0178-2.

Budde, L. E.; Kullmann, T.; Iwaszczuk, D. (2024): BeMeDa database and Repository Finder. https://benchmedata.org/; last accessed: 30.01.2024.

Budde, L. E.; Kullmann, T.; Iwaszczuk, D. (2023a): Experience with benchmark datasets in the heterogeneous data landscape of photogrammetry and remote sensing. In Stefanie Roski, Regine Gerike (Eds.): Proceedings of the NFDI4Ing Conference 2023 - Innovation in Research Data Management: Bridging the gaps between disciplines and opening new perspectives for research in engineering science. Zenodo, pp. 12–13.

Budde, L. E.; Kullmann, T.; Iwaszczuk, D. (2023b): On the Development of a Dataset Publication Guideline: Data Repositories and Keyword Analysis in ISPRS Domain. In Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLVIII-1/W3-2023, pp. 17–23. DOI: 10.5194/isprs-archives-XLVIII-1-W3-2023-17-2023.

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. In ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. V-1-2022, pp. 187–193. DOI: 10.5194/isprs-annals-V-1-2022-187-2022.

Budde, L. E.; Schmidt, J.; Kullmann, T.; Iwaszczuk, D. (2023c): Current Status of the Benchmark Database BeMeDa. In Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLVIII-1/W3-2023, pp. 25–30. DOI: 10.5194/isprs-archives-XLVIII-1-W3-2023-25-2023.

Felden, J.; Möller, L.; Schindler, U.; Huber, R.; Schumacher, S.; Koppe, R. et al. (2023): PANGAEA - Data Publisher for Earth & Environmental Science. In Scientific data 10 (1), p. 347. DOI: 10.1038/s41597-023-02269-x.

re3data.org (2024): re3data.org (2024): Registry of Research Data Repositories. DOI: 10.17616/R3D; last access: 29.01.2024

Swab, M. (2016): Mendeley Data. In J Can Health Libr Assoc 37 (3). DOI: 10.5596/c16-022.

Thelwall, M.; Kousha, K. (2016): Figshare: a universal repository for academic resource sharing? In Online Information Review 40 (3), pp. 333–346. DOI: 10.1108/OIR-06-2015-0190.

Appendix: Dataset Publication Guideline

This guide provides a brief introduction to some terms and background information in the first part, an overview of the considered data repositories in the second part, and lastly instructions on how to use our Repository Finder in the field of remote sensing and photogrammetry. While part 1 is also useful at the beginning of a research project, parts 2 and 3 assume that data is already available and ready to publish.

PART 1: Terms & Background

If you are unfamiliar with the concept of research data management (RDM), this short introduction should help. In this section, some basic terms in the cosmos of open science and RDM are briefly explained.

1. Metadata

   Metadata is a description of the data. It contains all relevant information about the creator and the data creation process, as well as how the data is structured and stored. Some important standards for metadata include Schema.org[1], DublinCore[2] and DataCite Schema[3].

2. FAIR principle

   One of the basic concepts in the context of RDM is the so-called FAIR principle. The acronym stands for Findable, Accessible, Interoperable and Reusable[4]. The data should be as open as possible. However, restrictions can be implemented to ensure data privacy[5]. If you want to check how FAIR your data is, you can use the SATIFYD tool[6].

3. Licenses

   Licenses are important when working with data. The Creative Commons licenses[7] are widely used. However, the specific constraints of each category may not be as well-known. If you have no other restrictions, the CC-BY licenses might be suitable, where the data can be shared, remixed, adapted and further developed with your credit. However, the CC License Chooser tool on the provider's website will help you choose an appropriate license.

4. Data management plan

   Ideally, such a plan should be created at the beginning of your research project. It is a documentation strategy that allows you to write down your information step by step throughout your research process. Among other things, file types, licenses, project users and data volume are documented as metadata. Then you have already collected important metadata about your research and your data, which can be used in a possible future publication. For further information, you can visit the MIT libraries[8] website.

5. Data Repository

   Data repositories offer the option of long-term data storage. There are three types of repositories: institutional, general and subject-specific. Beside the long-term preservation, repositories provide

---

[1] Schema.org. https://schema.org/. (last access: 09.01.24)

[2] Dublin Core Metadata Initiative. https://dublincore.org/ (last access: 06.02.24)

[3] DataCite Schema. https://schema.datacite.org/ (last access: 06.02.24)

[4] Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J. J. et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. Scientific data 3, 1–9. DOI:10.1038/sdata.2016.18.

[5] Kinkade, D.; Shepherd, A., 2022. Geoscience data publication: Practices and perspectives on enabling the FAIR guiding principles. Geoscience Data Journal

[6] https://satifyd.dans.knaw.nl/ (last access: 06.02.24)

[7] https://creativecommons.org/ (last access: 06.02.24)

[8] https://libraries.mit.edu/data-management/plan/ (last access: 15.01.2024)

a persistent identifier. Thus, the data is easily cited and shared. In addition, the necessary metadata are important for finding and reusing the dataset. As a result, repositories can improve the data quality, especially if they include some data review processes[9]. In this way, repositories can contribute to the FAIR principle.

In general, a detailed documentation of each component of the research process enables a higher transparency and reusability. Proper RDM can considerably simplify working with research data and contributes to our path to open science. National or institutional organizations often provide an infrastructure and various tools to simplify the steps of a research documentation process and RDM, such as platforms to create and share data management plans. We often work with large amounts of data in different file formats in the field of photogrammetry and remote sensing. Thus, starting your documentation with a data management plan before and during data collection can save time when you want to publish your data later on. Regardless of whether the data is made public, research data should be stored long-term in a structured manner to ensure traceability for future work.

Further Reading:

Van den Eynden, V.; Corti, L.; Woollard, M.; Bishop, L.; Horton,L. (2011): Managing and sharing data: Best practice for researchers. UK Data Archive, ISBN: 1-904059-78-3, https://dam.ukdataservice.ac.uk/media/622417/managingsharing.pdf

PART 2: Data Repositories

Data repositories form an important infrastructure for the publication of data. However, the large number of existing repositories makes it difficult to find the most appropriate one. To present alternatives to institutional repositories for greater reach, we have currently selected 10 data repositories and are comparing them based on 6 main attributes (Table 1). In addition, the selected repositories are evaluated for already existing datasets with photogrammetry and remote sensing specific keywords. However, this evaluation is only a snapshot (Table 2). This information about the different data repositories is used for our Repository Finder (see PART 3: Repository Finder). It should be noted that the geospatial-specific datasets listed are developed for the geosciences and are therefore often oriented towards specific applications, e.g. maps for socioeconomic research in SEDAC[9].

Table 1: Overview of the selected data repositories. All of them have a DOI and contribute to the FAIR principle

Name	Category	Supported Metadata Types	Data Size Limits	Special Features	Access Restriction	Licensing Options
Figshare[10]	General Repository	JSON-LD, Schema.org, Dublin Core, DataCite	20GB free, 500GB paid	N/A	Group Restriction	CC0, CC-BY, Others

[9] Kindling, M.; Strecker, D. (2022): Data Quality Assurance at Research Data Repositories. In *Data Science Journal* 21, Article 18. DOI: 10.5334/dsj-2022-018.

[10] https://figshare.com/ (last access: 05.02.24)

Mendeley Data[11]	General Repository	Dublin Core, Schema.org	10GB free	Long-term Preservation	Access On Request	CC0, CC-BY, Others
SEDAC[12]	Geospatial Specific Repository	CSDGM	N/A	N/A	N/A	CC-BY
Dryad[13]	General Repository	Dublin Core, DataCite	300GB paid	N/A	N/A	CC0, CC-BY
Harvard Dataverse[14]	General Repository	N/A	1TB free	N/A	N/A	CC0, Others
Open Science Framework[15]	General Repository	N/A	50GB free	Version Control	Access On Request	CC0, CC-BY, Others
Zenodo[16]	General Repository	N/A	50GB paid	N/A	Embargo	CC0, CC-BY, Others
GFZ Data Services[17]	Geospatial Specific Repository	N/A	N/A	N/A	Embargo	CC0, CC-BY, Others
Pangaea[18]	Geospatial Specific Repository	N/A	10GB free	Long-term Preservation	Password Protection	CC0, CC-BY, Others
GeoLeo[19]	Geospatial Specific Repository	N/A	N/A	Long-term Preservation	Access On Request	CC0, CC-BY, Others

[11] https://data.mendeley.com/ (last access: 05.02.24)
[12] https://sedac.ciesin.columbia.edu/ (last access: 05.02.24)
[13] https://datadryad.org/stash (last access: 05.02.24)
[14] https://dataverse.harvard.edu/ (last access: 05.02.24)
[15] https://osf.io/ (last access: 05.02.24)
[16] https://zenodo.org/ (last access: 05.02.24)
[17] https://bib.telegrafenberg.de/dataservices (last access: 08.02.24)
[18] https://pangaea.de/ (last access: 05.02.24)
[19] https://geo-leo.de/ (last access: 05.02.24)

Table 2: Snapshot of the count of existing datasets in the repositories for the keyword "LiDAR".

Name	Dataset counts for keyword "LiDAR" (status: 09.02.2024)
Figshare	3163
Mendeley Data	8487
SEDAC	0
Dryad	97
Harvard Dataverse	144
Open Science Framework	1
Zenodo	1085
GFZ Data Services	10
Pangaea	574
GeoLeo	12

PART 3: Repository Finder

Even though the tool re3data[20] provides a tool for searching for repositories, it is still difficult to find suitable repositories in the field of remote sensing and photogrammetry[21]. For the beginning we have selected a few repositories that we believe can be used in the ISPRS domain (see PART 2: Data Repositories). With our repository selection we build up a web application called “Repository Finder”. Our tool can be used to generate a recommendation list for suitable data repositories based on our database. However, further repositories that meet our criteria can be added in the future. In this section we will show you how to use our Repository Finder.

Step 1: Visit our website https://benchmedata.org/ and go to the Repository Finder (may take a while to load due to downtime).

Step 2: Select the filters that match your requirements for your data, such as data volume or licensing options. Use the drop-down menu to see more details about each repository.

[20] GFZ German Research Centre For Geosciences; Humboldt-Universität Zu Berlin; Germany Karlsruhe Institute Of Technology; Purdue University Libraries; Bertelmann, Roland; Buys, Matt et al. (2013): Registry of Research Data Repositories. Available online at https://www.re3data.org/.

[21] Budde, L. E.; Kullmann, T.; Iwaszczuk, D. (2023b): On the Development of a Dataset Publication Guideline: Data Repositories and Keyword Analysis in ISPRS Domain. In *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLVIII-1/W3-2023, pp. 17–23. DOI: 10.5194/isprs-archives-XLVIII-1-W3-2023-17-2023.

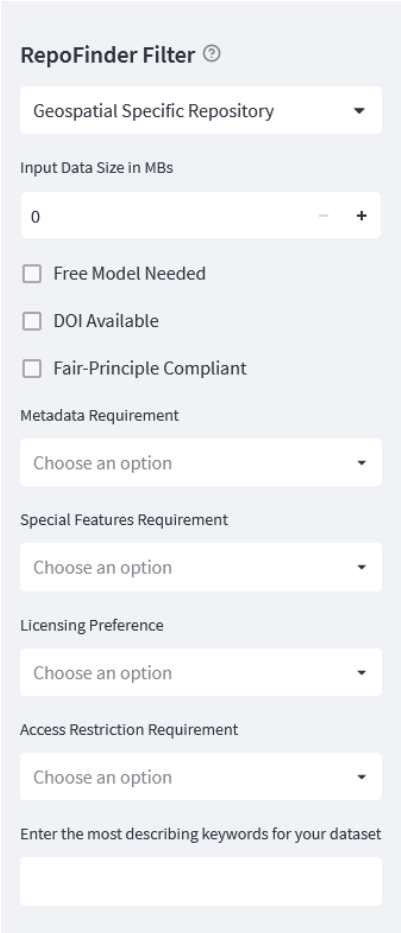

Figure A: Screenshot of the filter menu to choose specific properties of the preferred data repositories.

Step 3: View the best matching repositories. Once you decided for one, the link provided will take you to the selected repository. There you will find more information about the submission process. In general, an account for the specific platform is needed to upload data.