The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W3-2023
2nd GEOBENCH Workshop on Evaluation and BENCHmarking of Sensors, Systems and GEOspatial Data
in Photogrammetry and Remote Sensing, 23–24 October 2023, Krakow, Poland

CURRENT STATUS OF THE BENCHMARK DATABASE BEMEDA

L. E. Budde, J. Schmidt, T. Kullmann, D. Iwaszczuk*

Technical University of Darmstadt, Dept. of Civil and Environmental Engineering Sciences,
Remote Sensing and Image Analysis, Darmstadt, Germany
(lina.budde, jakob.schmidt1, dorota.iwaszczuk)@tu-darmstadt.de, timo.kullmann@stud.tu-darmstadt.de

Commission I

KEY WORDS: FAIR principle, Benchmark, Database, Dataset, Metadata.

ABSTRACT:

Open science is an important attribute for developing new approaches. Especially, the data component plays a significant role. The FAIR principle provides a good orientation towards open data. One part of FAIR is findability. Thus, domain specific dataset search platforms were developed: the Earth Observation Database and our Benchmark Metadata Database (BeMeDa). In addition to the search itself, the datasets found by this platforms can be compared with each other with regard to their interoperability. We compare these two platforms and present an update of our platform BeMeDa. This update includes additional location information about the datasets and a new frontend design with improved usability. We rely on user feedback for further improvements and enhancements.

1. INTRODUCTION

In today's information age, data is an important fundamental. For successful and sustainable use of data, metadata and meta-data management are crucial (Elouataoui et al., 2022). This supports not only the human users, but also the machine (Wilkinson et al., 2016). Besides a better understanding of the respective dataset, metadata supports the FAIR principle i.e. metadata helps regarding to findability, accessibility, interoperability, and reusability. In addition, the quality of metadata have also an impact on the data quality with regard to their reusability (Kindling and Strecker, 2022). The evaluation of metadata quality according to the FAIR principle is an important task. Elouataoui et al. (2022) describes different metadata quality factors such as completeness and usefulness. To ensure such factors improves the trust into the data and promotes the use of the datasets (Trisovic et al., 2021).

The introduction of a domain-specific search tool makes it easy to find existing datasets, especially for beginners in research. Compared to general search engines such as Google dataset search (Brickley et al., 2019), domain-specific tools allows specialized filter options. In addition, pure metadata databases are lightweight with respect to the stored data size, compared to data repositories and databases which stores the dataset itself. For example our metadata database storage is currently less then 1 GB. In contrast, a single image dataset such as the Potsdam dataset (Markus Gerke et al., 2014) has a size of about 12 GB.

For the dissemination of benchmark datasets created by the ISPRS community, the FAIR principle must be given greater consideration, especially the metadata aspect. Due to missing standards for data and metadata, dataset are hard to find and interoperable and thus insufficient for reuse (Crystal-Ornelas et al., 2022). To improve such limitations in the findability aspect and to create a base for the interoperability and reusability, we developed the Benchmark Metadata Database (BeMeDa) (Budde et al., 2022).

* Corresponding author

However, the database contains some limitations that are noted in discussion and outlook of Budde et al. (2022). Therefore, in this contribution we present a detailed comparison with the Earth Observation Database (EOD) platform (Schmitt et al., 2022)[1]. In addition, updates in the search functionalities and metadata of BeMeDa are presented.

The paper is structured as follows. Section 2 present the background of this database project. In section 3, we discuss the commonalities and differences between EOD platform and BeMeDa. The description of the innovations in our database are included in section 4. A new possibility for participation is introduced in section 5. Section 6 concludes the paper with some future work.

2. BACKGROUND

Started 2021, the BeMeDa database was introduced during the ISPRS congress 2022. The aim of BeMeDa is to enable a simple search for datasets published with relation to the ISPRS and corresponding topics. One of the drawbacks of conventional search engines is their low domain-specificity. So, it is difficult to narrow down the results to relevant datasets. In addition, due to self-hosted datasets and, more important, due to missing metadata the comparison of different datasets is time consuming. Which comes on top of the time consuming part of harmonization of different data sources (Mons et al., 2017).

However, by developing a metadata schema based on our domain terminology, we collected the respective information for each dataset (Budde et al., 2022). In addition to technical terms, metadata from the schema.org metadata schema (Data and Datasets - schema.org, 2021) is also used. Thus, through such a standardized procedure, direct comparisons between datasets can be made. In total, we used 12 attributes for the datasets, with the "paper" attribute is consisting of 4 further attributes.

[1] Earth Observation Database, 2022. `https://eod-grss-ieee.com/home` (last access July 2023).

Table 1. The number in brackets represented the number of predefined selection options to filer datasets in BeMeDa and EOD database.

BeMeDa (ours)	EOD
Sensors (8)	Sensor (8 + other)
Tasks (17)	Task (11 + other)
Location (new)	Location (map provided)
Platform (9)	
Configuration (9)	
Application (11)	
Dimensionality (5)	

The subject-specific attributes are: task, sensor, acquisition configuration, acquisition platform, application environment and dimension. (Budde et al., 2022)

Due to the use of a NoSQL document database, heterogeneous data can be easily stored and efficiently queried (Meier and Kaufmann, 2019). Especially, we used the advantage that NoSQL database can contain empty fields and the structure of each document is flexible for new attributes and values (Kaur and Rani, 2013).

This results in a publicly available search tool[2]. Here, the previously mentioned subject-specific attributes from the metadata schema are used for filtering in the database by selecting the appropriate filter value.

3. COMPARISON OF METADATA DATABASES

The EOD platform was developed for the IEEE GRSS community (Schmitt et al., 2022). Thus, this platform has a high degree of overlap with topics and data in the ISPRS community. To the best of our knowledge, the EOD platform and our BeMeDa are the only databases that exclusively cover the fields of remote sensing and photogrammetry and only contain metadata and not the data itself. However, to compare BeMeDa with the EOD platform, the different filter options are considered. As shown in Table 1, BeMeDa provides more filter options.

Nevertheless, EOD enables map and location based search while BeMeDa only displays the location since the new version (see section 4). In both databases, datasets can be selected by defining sensor type and task. However, in BeMeDa, additional sensors are considered which are often used for referencing such as GNSS. In contrast, the predefined sensor types in EOD are mixed up with the configuration option in BeMeDa e.g. multispectral.

For each dataset, EOD provides information about the dataset name, location, sensor, dataset size, date, task, link, number of views and a short description. The compare functionality in EOD simplifies the search of interoperable datasets. BeMeDa presents name, sensor, year, task, link and also a short description. In addition, BeMeDa shows the attributes of the other filters such as application and the corresponding paper. In particular, this paper link can increase confidence in the dataset (Trisovic et al., 2021). In total, BeMeDa contains 61 and EOD 133 datasets (July 2023 status). While the datasets on EOD are only related to remote sensing, BeMeDa also offers datasets in the field of photogrammetry.

Overall, both databases are useful for dataset search. They complement each other. To compare the distribution of the dataset by their location, this information is added to the updated version of BeMeDa (section 4). However, we provided detailed information about the implementation in Budde et al. (2022) compared to Schmitt et al. (2022). In addition to the increased transparency, this allows others to implement their own databases.

One of the few datasets that is included in both databases is PASTIS (Garnot et al., 2021). The corresponding database entry is shown in Figure 1. In contrast, the BeMeDa entry with the table view cannot display all metadata information well and thus only a part of the metadata is viewed. However, with the machine-readable json or csv file, the full information can be downloaded and reused. Preference is given to a view that can display all the information about the dataset as well as a compact comparison showing the differences and similarities between different datasets.

In detail, the entries differ in the description, sensors and task. However, the same information are delivered. The less detailed description in BeMeDa is compensated by the additional attributes, such as application, configuration, platform and dimensionality. Both platforms provide additional information, such as an image example, likes, views and sizes in EOD and paper reference in BeMeDa. The location is displayed via map in BeMeDa, EOD named the location “France”.

4. BEMEDA DATABASE UPDATES

To accommodate the requirements of additional functionalities in the updated version of BeMeDa, the website was redesigned. Not only the frontend design was changed, but also the implementation. While the MongoDB database is still in use (Budde et al., 2022), for the web development the python package streamlit is now used (Khorasani et al., 2022). As displayed in Figure 3, the filter menu moves into a sidebar which is collapsible. Due to the metadata concept from Budde et al. (2022) which can easily extended by new attributes, a location attribute now supplements the dataset information (Figure 2). Thus, the database overview is enhanced by a map view in which marker represent inserted datasets. If a dataset contains data belonging to different locations, each location is marked. By clicking on the marker, the name of the dataset to which it belongs is presented. Such a map is also displayed with the search results.

With the new filter menu, the website automatically updates the search results when changing the input values. Thus, an extra search button becomes obsolete. Instead, two new buttons are inserted. The one button with the label ”Get all dataset” outputs all datasets present in the database. The second button ”Reset all filter” removes all selected filter elements at once. In addition, for each filter category a help information is added. Furthermore, selecting a time period via a slider allows intuitive customization of the search.

The resulting datasets based on the used filter are now displayed in a compact table view. The table can also be displayed in full screen mode. In addition, each column of the table can be sorted alphabetically. In addition, all available information for the found dataset such as the description and the coordinates can be downloaded as csv or json file (Figure 4). This allows, to reuse the metadata information automatically.

5. PARTICIPATION

The updated BeMeDa version includes a new possibility to participate. The database search is extended by a submission

[2] `https://benchmedata.org/` (last access July 2023)

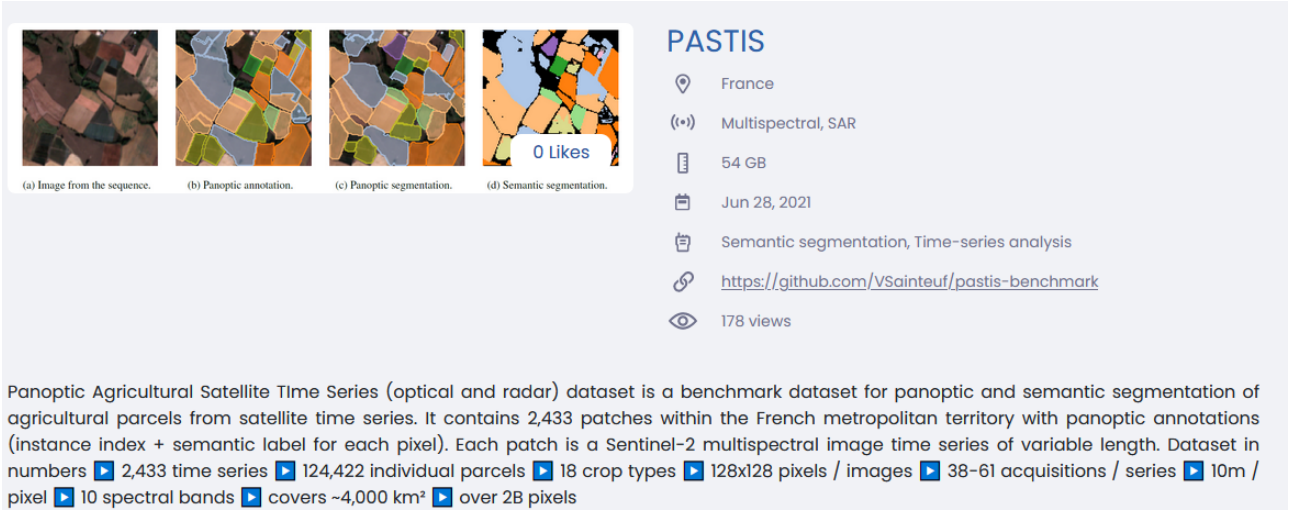

Figure 1. Screenshot of the PASTIS entry from the EOD platform (Schmitt et al., 2022) (top) and the table from BeMeDa (bottom)

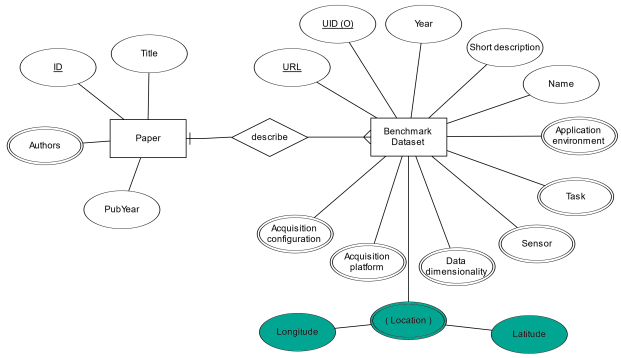

Figure 2. Extension of the ER diagram adapted from Budde et al. (2022) by the location attribute which consist of longitude and latitude coordinates.

form. This allows researchers to suggest data sets to be added to the database. Figure 5 displays the empty form. The desired metadata information is either entered as free text or applicable values are selected. The latter are multiple selection fields, i.e. several terms can be selected by pressing the shift key. When the send button is pressed, the request is submitted to the BeMeDa team. After a check of the data, the dataset will be added to the database.

6. CONCLUSION

Both databases, BeMeDa and EOD, allow an effective search of benchmark datasets compared to more general dataset search platforms such as Google Dataset search (Brickley et al., 2019). While EOD is very easy to use, BeMeDa offers more domain specific filtering and further statistics and information about the database itself.

The creation of a database such as BeMeDa needs continuous development. Therefore, user feedback is important for further improvements. However, some feedback from users has already been implemented. Thus the number of colors is reduced. More important, the unintuitive use of the previous search button could removed. Thus, the results are updated automatically by filter changes. The use of a flexible and easy-to-use implementation enables long-term maintenance and further adaptions. Furthermore, the submission form simplifies the participation of BeMeDa and hopefully will extend the amount of datasets more rapidly. Although so far the focus has been on the human user, the ability to download the search results into machine-readable formats such as csv or json also improves the findability aspect of the FAIR principle.

However, lack of unique identifiers still limits the long-term availability of datasets, e.g., the ISPRS indoor modeling benchmark is no longer accessible due to restructuring in ISPRS (Khoshelham et al., 2017). The use of persistent identifiers improves the quality of the metadata and thus the reusability and reuse of the dataset (Trisovic et al., 2021; Kindling and Strecker, 2022). This can also be considered as a quality factor of the data. Thus, for data quality assessment such factors can be used too.

In the future, we are going to link BeMeDa to the ISPRS scientific initiative “Publishing dataset guideline: gaps and trends in research data management in the ISPRS community” (Budde et al., 2023). Furthermore, the importance of databases for benchmarking is also visible in other scientific initiatives, for example in “NAUTILUS uNder And throUgh waTer datasets for geospatIaL stUdieS”[3].

References

Brickley, D., Burgess, M., Noy, N., 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. L. Liu, R. White (eds), *The World Wide Web Conference*, ACM, New York, NY, USA, 1365–1375. https://doi.org/10.1145/3308558.3313685.

Budde, L. E., Kullmann, T., Iwaszczuk, D., 2023. On the development of a dataset publication guideline: data repositories and keyword analysis in ISPRS domain. *ISPRS Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-1/W3-2023.

[3] https://nautilus-isprs.fbk.eu/ (last access September 2023)

Figure 3. The database statistics are extended by a world map with all available location information.

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., Welch, J. N., Whitenack, K., Agarwal, D. A., 2022. Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats. *Scientific data*, 9(1), 700. 10.1038/s41597-022-01606-w.

Data and Datasets - schema.org, 2021. https://schema.org/docs/data-and-datasets.html (accessed 24.11.21).

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. 10.1007/978-3-030-91738-8_11.

Garnot, F., Sainte, V., Landrieu, L., 2021. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4852–4861. 10.1109/ICCV48922.2021.00483.

Kaur, K., Rani, R., 2013. Modeling and querying data in nosql databases. *2013 IEEE International Conference on Big Data*, 1–7. https://doi.org/10.1109/BigData.2013.6691765.

Khorasani, M., Abdou, M., Hernández Fernández, J., 2022. Streamlit basics. M. Khorasani, M. Abdou, J. Hernández Fernández (eds), *Web Application Development with Streamlit*, 1st ed. 2022 edn, Apress and Imprint Apress, 31–62. 10.1007/978-1-4842-8111-6_2.

Khoshelham, K., Díaz Vilariño, L., Peter, M., Kang, Z., Acharya, D., 2017. The ISPRS Benchmark on Indoor Modelling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W7, 367-372.

Kindling, M., Strecker, D., 2022. Data Quality Assurance at Research Data Repositories. *Data Science Journal*, 21. 10.5334/dsj-2022-018.

Markus Gerke, Franz Rottensteiner, Jan D Wegner, Gunho Sohn, 2014. ISPRS semantic labeling contest. 10.13140/2.1.3570.9445.

Meier, A., Kaufmann, M., 2019. *NoSQL Databases*. Springer Fachmedien Wiesbaden, Wiesbaden, 201–218. https://doi.org/10.1007/978-3-658-24549-8_7.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., Wilkinson, M. D., 2017. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, 37(1), 49–56. 10.3233/ISU-170824.

Schmitt, M., Ghamisi, P., Yokoya, N., Hansch, R., 2022. EOD: The IEEE GRSS Earth Observation Database. *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 5365-5368.

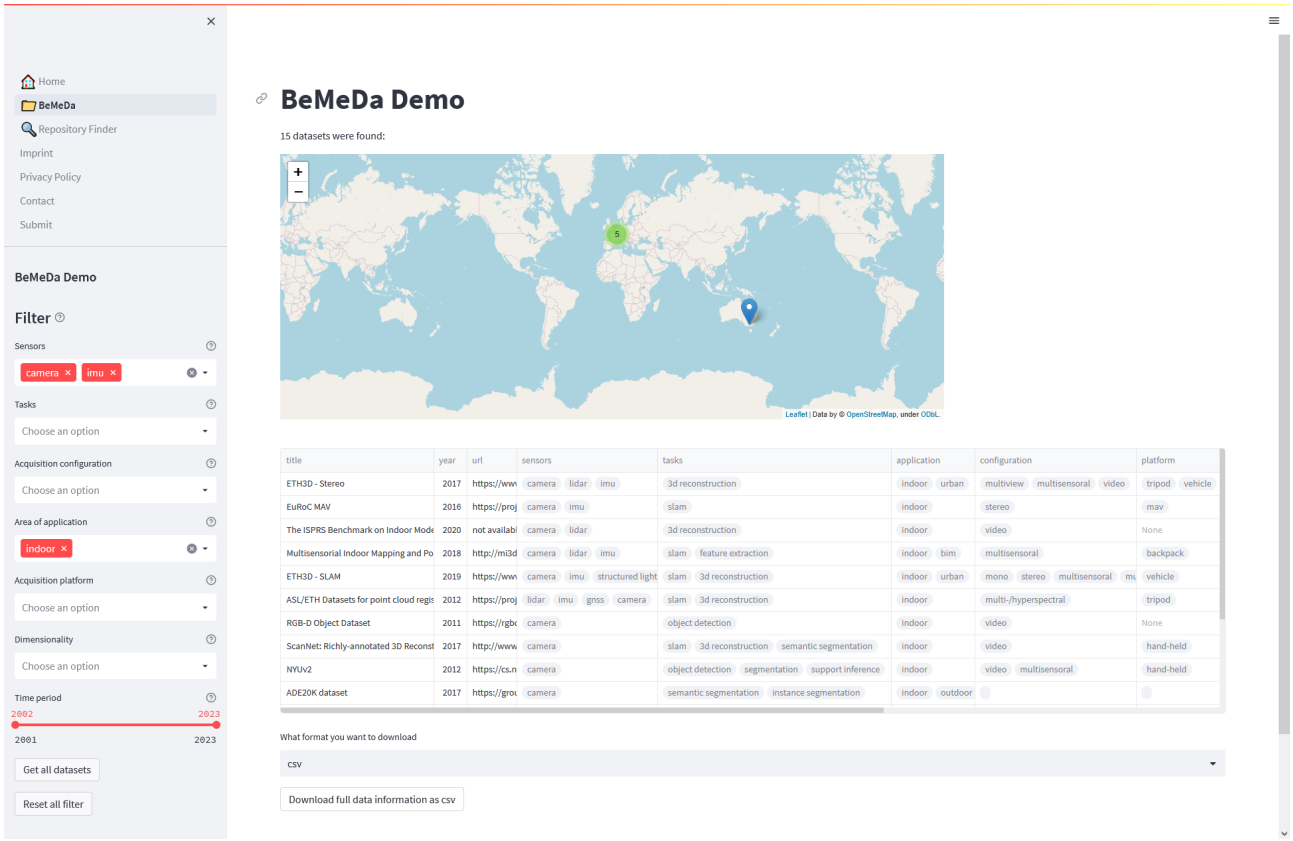

Figure 4. New design of the search results. The table view present the key attributes in a compact way. In addition, the world map visualize the locations of found datasets.

Trisovic, A., Mika, K., Boyd, C., Feger, S., Crosas, M., 2021. Repository Approaches to Improving the Quality of Shared Data and Code. *Data*, 6(2), 15. 10.3390/data6020015.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3, 1–9. 10.1038/sdata.2016.18.

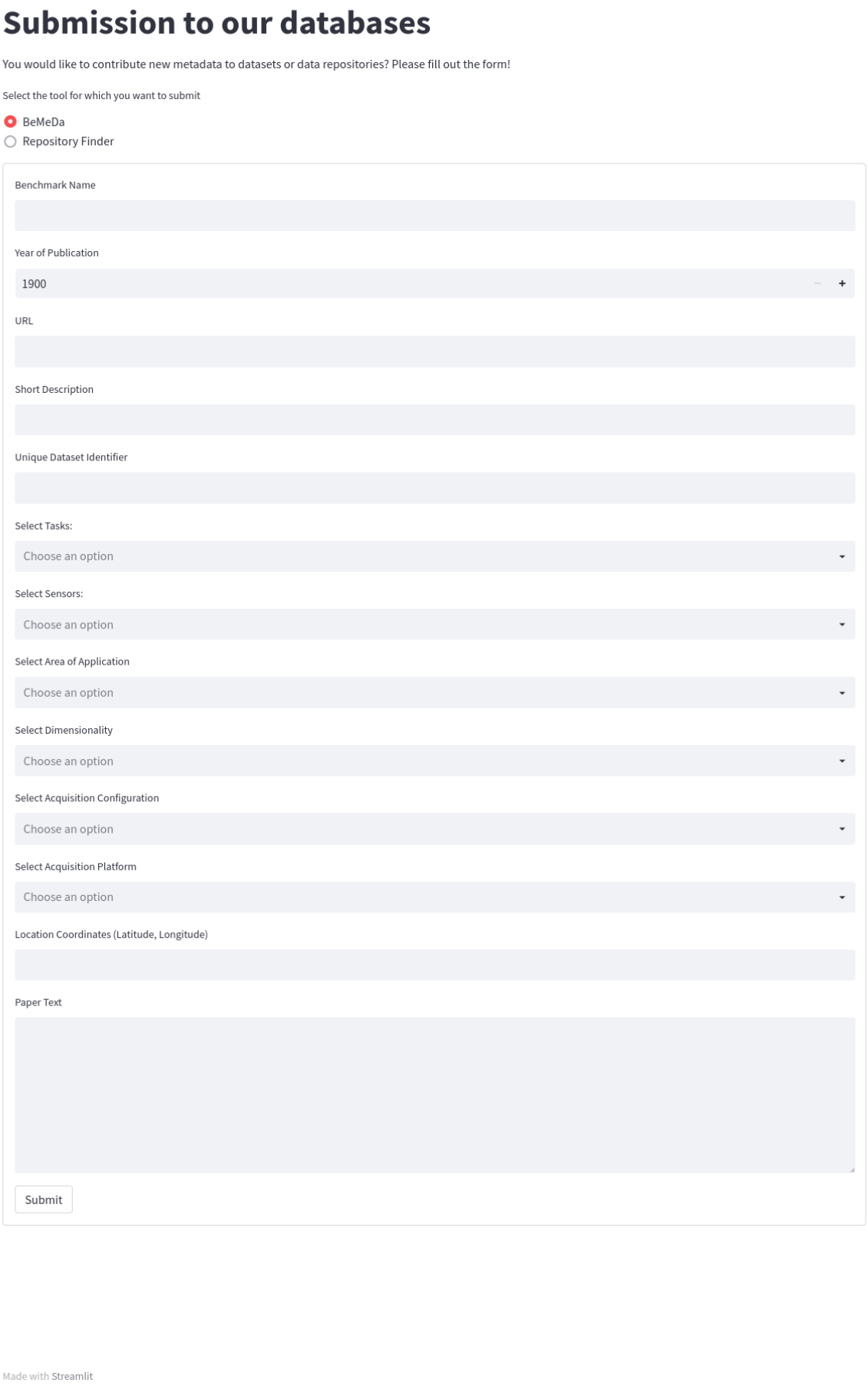

Figure 5. An excerpt from the submission form. Currently, the form contents free text fields and selection fields which correspond to the defined filter options.