Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-15-AC2, 2020

Open Access Earth System Science Data Discussions

Interactive comment on "P^3 – PetroPhysical Property Database – a global compilation of lab measured rock properties" *by* Kristian Bär et al.

Kristian Bär et al.

baer@geo.tu-darmstadt.de

Received and published: 16 May 2020

We would like to thank the anonymous referee #2 very much for the valuable comments and helpful suggestions to our manuscript. In the following, we would like to adress the general questions and questions more related to the text and consequently to the dataset and hope that our answers provide the requested clarification.

Referee question 1: Did the authors consider the possibility to use the GeoSciML model of geological features to build the database? (...)

Thank you very much for this question. We did indeed consider the possibility to use the GeoSciML model of geological features to build the database. However, the initial database of our own measurements was already based on the petrographic and

Printer-friendly version

Discussion paper

stratigraphic classification schemes as published by Bär et al. (2019). Reason for that is the direct link to the well database of the federal Geological Survey of Hessen, Germany, where the majority of our initial samples originated from. Additionally, these classifications with their internal hierachical structure allows for a much more detailed classification compared to the GeoSciML petrographic terms. For the next release of the P^3 database however, we prioritize to implement a direct link of our classifications to the GeoSciML geological features to allow for the proposed future compliance with INSPIRE. Since for both classifications, our own and the GeoSciML the terms are defined by geological vocabularies, such a link will be easy to implement but on the other hand will have no direct impact on the main contents of P^3 - the petrophysical properties.

Referee comment 2: It is worthwhile to mention the important fact in the text, that the dataset presented by the authors, follows almost all the FAIR (Findable Accessible Interoperable Reusable) requirements (...).

Thank you for this comment. We fully agree and adapted the text accordingly. New text: Page 3 line 13 (adapted): P^3 is a publicly accessible database containing physical rock properties measured in laboratory experiments. It is licensed under a creative commons (CC-BY 4.0) license and its structure follows the FAIR guiding principles for scientific data management and stewardship (Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18).

Referee comment 3 and 4: Regarding the datasets provided together with the manuscript, I would put more emphasis on the interoperable file format (...).

Thank you very much for this comment. We will remove the pdf-file of P^3 since both referees have mentioned that it is not useable. Additionally, we will provide a .csv file as interoprable file format as specified in the P^3 Readme. We will remove the TAB-separated txt file and check the correct use of the decimal marker as dot "." instead of ","

Referee comment 5: Page 4 – Line 24 to 35 – The authors explain the choice to use a flat file instead of a relational database with the pros and cons. (...) Did the author create an Entity-Relation (ER) model? (...) Did they perform any check onthe xls file to guarantee the database consistency?

Thank you very much for this comment. We reworded the sentence in order to clarify this point. We did create an Entity-Relation (ER) model already during the development of P^3 but did not change the file format for this first release from excel to another one. In the revised manuscript, we will add an Entity-Relationship-Diagram as Appendix. We checked for dublicate entries manually and adjusted the paragraph in the manuscript accordingly: 'The database was developed as flat-file format using Microsoft Excel to keep it as simple and easy to handle even by the unexperienced user as possible. While other database structures are in comparison much more efficient, their database management schemes may render it too difficult for users not familiar with SQL to recover the desired data. However, the internal design of P^3 with multiple sub-entities and tables is structured following a relational database management system (RDBMS, Codd, 1970) with an Entity-Relation (ER) model (see Appendix 1), so that it could easily be transferred to e.g. the well-established structured query language (SQL, Chamberlin and Boyce, 1974). Following this ERM the database could easily be organised into multiple tables using the names of the tables as unique keys as links to other sub-tables. The main advantages of a relational database over a flat file format are that data is uniquely stored just once, eliminating data duplication, as well as performance increases due to greater memory efficiency and easy filtering and rapid queries (Gard et al., 2019). However, the current flat-file structure allows for easy modification and extensions as new requirements emerge, as for example by adding more sub-tables for newly developed property measurements not fitting to any of the already included properties could be added at later stages. On the other hand, filtering and quality control to ensure that data is entered into the database only once and that no duplicates exist had to be done manually. In our case data duplicates where removed by checking the coordinates of each data point with a radius of uncertainty of

1 km and, if necessary, manually removing every double entry identified.'

Referee comment 6: - Page 7 – Line 7 – It is mentioned that the latitude/longitude coordinates are UTM based. The UTM system is a projected system that implies planar coordinates (i.e., easting and northing). In the database the field related to coordinates report geographical coordinates as decimal degree. It is not clear which is finally the used coordinate system or why there is this difference between the text and the database. This has to be sort out. Beside that it would be good to explain also how the coordinates retrieved from literature were treated if samples had different coordinate typology.

Thank you very much for this valuable comment. Indeed the manuscript text is misleading at this point. It was planned in an earlier version to use UTM. However, in the final version we used latitude/longitude in decimal degrees with a WGS84 reference system. For a conversion from literature, we used either Google Earth (Web Mercator Projection) or ArcGIS to allocate a latitude/longitude value in decimal degrees and the associated uncertainty to each data point. We are aware that this 'Google maps method' is not accurate but exact geographic information is quite often not provided in the literature used for this compilation. Typically in this case, only location names are provided. For all literature where both the exact coordinates and the reference system was given, or where the location of the datapoints was given on a georeferenced map with the required information on the corrdinate system used, we used ArcGIS. Therein, we used the same geographic projection as given in the original literatue and either included the points as tabular values or we geofrerenced the given maps accordingly and picked the points on the maps. Afterwards, the resulting coordinates were transferred to decimal degrees in the WGS84 reference with the transformation as suggested by ArcGIS. We have not documented the exact coordinate transformation used in each case.

We additionally would like to thank the anonymous referee #2 for pointing out some typos and sentences with problematic grammar as well as suggestions for improvements

of the figures and we will change these in the revised version of the manuscript.

Last referee comment: In the conclusions and perspectives the plan to develop a public accessible web-based interface is highly recommended for the future and should be prioritized, because the high number of rows (more than 75000) and columns (around 300) doesn't make the database so easily query-able and browsable, even in the excel software package.

Thank you for the remark. We will prioritize this work in the future as recommended. Please be referred here to our answers to the comments of the first anonymous referee: 'The database in its current version will be permanently available and accessible via the doi and GFZ data services. The future development and extensions of the database as almost everywhere in the scientific world strongly depends on ongoing project funding and successful research proposals in the future. Funding for extensions is currently available until the end of 2022. The online portal is planned to be developed within the next four years. A research proposal for a project where this will be developed is currently under review.'

Kind regards

Kristian Bär, Thomas Reinsch, Judith Bott

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-15, 2020.

Printer-friendly version

Discussion paper