Geoscientific
Instrumentation
Methods and
Data Systems

A universal and multi-dimensional model for analytical data on geological samples

Yutong He[1], Di Tian[2], Hongxia Wang[2], Li Yao[3], Miao Yu[1], and Pengfei Chen[4]

[1]School of Electronic Information Engineering, University of Electronic Science and Technology, Zhongshan Institute, Zhongshan, Guangdong, China
[2]College of Instrumentation & Electrical Engineering, Jilin University, Changchun, Jilin, China
[3]College of Earth Science, Jilin University, Jilin, China
[4]The 41st research institute CETC, Qingdao, China

Correspondence: Di Tian (tiandi@jlu.edu.cn) and Miao Yu (yumiao@zsc.edu.cn)

Abstract. To promote the sharing and reutilization of geoanalytical data, various geoanalytical databases have been established over the last 30 years. Data models, which form the core of a database, are themselves the subjects of intensive studies. Data models determine the contents stored in the databases and applications of the databases. However, most geoanalytical data models have been designed for specific geological applications, which has led to strong heterogeneity between databases. It is therefore difficult for researchers to communicate and integrate geoanalytical data between databases. In particular, every time a new database is constructed, the time-consuming process of redesigning a data model significantly increases the development cycle. This study introduces a new data model that is universally applicable and highly efficient. The data model is applied to various geoanalytical methods and corresponding applications, and comprehensive analytical data contents together with associated background metadata are summarized and catalogued. Universal data attributes are then designed based on these metadata, which means that the model can be used for any geoanalytical database. Additionally, a multi-dimensional data mode is adopted, providing geological researchers with the ability to analyze geoanalytical data from six or more dimensions with high efficiency. Part of the model is implemented with the typical database system (MySQL) and comprehensive comparison experiments with existing geoanalytical data model are presented. The result unambiguously proves that the data model developed in this paper exceeds existing models in efficiency.

1 Introduction

Geoanalytical data include measurements of major and trace elements, rare Earth elements (REEs), isotopes, and structures and morphology of geological samples analyzed by various analytical instruments and techniques such as ICP (inductively coupled plasma), LA-ICP-MS (laser ablation inductively coupled plasma mass spectrometry), ICP-MS (inductively coupled plasma mass spectrometry), EPMA (electro-probe microanalyzer), SIMS (secondary ion mass spectroscopy), SEM (scanning electron microscope), TEM (transmission electron microscopy), XRF (X-ray fluorescence), and XRD (X-ray diffraction). Geoanalytical data effectively reflect the material composition, internal structure, external characteristics, interaction, and evolution history of the Earth and represent the most important support for geological researchers in their aim to understand the Earth and exploit its resources for the survival and development of human society. Enormous financial, material, and human resources have been invested into the geological surveys and geoanalysis required to acquire more comprehensive and abundant geoanalytical data. Over time, tremendous volumes of geoanalytical data have been created, and these volumes continue to increase at a high rate. It is of paramount importance that these data are curated effectively and that adequate background information, such as sample description, sampling information, and analysis information, is included, so that geological researchers can utilize the data according to their requirements. This will also facilitate the reutilization of the precious geoanalytical data. In addition, with the

accumulation of large volumes of data, statistical analysis and data mining can be conducted on these data to provide a more comprehensive and advanced scientific understanding of the Earth. Hence, a variety of geoanalytical databases aimed at managing, sharing, and reutilizing geoanalytical information has been constructed and used as advanced tools in geological studies. The analysis and comparison of existing geoanalytical data models, as well as the development of improved models, are therefore a worthy and significant study to be conducted. Over the last decades, several studies of geoanalytical data models have been conducted. As early as 1977, Jeorge Van Trump and colleagues described a data model for environmental geochemical surveying and mineral resource exploration in the United States of America (Jr and Miesch, 1977). Lehnert et al. (2000) suggested a data model for the storage of global geochemical data of rocks. Their data model provides a complete summary of essential geochemical data contents and a robust structure with a relational database management system (RDBMS). Numerous databases such as GEOROC (Geochemistry of Rocks of the Oceans and Continents), NAVDAT (the North American Volcanic and Intrusive Rock Database), and PetDB (the interactive web-based Petrological Database of the Ocean Floor) have since been constructed based on this model, and it is used by geological researches worldwide. In particular, PetDB has been used for a considerable amount of high-impact research such as *Nature* (Brandl et al., 2013; Carbotte et al., 2013; Cheng et al., 2016; Dick and Zhou, 2014; Helo et al., 2011; Hoernle et al., 2011; Kamenov et al., 2011; Kelley, 2014; Samuel and King, 2014; Schlindwein and Schmid, 2016; Straub et al., 2009) and *Science* (Cottrell and Kelley, 2013; Greber et al., 2017; Joy et al., 2012; Kelley and Cottrell, 2009; Mcnutt et al., 2016). A limitation of existing geoanalytical data models is their specificity to particular applications or geological domains and their focus on the description and curation of only a certain portion of geoanalytical data. For example, RU_CAGeochem is specifically focused on major and trace element concentrations and Sr, Nd, and Pb isotopic ratios of American volcanic rocks (Carr et al., 2014). Another database is focused on lead isotopes of copper ores from the southeastern Alps (Artioli et al., 2016). Many other examples of similarly specific geoanalytical databases and associated models exist (e.g., Artioli et al., 2016; Hellström, 2016; Lopes et al., 2014; Siegel et al., 2012; Strong et al., 2016). The consequence of this development is that each database exists as a separate island, and it is difficult for researchers to communicate and integrate geoanalytical data between databases. In particular, every time a database is constructed, a data model has to be redesigned. This consumes considerable amounts of time and prolongs the development cycle. In addition, the vast majority of models are designed based on relational models, which focus on the construction of relations between different data categories. When users query and utilize the geoanalytical data from different dimensions, these types of models utilize compli-

cated joints between different tables to query the target data, which decrease efficiency as the amount of data increases. However, the exploration of such data models including the background items has laid a solid foundation for later study of advanced geoanalytical data models. At present, the development of various new techniques provides us with the opportunities to design more comprehensive and advanced geoanalytical data models. In this study, we introduce a novel, universal, and efficient geoanalytical data model. First, we provide an overview of geoanalytical methods and applications to summarize the geoanalytical data available. Then, we design universal data attributes based on these data and develop a multi-dimensional data model. Finally, we evaluate the model to validate its efficiency.

2 Overview of geoanalytical data contents

In recent years, many new geoanalytical methods and instruments have been developed, creating novel kinds of data (Linge et al., 2017). A truly universal data model should have the ability to accommodate all kinds of geoanalytical data. In addition, the data model should be capable of making all stored data readily available for reutilization by geological researchers. In order to develop a model with such capabilities, a comprehensive set of geoanalytical data, together with related background information required for reutilization of the data, was summarized and categorized, as outlined below. First, analytical techniques and their applications were studied to comprehensively summarize geoanalytical measurement data. This process is outlined in Fig. 1. Because of the great diversity of analytical methods and geological applications, Fig. 1 only shows a few examples to indicate the method adopted in this paper. The five categories (namely, bulk analysis; microanalysis; isotope analysis; morphology, structure, and valence analysis; and organic analysis) were divided according to the analytical technique used. In this way, data from each category were categorized according to analytical instruments (e.g., SEM, SNM, and EPMA for microanalysis). In the next step, the data were grouped according to geological applications. The comprehensive list of geoanalytical measurement data items used in the present study, compiled from a thorough literature review, is presented in Fig. 2. In the case of bulk analysis, most measurements ultimately provided major, trace, and ultra-trace element concentration data. Microanalysis can yield data of elemental concentrations in a microregion, as well as structural information of geological samples acquired by secondary electron and backscattered electron techniques, commonly stored as image files. For geochronology and stable isotopic analysis (GSI analysis), most measurement data are isotopic ratios. For morphology, structure and valence analysis (MSV analysis), the most common measurement data are image files such as X-ray photoelectron spectroscopy (XPS) spectra or XRD patterns. Organic analysis is a new analyt-
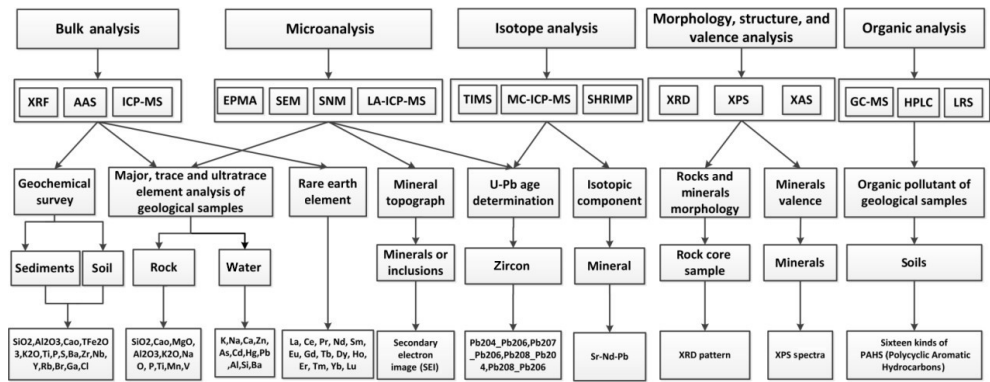
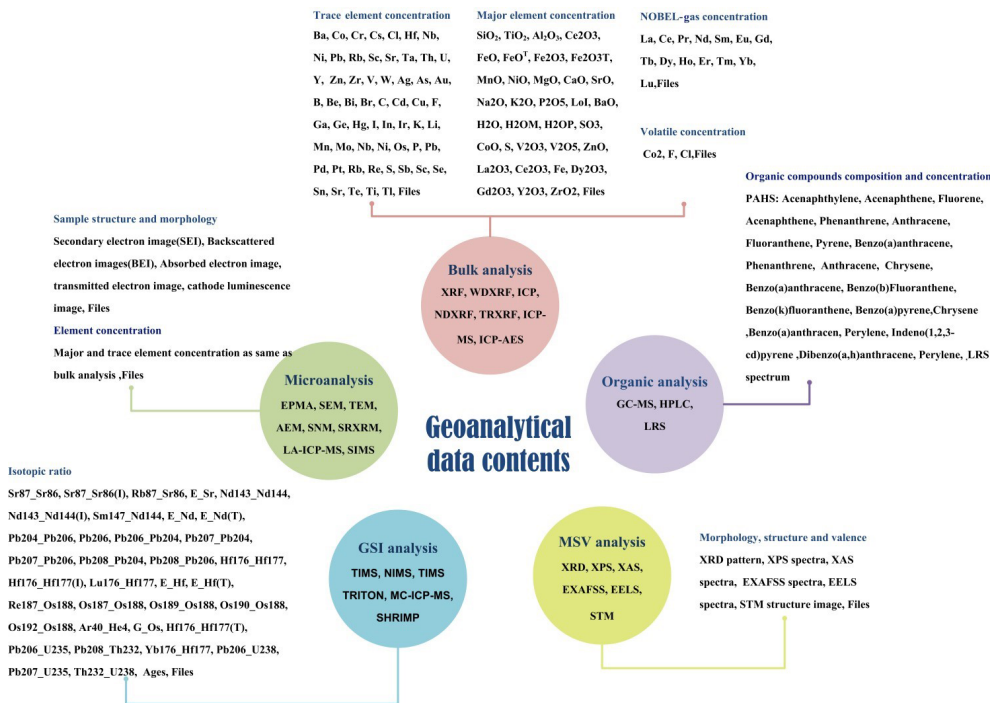**Figure 1.** Process of summarizing the geoanalytical data contents.

Figure 2. Lists of geoanalytical methods and measurement data items.

ical method which is used for the analysis of environmental geological samples. The most common application of this method in the geological literature is the analysis of the 16 kinds of polycyclic aromatic hydrocarbons (PAHs) in soils.

Background information describing the analyzed samples and data quality has to be incorporated, because it is indispensable for proper evaluation, efficient recovery, and sorting of the compiled data. Hence, background metadata are summarized based on the investigations of the geological researchers and the contents of existing databases (Adcock et al., 2003; Lehnert et al., 2000). Table 1 lists details of the background metadata used during the present study. In this study, the background metadata are divided into three parts: sample metadata provide geological researchers with infor-

Table 1. Background metadata of geoanalytical data.

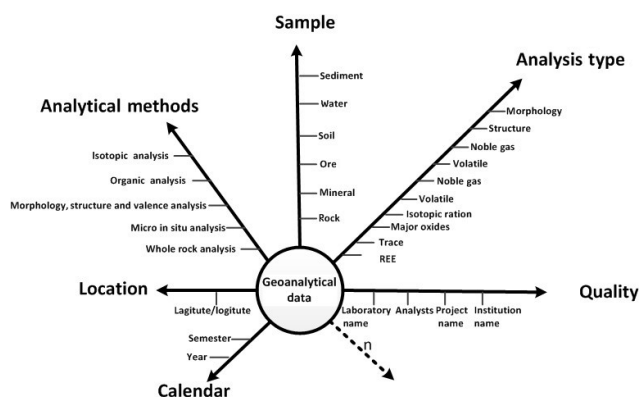| Background metadata	
Sample metadata	sample type, sample name, sample description
Sampling metadata	sampling site description, longitude, latitude, sampling methods, sampling depth, sampling institution
Quality metadata	laboratory name, project name, publish link, analytical method, analysts

Figure 3. Multi-dimensional conceptual data model (CDM) for geoanalytical data.

Figure 4. Logical data model (LDM) of the geoanalytical data model.

mation about geological materials, sampling metadata provide information about environmental conditions in the field, and quality metadata allow geological researchers to make an assessment of data quality and usability (Table 1). The background metadata items listed in Table 1 are the most essential information required for every kind of geoanalytical measurement data. More specific attributes are not included in our model.

Geoanalytical data modeling

This section outlines how the novel geoanalytical data model was designed, utilizing the data summarized above. Despite their limitations, the currently relational data mode is the most commonly used pattern for geoanalytical data models. The relational data mode constructs relations between each group of data within the database. This means that more data categories inevitably lead to much more data relations, increasing storage demands and the time required to query the database. Compared to such conventional relational data models (Beynon-Davies, 2004), multi-dimensional models (MDMs), which are widely utilized during the development of big data science and data mining, are single subject-oriented sources for analyzing data based on various dimensions (Niemi and Hirvonen, 2003). Multi-dimensional modeling approaches share characteristics with fast analysis of shared multi-dimensional information (FASMI). In particular, MDM offers the advantage of a relatively simple and straightforward database design, which nevertheless supports powerful analyses and is relatively well understood by the end users (Hoberman, 2005). As a modeling framework, MDM has a conceptual and a logical phase of design, composed of a fact table and several dimension tables (Höpken et al., 2013). Facts comprise numeric and additive characteristics of the data, which can be accumulated along multiple dimensions. Frequently, researchers are interested in analyzing geoanalytical measurement data from different metadata
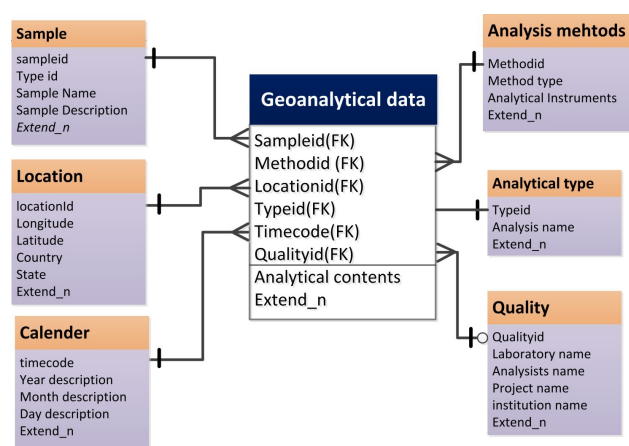
perspectives. Hence, the MDM approach is ideally suited for the design of geoanalytical data models. Here, the geoanalytical data are the fact data, and other background information are dimension data. The use of the MDM modeling framework applied in the present study will allow geological researchers to rapidly analyze geoanalytical data based on numerous metadata criteria.

2.1 Conceptual data model (CDM)

A conceptual data model (CDM) includes the definition of its universal attributes and a rough design of its structure. It represents the primary phase of data model design, independent from the detailed techniques of computer systems. Figure 3 presents the multi-dimensional CDM we developed for geoanalytical data. Here, with the abstraction of universal concepts present in geoanalytical data, the model becomes more flexible and universally applicable. The geoanalytical data are placed in the center of the model, in the form of a fact table. The associated background information is categorized and abstracted as various dimensions which are represented by different axes in Fig. 3. The six dimensions of our CDM are sample, analysis type, analytical methods, location, time, and quality. This arrangement allows geological researchers to analyze geoanalytical data from six different dimensions or any combination thereof. The marks in each dimension represent the detailed measurement conditions. The "n" dimension is an expansible dimension, which can be added according to the specific model application.

2.2 Logical data model (LDM)

A logical data model (LDM) is a CDM written in unified modeling language (UML) (Evans et al., 2014). Logical model design leads to a logical scheme, defining objects, attributes, and relationships (Chmura and Heumann, 2005). The LDM scheme can be easily implemented by any DBMS.
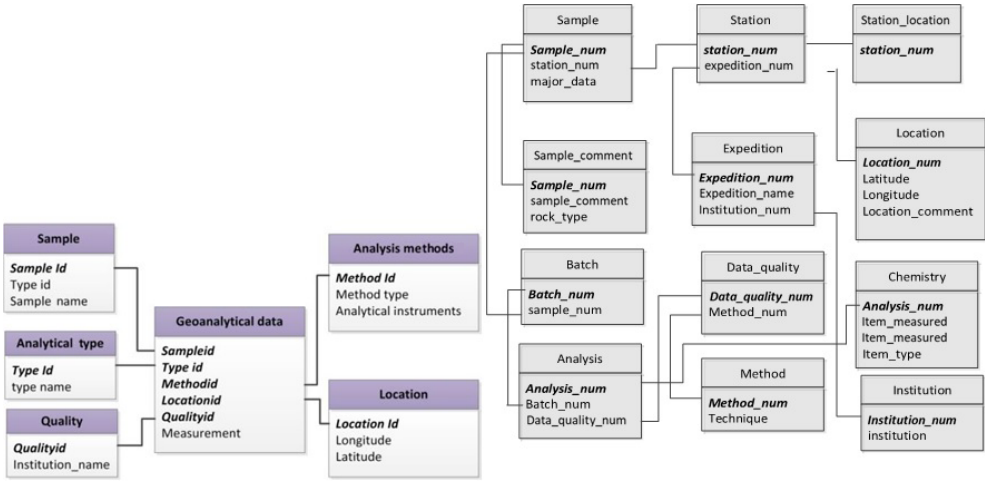
Figure 5. Physical data model (PDM) structure comparison with the Lehnert rock data model.

Figure 6. Comparison of insert and query operations in structured query languages (SQLs).

Figure 4 shows the LDM scheme designed for geoanalytical data. Each box in the LDM represents an object, and items in the box are its attributes. The relations between object are represented with lines. There are three kinds of symbols associated with the lines. The short line denotes "1", the circle denotes "0" (which means "maybe"), and the triangle denotes "many". Lines and symbols define the relations between objects. The additional notation foreign key (FK) is added if the attribute in one object uniquely identifies an attribute in another object. For example, the sample ID in the geoanalytical data object is a foreign key of "sampleid" in the sample object, because they have the same value. By means of this foreign key, the data contents of the two objects are connected. For each object, a few extended attributes are added

(*Extend_n* in Fig. 4). This feature allows developers to add database-specific attributes to this model, increasing its flexibility and universal applicability.

3 Implementation and evaluation

In order to evaluate the performance of our model, we carried out a comparison experiment with the widely used Lehnert rock geochemical data model (Lehnert et al., 2000). In order to conduct the experiment, a physical data model (PDM) needed to be created with a database management system. As RDBMS is the most common technique used in geoanalytical databases, MySQL, which is a widely used RDBMS,
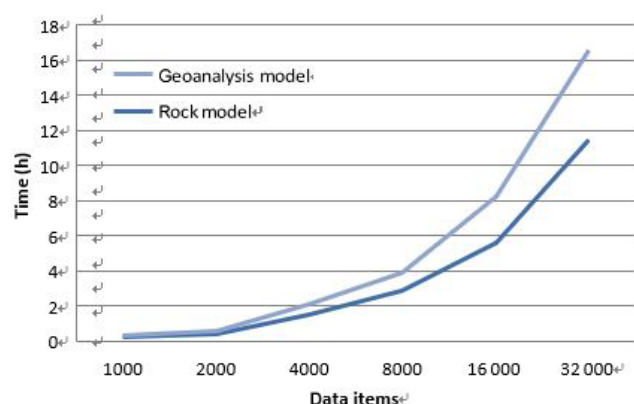
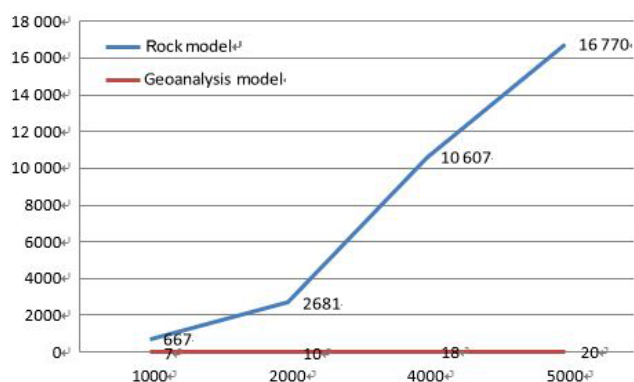**Figure 7.** Time spent on data insert operations with increasing amounts of data.

Figure 8. Time requirement for data queries (latitude and longitude).

Figure 9. Space usages of the two data models with increasing amount of data items.

and efficient. In contrast, for the rock data model, query time costs increased exponentially with the increasing amount of data items.

4 Conclusions

The geoanalytical data model presented herein is flexible and appropriate for a broad range of applications to geoanalytical data. The model has the following general characteristics:

1. Its universality allows the model to accommodate any type of geoanalytical data for various geological materials, as well as all significant metadata.

2. The adoption of a multi-dimensional data model framework provides geological researchers with the ability to analyze geoanalytical data from different dimensions. In addition to the sample description and location criteria commonly used in existed databases, this model provides four additional query criteria (method, quality, time, and analysis).

3. There are minimum data relations between different objects. Relations between different background metadata objects have been avoided in order to construct robust relations between background metadata and measurement data. This increases the model's efficiency when geoanalytical data are inserted or queried while simultaneously decreasing its space usage.

It is hoped that the design of this model will allow for the unified construction of geoanalytical databases. The model enables the accumulation and integration of significant amounts of diverse geoanalytical data. By utilization of the big data analysis techniques described in our study, geological researchers could analyze geoanalytical data with high efficiency and develop novel methods to conduct Earth science studies.

was adopted to implement the two models. A specific data item (rock type: andesite; location: Sycamore Hall; latitude: 36.27.12° N, longitude: 83.34.12° W; institution: Jilin University; method: ICP-MS; SiO2:58.9; FiO2:1.13) was used as test data and tables related to the data contents were implemented. We analyzed the two models from two perspectives: developers and users. For developers, the comparison of the PDM structure is shown in Fig. 5, and query operation descriptions are presented Fig. 6. The comparison clearly indicates that the geoanalytical data model is more succinct than rock data model and saves time and computer resources. Three model performance indicators (insert time, storage space usage, and retrieval time) were evaluated with the increasing of amounts of data. The results are shown in Figs. 7, 8, and 9, respectively. Figure 7 shows clearly that the process of data insertion is considerably faster for the geoanalytical data model when compared to the rock data model. Figure 9 shows clearly that the storage space usage is relatively less than rock data model. In the case of data query (Fig. 8), the difference in time consumption is even more striking. With an increasing amount of data items, the query time of the geoanalytical data model remains very fast
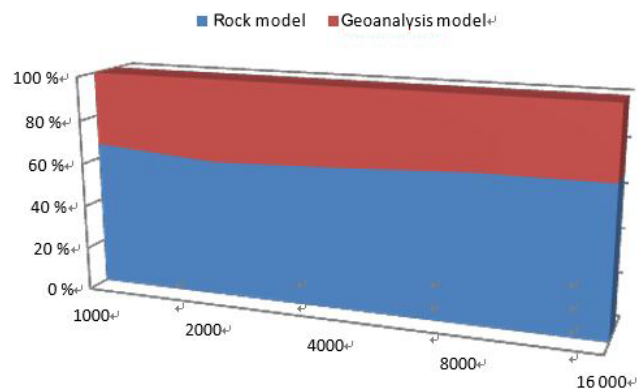
References

Adcock, S., Grunsky, E., and Laframboise, R.: A Universal Geochemical Survey Data Model, available at: https://www.researchgate.net/publication/266066484_A_Universal_Geochemical_Survey_Data_Model, last access: 1 October, 2019.

Artioli, G., Angelini, I., Nimis, P., and Villa, I. M.: A lead-isotope database of copper ores from the Southeastern Alps: A tool for the investigation of prehistoric copper metallurgy[J], J. Archaeol. Sci., 75, 27–39, https://doi.org/10.1016/j.jas.2016.09.005, 2016.

Beynon-Davies, P.: Relational Data Model, in: Database Systems, Palgrave, London, https://doi.org/10.1007/978-0-230-00107-7_7, 2004.

Brandl, P. A., Regelous, M., Beier, C., and Haase, K. M.: High mantle temperatures following rifting caused by continental insulation, Nat. Geosci., 6, 391–394, https://doi.org/10.1038/ngeo1758, 2013.

Carbotte, S. M., Marjanović, M., Carton, H., Mutter, J. C., Canales, J. P., Nedimović, M. R., Han, S., and Perfit, M. R.: Fine-scale segmentation of the crustal magma reservoir beneath the East Pacific Rise, Nat. Geosci., 6, 866–870, https://doi.org/10.1038/ngeo1933, 2013.

Carbotte, S. M., Marjanović, M., Carton, H., Mutter, J. C., Canales, J. P., Nedimović, M. R., Han, S., and Perfit, M. R.: Fine-scale segmentation of the crustal magma reservoir beneath the East Pacific Rise, Nat. Geosci., 6, 866–870, https://doi.org/10.1038/ngeo1933, 2013.

Carr, M. J., Feigenson, M. D., Bolge, L. L., Walker, J. A., and Gazel, E.: RU_CAGeochem, a database and sample repository for Central American volcanic rocks at Rutgers University[J], Geosci. Data J., 1, 43–48, https://doi.org/10.1002/gdj3.10, 2015.

Cheng, H., Zhou, H., Yang, Q., Zhang, L., Ji, F., and Henry, D.: Jurassic zircons from the Southwest Indian Ridge, Sci. Rep, 6, 26260, https://doi.org/10.1038/srep26260, 2016.

Chmura, A. and Heumann, J. M.: Logical Data Modeling, Integrated, 5, 179–203, https://doi.org/10.1007/b100064, 2005.

Cottrell, E. and Kelley, K. A.: Redox heterogeneity in mid-ocean ridge basalts as a function of mantle source, Science, 340, 1314, https://doi.org/10.1126/science.1233299, 2013.

Dick, H. J. B. and Zhou, H.: Ocean rises are products of variable mantle composition, temperature and focused melting, Nat. Geosci., 8, 68–74, https://doi.org/10.1038/ngeo2318 , 2014.

Evans, A., France, R., Lano, K., and Rumpe, B.: The UML as a Formal Modeling Notation, Comput. Stand. Interf., 19, 325–334, https://doi.org/10.1016/s0920-5489(98)00020-8, 1998.

Evans, A., France, R., Lano, K., Francea, R., Evansb, A., Lanoc, K., and Rumped, B.: The UML as a Formal Modeling Notation[J], Comput. Stand. Interf., 19, 325–334, https://doi.org/10.1007/978-3-540-48480-6_26, 2014.

Greber, N. D., Dauphas, N., Bekker, A., Ptáček, M. P., Bindeman, I. N., and Hofmann, A.: Titanium isotopic evidence for felsic crust and plate tectonics 3.5 billion years ago, Science, 357, 1271–1274, https://doi.org/10.1126/science.aan8086, 2017.

Hellström, F.: The Swedish bedrock age database, https://doi.org/10.13140/RG.2.1.1528.8085, 2016.

Helo, C., Longpré, M. A., Shimizu, N., Clague, D. A., and Stix, J.: Explosive eruptions at mid -ocean ridges driven by CO_2-rich magmas, Nat. Geosci., 4, 260–263, https://doi.org/10.1038/ngeo1104, 2011.

Hoberman, S.: Data Modeling Essentials, 3rd, The Morgan Kaufmann Series in Data Management Systems ,Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 560, available at: https://dl.acm.org/citation.cfm?id=1211351 (last access: 1 October 2019), 2005.

Hoernle, K., Hauff, F., Werner, R., Bogaard, P. V. D., Gibbons, A. D., Conrad, S., and Müller, R. D.: Origin of Indian Ocean Seamount Province by shallow recycling of continental lithosphere, Nat. Geosci., 4, 883–887, https://doi.org/10.1038/ngeo1331, 2011.

Höpken, W., Fuchs, M., Höll, G., Keil, D., and Lexhagen, M.: Multi-Dimensional Data Modelling for a Tourism Destination Data Warehouse, https://doi.org/10.1007/978-3-642-36309-2_14, 2013.

Joy, K. H., Zolensky, M. E., Nagashima, K., Huss, G. R., Ross, D. K., Mckay, D. S., and Kring, D. A.: Direct detection of projectile relics from the end of the lunar basin-forming epoch, Science, 336, 1426–1429, https://doi.org/10.1126/science.1219633, 2012

Jr, V. T. and Miesch, A. T.: The U.S. geological survey rass-statpac system for management and statistical reduc-

tion of geochemical data, Comput. Geosci., 3, 475–488, https://doi.org/10.1016/0098-3004(77)90025-5, 1977

Kamenov, G. D., Perfit, M. R., Lewis, J. F., Goss, A. R., Jr, R. A., and Shuster, R. D.: Ancient lithospheric source for Quaternary lavas in Hispaniola, Nat. Geosci., 4, 554–557, https://doi.org/10.1038/ngeo1203, 2011

Kelley, K. A.: Inside Earth Runs Hot and Cold, Science, 344, 51–52, https://doi.org/10.1126/science.1252089 , 2014

Kelley, K. A. and Cottrell, E.: Water and the oxidation state of subduction zone magmas, Science, 325, 605–607, https://doi.org/10.1126/science, 2009.

Lehnert, K., Su, Y., Langmuir, C. H., Sarbas, B., and Nohl, U.: A global geochemical database structure for rocks, Geochem. Geophys. Geosyst., 1, 179–188, https://doi.org/10.1029/1999gc000026, 2000.

Linge, K. L., Bédard, L. P., Bugoi, R., Enzweiler, J., Jochum, K. P., Kilian, R., Liu, J., Marin-Carbonne, J., Merchel, S., and Munnik, F.: GGR Biennial Critical Review: Analytical Developments Since 2014, Geostand. Geoanal. Res., 36, 337–398, https://doi.org/10.1111/ggr.12200, 2017.

Lopes, C., Ferreira, A., Chichorro, M. A., Pereira, M. F. C., Almeida, J. A., and Solá, A. R.: Chroniberia: The Ongoing Development of a Geochronological GIS Database of Iberia[J], Strati, 2013, https://doi.org/10.1007/978-3-319-04364-7_138, 2014.

Mcnutt, M. K., Lehnert, K., Hanson, B., and Nosek, B. A.: Liberating field science samples and data, Science, 351, 1024, https://doi.org/10.1126/science.aad7048, 2016.

Niemi, T. and Hirvonen, L.: Multidimensional data model and query language for informetrics, John Wiley & Sons, Inc., 939–951, https://doi.org/10.1002/asi.10290, 2003.

Samuel, H. and King, S. D.: Mixing at mid-ocean ridges controlled by small-scale convection and plate motion, Nat. Geosci., 7, 602–605, https://doi.org/10.1038/ngeo2208, 2014.

Schlindwein, V. and Schmid, F.: Mid-ocean-ridge seismicity reveals extreme types of ocean lithosphere, Nature, 535, 276–279, https://doi.org/10.1038/nature18277, 2016.

Siegel, C., Bryan, S.E., Purdy, D., Gust, D., Allen, C., Uysal, T., and Champion, D.: A new database compilation of whole-rock chemical and geochronological data of igneous rocks in Queensland: a new resource for HDR geothermal resource exploration[C], in: Proceedings of the 2011 Australian Geothermal Energy Conference, editedy by: Rudd, A., Geoscience Australia, Sydney, 239–244, 2012.

Straub, S. M., Goldstein, S. L., Class, C., and Schmidt, A.: Mid-ocean-ridge basalt of Indian type in the northwest Pacific Ocean basin, Nat. Geosci., 2, 286–289, https://doi.org/10.1038/ngeo471, 2009.

Strong, D. T., Turnbull, R. E., Haubrock, S., and Mortimer, N.: Petlab: New Zealand's national rock catalogue and geoanalytical database[J], New Zealand, J. Geol. Geophys., 59, 475–481, https://doi.org/10.1080/00288306.2016.1157086, 2016.

Geosci. Instrum. Method. Data Syst., 8, 277–284, 2019

www.geosci-instrum-method-data-syst.net/8/277/2019/