Software Dedicated for the Curation of Geochemical Data Sets in Analytical Laboratories

Yusuke Yachi, Hiroshi Kitagawa, Tak Kunihiro* and Eizo Nakamura

The Pleasance Memorial Laboratory for Geochemistry and Cosmochemistry, Institute for Study of the Earth’s Interior, Okayama University, Misasa, Tottori 682-0193, Japan

* Corresponding author, e-mail: sk@misasa.okayama-u.ac.jp

Software designed for analytical laboratories to guarantee traceability and accessibility of rocks with their geochemical properties has been developed. The software documents the sample origin, current sample location and the location of any sample subsets (e.g., thin sections, solutions, etc.), and archives all associated geochemical data sets. The software can be installed on a personal computer so it is available for use in any laboratory and allows curation before and after publication. The software will be of use in integrating and sharing geological reference materials within and among institutes. In this article, the system design and implementation are detailed. All source codes for the software are available at http://dream.misasa.okayama-u.ac.jp/.

Keywords: petrology, geochemistry, curation, data archive, software

Received 18 Jul 12 – Accepted 27 Jan 13

Un logiciel conçu pour les laboratoires d'analyse afin de garantir la tracabilité et l'accessibilité des échantillons de roches et de leurs caractéristiques géochimiques a été développé. Le logiciel documente l'origine de l'échantillon, son emplacement actuel et l'emplacement de tous ses dérivés (par exemple, des lames minces, des solutions, etc.), et toutes les bases de données, géochimiques associées. Le logiciel peut être installé sur un ordinateur personnel et disponible pour une utilisation dans n'importe quel laboratoire et permet la conservation avant et après publication. Le logiciel sera utile dans l'intégration et le partage de matériaux de référence géologiques à l'intérieur et entre des instituts. Dans cet article, la conception et la mise en œuvre du système sont détaillées. Tous les codes sources du logiciel sont disponibles à l'adresse: http://dream.misasa.okayama-u.ac.jp/.

Mots-clés : pétrologie, géochimie, conservation, archivage des données, logiciel.

Over the past several decades, Earth scientists have collected vast amounts of numerical data on the material properties of minerals. With the assistance of modern technology, several different projects have been conducted to create combined data sets that facilitate the utilisation of such data. Some of the better known data sets are MagIC (magnetic data, Koeppe et al. 2005), RRUFF (spectroscopic data, Daws 2006), LEPR (phase equilibria data, Hirschmann et al. 2008) as well as various thermodynamic data sets (e.g., Holland and Powell 1998). These various data sets have been applied to a broad range of problems in the Earth sciences, especially in the fields of petrology, geophysics and tectonics. Subsequently, various online database systems that focus on integration of existing geochemical data sets have been developed (Table 1). Some of the better known ones include: (a) GEOROC (Shabas 2002), GeoReM (Oschum et al. 2005) and Geochron (http://www.geochron.org) which integrate data sets of volcanic rocks, reference materials, and geochronology, respectively, (b) PetDB (Lehnert et al. 2000) and SedDB (http://www.seddb.org/) which integrate data sets from the ocean seafloor and sediment, and (c) NAVDAT (Carlson et al. 2001): http://www.navdat.org/ which stores elemental, isotope, and age data sets for Mesozoic and younger igneous rocks in North America. A common feature of all of these systems is that they are based solely on published data sets from the literature.

For certain areas within the Earth sciences, it is not only the preservation and cataloguing of existing geochemical data that is important, but also preservation of the samples themselves. Indeed, preservation of original sample material should be a crucial component in geochemistry given the likelihood that future state-of-the-art technologies will permit re-examination of existing sample material. This is especially
Table 1.
Comparison of geochemical database systems

<table>
<thead>
<tr>
<th>Database system</th>
<th>Target</th>
<th>Geochemical dataset</th>
<th>In situ visual</th>
<th>Curation</th>
<th>Open source</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEOROC&lt;sup&gt;®&lt;/sup&gt;</td>
<td>igneous</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>GeoRpm&lt;sup&gt;®&lt;/sup&gt;</td>
<td>RMs&lt;sup&gt;®&lt;/sup&gt;</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>PerDB&lt;sup&gt;®&lt;/sup&gt;</td>
<td>seafloor</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>SedDB&lt;sup&gt;®&lt;/sup&gt;</td>
<td>sediment</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>NAVDAT&lt;sup&gt;®&lt;/sup&gt;</td>
<td>igneous&lt;sup&gt;®&lt;/sup&gt;</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>GRL&lt;sup&gt;®&lt;/sup&gt;</td>
<td>any kind</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>SEASAR&lt;sup&gt;®&lt;/sup&gt;</td>
<td>metamorphic</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>MetPebDB&lt;sup&gt;®&lt;/sup&gt;</td>
<td>any kind</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Medusa DB&lt;sup&gt;®&lt;/sup&gt;</td>
<td>any kind</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

GEOROC, GeoRpm, PerDB, SedDB, NAVDAT and GRL handle bulk and in situ geochemical datasets. MetPebDB handles both geochemical dataset and curation, but the target is limited to metamorphic rocks. Target denotes the target rock type. ‘Curation’ indicates the ability to trace the current sample location and the location of any sample subsets (e.g., thin section, solution, etc.). ‘In situ visual’ indicates the ability to visualise locations of in situ analysis on images.

<sup>®</sup> Online database system.
<sup>®</sup> Any kind of rock but limited to reference materials.
<sup>®</sup> Sampled from North America only.
<sup>®</sup> Available at http://www.earthchem.org/gvl.
<sup>®</sup> Lehner et al. (2000).
<sup>®</sup> Soor et al. (2009).
<sup>®</sup> Database system run by software proposed in this study.

true for specimens that are very difficult to obtain such as those sampled by deep-sea drilling (e.g., Barnitt and Stott 1991) or extraterrestrial sample return missions such as Hayabusa (Fujiwara et al. 2006), Stardust (Brownlee et al. 1996) and Genesis (Barnett et al. 2003), which brought materials from an asteroid, a comet and the Sun, respectively. For such missions, curation of materials is considered to be one of the major objectives, and the materials are stored and curated at dedicated facilities.

Various museums have utilised commercial database systems that handle both curatorial information and geochemical data sets. However, they are not applicable for a laboratory where geochemical analyses take place because preservation of entire geochemical data sets including all associated information such as analytical method, calibration information, acquisition conditions, digital image etc. (i.e., metadata) is not within their scope.

At present, a rock depository named DREAM (Depository of References for Earth and Analytical Materials) is being constructing in our institute. It has become clear that (a) to organise the rocks in the depository, (b) to curate the rocks themselves, and (c) to archive any and all geochemical data sets, specialised software is required. Our needs have shown that the software should document the origin of a rock, its current storage location, the relationship, if any between the original rock and any subsamples, and correlate any and all geochemical data sets associated with the rock. To meet these needs, we have developed a software package called Medusa. It can be installed on a PC and used in any laboratory and should help geochemists curate materials independently from publication and online database systems. In this article, the system design and implementation are detailed.

System design

Record classes

The software correlates rocks and properties including sample (i.e., rock) collection location, current storage locations, genetic relations to the original rock, geochemical data sets obtained in a laboratory (e.g., elemental abundances, isotopic ratios) and bibliographies. The database schema applied in this study is consistent with one used by PerDB (Lehner et al. 2000). Any given rock is described in terms of a combination of records. In our system, there are six major record classes including sample, storage, locality, experiment, analysis and bibliography (Table 2). Each record consists of a written description as summarised in Table 3 (e.g., type of rock, building or shelf where it is located, sample location, etc.). For any rock, records that belong to different classes are linked by the system to one another as shown in Figure 1.

Creation of records and links with geochemical workflow

Original information for a rock is described by a combination of locality and sample records. When a rock is collected, a sample record is created and linked to a
Table 2. Major classes for records

<table>
<thead>
<tr>
<th>Class</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>A physical entity of a rock</td>
</tr>
<tr>
<td>Storage</td>
<td>A physical entity of a tenant (e.g., building, room, shelf, container)</td>
</tr>
<tr>
<td>Locality</td>
<td>Where a rock was sampled</td>
</tr>
<tr>
<td>Experiment</td>
<td>How a sample was synthesized (e.g., pressure, temperature, oxygen fugacity)</td>
</tr>
<tr>
<td>Analysis</td>
<td>An action to make an analysis from a rock</td>
</tr>
<tr>
<td>Chemistry</td>
<td>Chemical properties of a rock (e.g., element abundances, isotopic ratios)</td>
</tr>
<tr>
<td>Bibliography</td>
<td>A publication</td>
</tr>
</tbody>
</table>

Table 3. Record attributes for major classes: Common and class-specific attributes are shown. A link between records is established by a link attribute

<table>
<thead>
<tr>
<th>Class</th>
<th>Major attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>Link-storage, Link-locality, Link-experiment, Link-parent</td>
</tr>
<tr>
<td>Storage</td>
<td>Link-storage</td>
</tr>
<tr>
<td>Locality</td>
<td>Latitude, longitude, elevation</td>
</tr>
<tr>
<td>Experiment</td>
<td>Device, capsule, pressure, temperature, oxygen-fugacity, duration</td>
</tr>
<tr>
<td>Analysis</td>
<td>Device, technique, operator, Link-sample</td>
</tr>
<tr>
<td>Chemistry</td>
<td>Link-analysis, item-measured, value uncertainty, unit</td>
</tr>
<tr>
<td>Bibliography</td>
<td>Authors, title, journal, year, doi, volume, page, published date</td>
</tr>
</tbody>
</table>

![Figure 1. Record classes that describe a rock in the database system. A ‘foot’ symbol at the end of line represents multiple possible relationship. Sample and storage classes can have a self-to-self relationship. A bibliography is linked when data sets are published.](image)

If a rock is saved after collection and subsequent analysis, the original rock as well as any subsequent items will be stored in appropriate media and location. For example, a powder may be located in a plastic container on a shelf with other rock powders in a clean room. The records and links involved in this plot are shown in Figure 3. A 1 inch mount, often used for in situ beam analysis, is considered as storage because it holds tiny minerals with different origins.

Analyses are divided into two types, bulk and in situ. Analysis is defined as an action to obtain chemical properties from a portion of a rock, such as trace element determination of a solution by ICP-MS, or in situ analysis by microprobe. Chemical properties obtained from an analysis (e.g., elemental abundance, isotopic ratio) are stored individually in a chemistry record and linked to an analysis record. For an analysis that includes multiple elements (e.g., Rb, Sm and Eu), a chemistry record is created and linked to an analysis record for each element (Figure 4a). In the case of in situ analysis, an analysis record is linked to a point record. The point record is linked to an image file with its coordinates. In the case of a line profile, multiple point records are linked to an image file. All records and links involved in this line profile are shown in Figure 4b.

If and when data are published, a link from a bibliography to a sample record is established (Figure 1). A third-party researcher can access any and all information via the bibliography record if the system is open. When a bibliography documents multiple rocks, and/or multiple bibliographies refer to a single sample, all possible links between the bibliography and sample records are established (Figure 1).

**Interface**

The primary interface for the software is a web browser that operates on any platform. Most functions including search and editing of data sets are provided through this interface.

It is also important for users to be able to develop their own software that is better suited to their particular needs. A web-service API (Application Programming Interface) is the most general interface to enhance interoperability. If a web-service API is available, Matlab, popular software for scientific computing, can be a client using its built-in commands. Another possible application using the API is software for mobile devices. With a mobile device, a new record of a rock can be created at a field site with a photograph and the location obtained by GPS. Software...
that automatically transfers data from an instrument to the system can be developed, if necessary.

Implementation

The software that has been developed to meet the system design described above (Medusa) can be installed on any PC. An initial version of Medusa has been in operation in our laboratory for 3 years. Currently, it handles 100000 rocks, 1000 storages, 300 localities, 1200 analyses and 11000 attached files consuming 1.5 GB. A typical rock occupies nominally 1.5 MB including associated records and attached files. Thus, 1 million records could be stored in a database that had a capacity of 1.5 TB. In this section, implementation of Medusa is outlined. First, the system architecture is described followed by a discussion of several features including visualisation, access control and utilisation of barcodes. Source codes for Medusa are available at http://dream.misasa.okayama-u.ac.jp/.
Figure 4. Records and links that describe three elemental concentrations obtained by (a) four bulk analyses and (b) a line profile with four spots on a thin section.

Architecture

Medusa employs an open-source web-application-framework called Ruby on Rails (http://rubyonrails.org/). Medusa consists of a front-end, a back-end database and a core. The front-end provides interfaces for a web browser and a web-service API for client software. The back-end database stores records. The database has multiple tables, and each table corresponds to a record class. A record in a table has a unique key, and by tracing the keys, correlation between records is dynamically reconstructed. Attachment files (image and PDF files) are not stored in the database, only the paths to them.

Upon request from a client, the core interacts with the back-end database and returns data sets formatted to HTML, XML (Extensible Markup Language), CSV (Comma Separated Values) or PDF.

Visualisation on web interface

A rock collected from eastern Iceland (Kitagawa et al. 2008) visualised on the web interface is shown in Figure 5. After its collection, the rock was subsampled creating a powder and a thin section. The original rock, powder and thin section were then stored in a container. The interface demonstrates not only sampling sites, but also the existence of
daughters by tree navigation. Current location of the rock is shown on a map. The rock was processed in several different ways (e.g., powdered, thin sectioned, put into solution) and stored in a container with other samples. The history of all subsampling can be traced by a tree structure labelled ‘relatives’.

Access control

Medusa can limit access for each record. Each record has an owner and belongs to an associated group. A user is a member of at least one research group, and only a group member can see that group’s record. An important benefit of this is that a user who is not a member of a certain group does not have to see rocks that are of little interest to him. For example, information related to reference materials should be shared with everyone, but records of mantle xenoliths can be hidden from a lunar researcher. In other words, access control increases signal-to-noise ratio for users.

Registration and updating records

Using the web-service API, two important utilities were developed that (a) registers rocks, and (b) correlates records using a scanning device. Both are available with the main source code.

When numerous rocks are registered at once, the rate-limiting procedure is to create a link between an image file and a record. A desktop utility creates a new record and makes a link to an image file that is shown on the screen.

All records in Medusa have a unique ID that consists of thirty characters. A label with its unique ID printed on it should always be placed with whatever sample is in storage. We suggest having a 2D barcode on it. When a user makes new links between rocks and storage, accuracy and efficiency of the scanning device become significant. With this utility that correlates rock and storage, a new relationship is set by two scans.

Summary

Technical achievement

The software named Medusa was developed to facilitate the curation and accessibility of rocks and all
associated data. Medusa correlates geochemical data sets with a rock and documents the rock’s origin, current location and any genetic relationships with subsequent materials and/or data generated from the original rock. Medusa can help any laboratory perform curation and will be of great use for sharing analytical information with the public. Using a web browser, any user can access most functions on Medusa. A database system powered by Medusa has been operational for 3 years as a component of our rock depository and handles 10,000 samples.

Future directions

The next step for Medusa is to organise and document a web-service API. At the same time, utility programs need to be developed. For example, a data reduction program is currently under development. This program will download, format, reduce and plot data, and should be especially useful for reference materials.

We also plan to develop an adapter for a community-driven web interface EarthChem (Walker et al. 2005). EarthChem distributes queries to major database systems such as GEOROC, GeoReM, PetDB, SedDB and NAVDIA. The adapter for EarthChem will make a link between Medusa and the community.

Intellectual merit

Anything we observe is an integration of phenomena in nature. A rock is a complex assemblage of materials originating by different processes from a variety of sources. To understand the origin, evolution and dynamics of the Earth and the Solar System, solving the multiple ‘equations’ stored within rocks are essential.

Even with the recognised importance of multidimensional investigations, intrinsic uncertainties inhibit interaction between analyses, approaches and scientists. We believe that Medusa works as a catalyst to enhance interaction and safeguards rock traceability and accessibility, for future analytical innovation. Interaction between specialists leads to comprehensiveness, and inter-operability and continuity promote understanding.

Acknowledgements

We are greatly indebted to J. Braphy, T. Tsujimori, D. Rumble III, and M. Kanzaki for numerous comments, which improved the manuscript. We thank T. Marigut, R. Tanaka, T. Ota, K. Kobayashi, and other PML members for maintaining the rock repository. This work was financially supported by the 21st Century Center Of Excellence (COE) Programme (EN) and Grants-in-Aid for Young Scientists (20740353) to TK from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References


references


Kennett J. and Stott L. (1991)


The magnetic information Consortium (MAGIC) online database: Uploading, searching and visualizing, palaeomagnetic and rock magnetic data. EOS Transactions AGU. 84, Fall Meeting Supplement, Abstract GP13A-0088.

A global geochemical database structure for rocks. Geochemistry Geophysics Geosystems. 1, 1012.

Geochemistry of oceanic island-arc and active continental margin volcanic suites. Some statistical evaluations and implications using the database GEOROC-FOS Transactions AGU. 83, Fall Meeting Supplement, Abstract V62B-1401.


EarthChem: International collaboration for solid Earth geochemistry in geoinformatics EOS Transactions AGU. 86, Fall Meeting Supplement, Abstract IN44A/03.