Geosci. Model Dev., 18, 4455–4467, 2025
https://doi.org/10.5194/gmd-18-4455-2025

Geoscientific Model Development
Open Access
EGU

The OpenMindat v1.0.0 R package: a machine interface to Mindat open data to facilitate data-intensive geoscience discoveries

Xiang Que[1], **Jiyin Zhang**[1], **Weilin Chen**[1], **Jolyon Ralph**[2], **and Xiaogang Ma**[1]

[1]Department of Computer Science, University of Idaho, Moscow, ID 83844, USA
[2]Hudson Institute of Mineralogy, Keswick, VA 22947, USA

Correspondence: Xiaogang Ma (max@uidaho.edu)

Received: 16 April 2024 – Discussion started: 11 June 2024
Revised: 31 January 2025 – Accepted: 8 May 2025 – Published: 23 July 2025

Abstract. Technologies such as machine learning and deep learning are powering the discovery of meaningful patterns in Earth science big data. In the field of mineralogy, Mindat ("mindat.org") is one of the largest databases. Although its front-end website is open and free, a machine interface for bulk data query and download had never been set up before 2022. Through a project called OpenMindat, an application programming interface (API) to enable open data query and access from Mindat was set up in 2023. To further lower the barrier between Mindat open data and geoscientists with limited coding skills, we developed an R package (OpenMindat v1.0.0) on top of the API. The Mindat API includes multiple data subjects such as geomaterials (e.g., rocks, minerals, synonyms, variety, mixture, and commodity), localities, and the IMA-approved (International Mineralogical Association) mineral list. The OpenMindat v1.0.0 package wraps the capabilities of the Mindat API and is designed to be user-friendly and extensible. In addition to providing functions for querying data subjects on the API, the package supports exporting data to various formats. In real-world applications, these functions only require minor coding for users to get desired datasets, and various other packages in the R environment can be used to analyze and visualize the data. The OpenMindat v1.0.0 package, which includes detailed tutorials and examples, is available on GitHub under the MIT license. The field of mineralogy and many other geoscience disciplines are facing opportunities enabled by open data. Various research topics such as mineral network analysis, mineral association rule mining, mineral ecology, mineral evolution, and critical minerals have already benefited from Mindat's open data efforts in recent years. We hope this R package can help accelerate those data-intensive studies and lead to more scientific discoveries.

1 Introduction

As machine learning and deep learning techniques thrive on their ability to discover complex patterns, data-driven geoscience studies yield increasingly more exciting results (Hazen et al., 2011; Bergen et al., 2019; Reichstein et al., 2019; Que et al., 2024). However, due to the complex and multifaceted nature of Earth's processes, high-quality data are required to enable the capacity of quantitative methods to make informed predictions across varying contexts (Chen et al., 2023). Open access to large and diverse datasets is imperative for data-driven geosciences and calls for attention and action (Hossain et al., 2016). Regarding the field of mineralogy, minerals provide many essential clues for exploring the complex geological history of the Earth and other planetary bodies (Hazen et al., 2019; Prabhu et al., 2021). A rapidly growing volume of mineralogical and geochemical data resources are available for research, such as the IMA (International Mineralogical Association) list of mineral species (https://rruff.info/ima, last access: 17 July 2025) (Prabhu et al., 2023), Mindat (https://www.mindat.org, last access: 17 July 2025) (Ralph et al., 2022; Ma et al., 2024), RRUFF (https://rruff.info, last access: 17 July 2025) (Yang et al., 2011), EarthChem (https://www.earthchem.org, last access: 17 July 2025) (Walker et al., 2005; Lehnert et al., 2007), the Evolutionary System of Mineralogy Database (ESMD; https://odr.io/esmd, last access: 17 July 2025) (Chiama et al., 2023), the Mineral Proper-

ties Database (https://odr.io/MPD, last access: 17 July 2025) (Morrison et al., 2023), and the Astromaterials Data System (https://astromat.org, last access: 17 July 2025) (Chamberlain et al., 2021). Thanks to these big and expanding open datasets, new scientific topics such as mineral evolution (Hazen et al., 2008, 2014), mineral ecology (Hazen et al., 2015), and mineral informatics (Prabhu et al., 2023) are emerging and developing quickly. Among those data sources, the Mindat, a crowd-sourced and expert-curated database that started running in 2000, is now one of the world's most widely used online databases for minerals and their distributions. By August 2023, Mindat had recorded 5960 minerals, 395 558 localities, 1 503 650 occurrences, and 1 291 077 photos, with a total data volume exceeding 25.8 TB (Ralph et al., 2022), and the records are actively expanding and updating.

Mindat is widely used by many individuals and communities. In 2021 alone, the Mindat website received 44 333 302 views from 10 148 136 unique visitors, and, as of August 2023, the number of registered users reached 72 488. The Mindat team provides a website portal (https://www.mindat.org/advanced_search.php, last access: 17 July 2025) for users to retrieve data by specifying parameters interactively. Although its website has always been open for searching and browsing datasets, a machine interface for Mindat data querying and downloading had never been fully established before 2023. Moreover, multiple constraints on the website require multiple interactions to be performed, and some pages cannot load all the filtered data records at once (due to the size of the data that meet the constraints) or cannot display them efficiently (e.g., in sorted order). In the past years, many researchers have reached out to the Mindat technical team requesting bulk datasets on certain topics, and those requests could only be addressed on a tedious case-by-case basis. To address the challenge, the OpenMindat project (Ma et al., 2024) was set up recently to implement a fully open access, machine-readable, and interoperable architecture for Mindat. Following the FAIR principles (i.e., findable, accessible, interoperable, and reusable) (Wilkinson et al., 2016), a roadmap of OpenMindat was laid out, including the technical approaches to upgrade and reuse existing data resources, tools, and infrastructure. In the spring of 2023, the preliminary RESTful API (Application Programming Interface) (Richardson and Ruby, 2008) of Mindat was established, which any registered users can access with an authorized API token (Zhang, 2024). While the API (https://api.mindat.org, last access: 17 July 2025) provides a structured and stable channel to Mindat open data (Zhang et al., 2024), users need to know the data subjects available in the API and the parameters of each data subject and have moderate coding skills to construct the commands for data retrieval. To further lower the barrier to accessing Mindat open data, we are constructing R software packages on top of the API. Such packages have several advantages. First, they wrap the capability of the API in a variety of functions for which users only need minimal coding to retrieve datasets of interest. Second, the data querying is fast, and the results can be returned in specified formats. Third, the packages can be easily integrated in workflow platforms such as R Studio and Jupyter, where many other packages can be used together for data analysis.

This paper presents our design and implementation of the R software package OpenMindat v1.0.0 to meet users' needs for quick and easy access to Mindat's open data. The package is open source for anyone to reuse, and we welcome feedback on improvement and extension.

2 Architecture of the OpenMindat R package

The primary objective of the OpenMindat R package is to provide an implementation mechanism to translate users' data requirements into Mindat API requests. Mindat datasets, especially those made machine-readable through the Mindat RESTful API, are structured records stored in a relational MySQL database. Table 1 lists the primary data subjects stored in Mindat and the number of their records.

The API server manages web requests for datasets. Currently, it provides a separate access endpoint for each data subject. While in the future, as the number of subjects and records in the crowdsourced Mindat database increases, the API server may open or update its data access endpoints. Some new features of our package will be updated according to the endpoints provided by the server. Accordingly, we designed an architecture (Fig. 1) to connect the user's data needs with the Mindat API server. (For a more detailed technical diagram of the architecture, please refer to the online documentation on our GitHub repository. Links are given in the "Code and data availability" section.) From our survey and interactions with geoscientists in the past years, most users' data requirements fall into the following categories: (1) queries about geomaterials (i.e., mineral, rock, commodity, and other natural geological materials). Users need to filter the geomaterials based on their physical properties (e.g., density, hardness, color, refractive index, and crystal structure), their chemical properties (e.g., element inclusion and exclusion states), or their entry types (e.g., synonym, variety, rock, mixture, mineral, and series). (2) There are queries about localities. Mindat's localities record textual addresses, coordinates, area boundaries, and other relevant attributes. They follow a specific hierarchical structure and naming rule (https://www.mindat.org/a/localityhierarchies, last access: 17 July 2025). A simple explanation is that the number of locality levels indicates the level of detail in the locality hierarchy. Larger values indicate more details. Thus, 0 indicates the top level, usually representing a country, a region, or a tectonic plate. Users may need to query the localities based on their number of levels, attributes (e.g., name, country, ID, description, and longitude/latitude), coordinate information, or geological ages. For example, studies of the mineral evolution (Hazen and Ferry, 2010; Hystad et al., 2019) and the coevolution of the geosphere and the biosphere (Hazen et al.,

Table 1. Primary data and the number of records stored in the Mindat database.

Data subjects	Brief description	Number
Mineral species	The official list of approved mineral species including their names, localities, and occurrences.	> 5800
Alternative mineral names	Alternative names that are not officially IMA-approved mineral species, including varieties, mixtures, and synonyms.	> 45 000
Localities	General information about a locality, which may include latitude and longitude or any other relevant information about the locality.	> 390 000
Occurrences	The links between the mineral data and the locality where they occur.	> 1.2 million
Photographs	Mineral photo	> 1.1 million
Mindat ID	Identifier for a mineral or related material (e.g., rock or mixture) in the mindat.org database	> 10.3 million
Locality age	The age of a mineral occurrence and its locality.	> 5500
Meteorites	A meteorite is a stony or metallic body that has fallen to the Earth's surface (or to any other planetary body on which it is found) from outer space.	> 1500

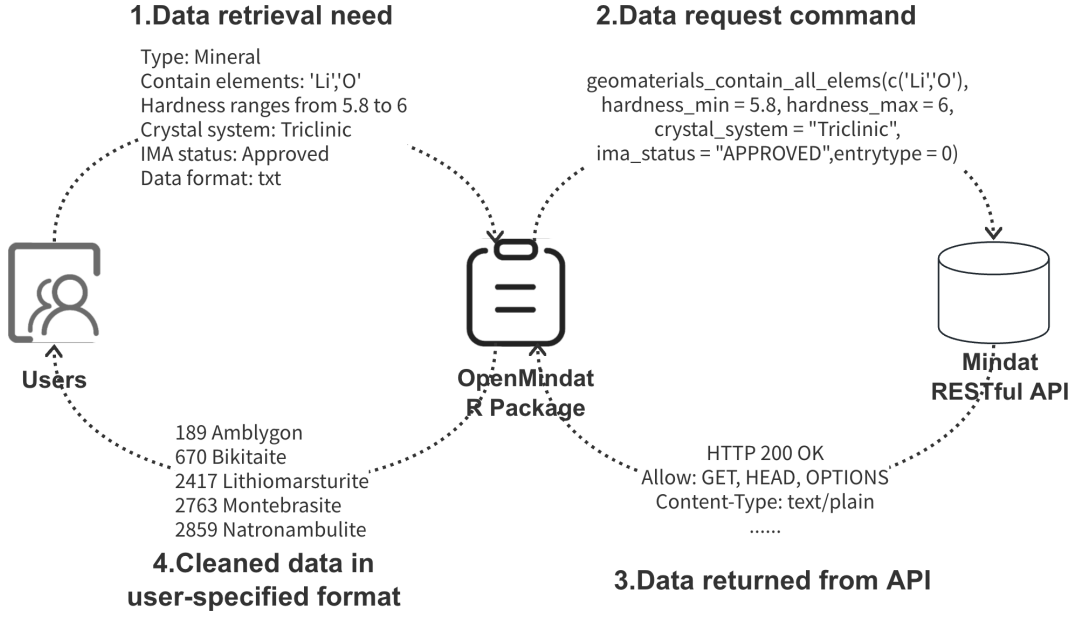

Figure 1. Architecture of the OpenMindat R package.

2014; Hazen and Morrison, 2020) ignited the need to retrieve localities with geological ages. (3) There are queries about IMA-approved minerals. IMA promotes the science of mineralogy and standardizes the nomenclature of mineral species. The IMA mineral list is updated frequently, and it is common to query mineral information by specific IMA status, such as A (approved), G (grandfathered), Rd (redefined), and Q (questionable). (4) There are data format needs. Some applications or analyses, including mineral association rule analysis (Morrison et al., 2023) and mineral network analysis (Liu et al., 2018; Morrison et al., 2020), require outputting filtered data in specified formats, such as CSV, TXT, TTL, and JSON-LD, for different use cases.

Following the designed architecture and user need analysis, we developed about 100 functions in the R package, and they are grouped into several classes. (To view all the functions, execute "help (package = OpenMindat)" or see the reference manual listed in Table 10.) Table 2 lists the main classes and functions related to data subjects and formats: (1) the geomaterials class, which is one of the main data subjects supported by Mindat API, includes sub-subjects of minerals, synonyms, varieties, mixtures, series, group

Table 2. A partial list of classes and functions in the OpenMindat R package.

Class	Functions	Brief description
Geomaterials	geomaterials_contain_any_elems	Retrieves geomaterials that contain any of the specified elements.
	geomaterials_cleavagetype	Retrieves geomaterials that match the specified cleavage type.
	geomaterials_color	Retrieves geomaterials that match the specified colors.
	geomaterials_crystal_system	Retrieves geomaterials that match the specified crystal system.
	geomaterials_bi_greater_than	Retrieves geomaterials that have higher birefringence than an input value.
	geomaterials_dens_range	Retrieves geomaterials that match the density within a given range.
	geomaterials_diapheny	Retrieves geomaterials that match the given diapheny.
Localities	localities_list_country	Retrieves the localities list that is found in a specified country.
	localities_list_elems_inc	Retrieves the localities that contain the given elements.
	localities_list_description	Retrieves the localities that contain the given description.
IMA minerals	minerals_ima_list	Retrieves the whole IMA mineral list.
	minerals_ima_list_ima	Retrieves IMA mineral lists with given authorization statuses.
	minerals_ima_retrieve	Retrieves IMA minerals with given IDs.
Data maker	saveMindatDataAs	Saves the data frame to a specified file format.
	ConvertDF2JsonLD	Converts the retrieved data frame into a JSON-LD format string.
	ConvertDF2TTL	Converts the retrieved data frame into a TTL format string.

lists, polytypes, rocks, and commodities. (A complete list of classes and functions is available via our online documentation on GitHub.) The current geomaterial record contains 146 attributes, including descriptions of physical properties, chemical information, optical properties, crystal structure information. Class (2) is the localities class. It consists of 37 attributes, including longitude, latitude, coordinate system, link, and area, which describe the information of textual address, coordinate point position, locality type boundary polygon, and occurrences. Class (3) is the IMA mineral class. This class is mainly for retrieving and managing IMA-approved mineral species names, chemical formulas, authorization status, and other attributes. Class (4) is the data maker class. It is for data format conversions and outputs. It can convert R data frames to required formats such as CSV, TXT, TTL, and JSON-LD.

These classes and functions can be applied flexibly to meet users' specific data needs. The geomaterials class provides functions that help easily filter records by the following relationships: "contains any", "contains all", "contains only", "does not contain", "contains all but not", and "contains any but not". It also provides functions to filter records by specifying physical properties, including, but not limited to, density, hardness, birefringence, optical 2v, crystal system, fracture type, color, streak, diaphaneity, luster type, optical sign, optical type, poly type, cleavage type, and tenacity. For some physical properties with numerical values or threshold ranges (e.g., Mohs scale (Broz et al., 2006) and density), it supports filtering records by relationships such as "greater than", "less than", and "within a given range". For other non-numerical physical properties, it provides functions for retrieving data records by specifying strings, enumeration variables, and special symbols. It also provides functions to retrieve geomaterial records based on wildcard names, non-null fields, Mindat IDs, mineral varieties, and more. The localities class also provides functions for retrieving records by specifying the chemical elements' inclusion and exclusion relationships. It can support filtering locality records by level, country name, Mindat ID, description, etc. The "age_id" attribute of the Mindat locality, if not null, shows a unique identifier that can be associated with a locality age record. This record contains geological time information about the locality. The IMA mineral class provides functions to retrieve IMA mineral records. It helps retrieve the complete list of IMA-authorized mineral names, including their chemical formulas, description information, and other related data. Users can also retrieve data by specifying the minerals' approved status or ID. The data maker class provides functions to help export the retrieved records into the required format. All the provided functions support the expansion of the input parameter, which enables data retrieval based on combined properties. Some examples will be presented in the next section.

3 Examples and results

3.1 Geomaterial data retrieval

To illustrate the capabilities of retrieving geomaterial records, Table 3 lists some basic use cases and their descriptions. In the list, each use case only involves the simple usage of one function from the R package.

In some other situations, even just one function can achieve a relatively heavy task. The code below demonstrates two such tasks: one is to retrieve a hierarchical taxonomy of

Table 3. Geomaterial data retrieval use cases. Last access for all URLs: 17 July 2025.

Function category	Function name and input	Output and description	Demo codes and results
Chemical elements' inclusion and exclusion relationships	geomaterials_contain_any_but_not_elem(c('Fe','S'), c('O'))	Records of geomaterials containing Fe and S but not O.	https://github.com/quexiang/OpenMindat/blob/main/notebook/Retriev_Geomaterials_by_elements.ipynb
	geomaterials_not_contain_elems(c('Fe','S','O'),fields = "id, name,mindat_formula,elements")	Records of geomaterials without Fe, S, and O that contain only the following fields: ID, name, mindat_formula, and elements.	
Physical properties with numerical values or threshold ranges	geomaterials_hardness_gt (9)	Records of geomaterials with a Mohs hardness greater than 9.	https://github.com/quexiang/OpenMindat/blob/main/notebook/Retrieve_Geomaterials_by_physical_prop_1.ipynb
	geomaterials_dens_range(3,3.2)	Records of geomaterials with a density ranging from 3 to 3.2.	
Non-numerical physical properties	geomaterials_color(c("bright blue"))	Records of geomaterials that have a bright-blue color.	https://github.com/quexiang/OpenMindat/blob/main/notebook/Retrieve_Geomaterials_by_physical_prop_2.ipynb
	geomaterials_cleavagetype (c("Poor/Indistinct"))	Records of geomaterials with a cleavage type of "Poor/Indistinct".	
Wildcard names, non-null fields, etc.	geomaterials_name("_u_r_z")	Records of geomaterials whose names have six characters where the second, fourth, and sixth characters were specified.	https://github.com/quexiang/OpenMindat/blob/main/notebook/Retrieve_Geomaterials_by_wildcard_names.ipynb
	geomaterials_name("qu*")	Records of geomaterials whose names have the first two characters "q" and "u".	
	geomaterials_field_exists ("meteoritical_code",TRUE)	Records of geomaterials whose "meteoritical_code" fields have non-null values.	
	geomaterials_varietyof(3337)	Records of geomaterials that were varieties of quartz (3337 is the Mindat ID of quartz).	

petrological names and their definitions (e.g., get the rock hierarchy information), and the other is to list mineral species containing nickel or cobalt with sulfur but without oxygen, which was discussed in Ma et al. (2024) as a typical use case.

R > mindat_geomaterial_list(ids = c (''), entrytype = 7, fields = c ("name", "description_short", "rock_parent", "rock_parent2"))

R > unique(rbind(geomaterials_contain_all_but_not_elems(c("Ni", "S"),c('O')),geomaterials_contain_all_but_not_elems(c("Ni", "S"), c('O'))))

With the appropriate combination of properties, a single function can also support complex data request needs. For example, if we need to retrieve records of IMA-approved minerals containing lithium (Li) and oxygen (O) elements, having Mohs hardness levels between 5.8 and 6, and having a triclinic crystal structure, we can use the following code. Table 4 shows the returned geomaterial records.

R > geomaterials_contain_all_elems(c('Li','O'), hardness_min = 5.8, hardness_max = 6, crystal_system = "Triclinic", ima_status = "APPROVED", entrytype = 0)

3.2 Locality data retrieval

Users can apply the package to retrieve locality data as needed. Table 5 lists some use cases. Once the locality dataset is retrieved, many other third-party packages and functions in the R environment can be leveraged in data visualization and analysis. For example, we can use a map window to view the spatial distribution of minerals containing certain elements (e.g., As (arsenic)) or geomaterials containing certain literal descriptions (e.g., volcano) (Fig. 2).

3.3 IMA mineral list retrieval

This package can support retrieval of records according to their IMA status. Table 6 lists some basic use cases. We can also use some other functions to validate alternative mineral/rock names. For example, if the name "amethyst" is input, the program would return that the correct mineral species is "quartz" and that "amethyst" is a varietal name (Ma et al., 2024). We can use the following code to realize that need.

R > df_gm_amethyst <− geomaterials_name("Amethyst")

R > mindat_geomaterial_list(ids = c(df_gm_amethyst$varietyof), entrytype = 0, ima_status = "APPROVED")

The functions in the OpenMindat package can be used together with many other packages and functions in the R environment to achieve data exploration or analysis needs, and many of them require just a few lines of code. For example, we can retrieve and visualize the top 10 IMA-approved mineral species (by occurrence count) found in a country,

Table 4. Results of geomaterial records retrieved by combined properties.

Records (only some relevant fields are shown)						
Name	Elements	Hmin	Hmax	Csystem	Ima_status	Entry type
Amblygonite	"Al", "Li", "O", "P", and "F"	5.5	6	Triclinic	"APPROVED", "GRANDFATHERED"	0
Bikitaite	"Al", "Li", "Si", "O", and "H"	6.0	6	Triclinic	"APPROVED", "GRANDFATHERED"	0
Lithiomarsturite	"Ca", "Li", "Mn", "Si", "O", and "H"	6.0	6	Triclinic	"APPROVED"	0
Montebrasite	"Al", "Li", "O", and "H"	5.5	6	Triclinic	"APPROVED", "GRANDFATHERED"	0
Natronambulite	"Na", "Li", "Mn", "Ca", "O", and "H"	5.5	6	Triclinic	"APPROVED"	0

The code and results shared on GitHub: https://github.com/quexiang/OpenMindat/blob/main/notebook/Retrieve_Geomaterials_by_combined_conds.ipynb (last access: 17 July 2025).

Table 5. Locality data retrieval use cases. Last access for all URLs: 17 July 2025.

Function category	Function name and input	Output and description	Demo codes and results
Name, ID, and description	localities_list_country ("China")	Records of localities in China	https://github.com/quexiang/OpenMindat/blob/main/notebook/Retrieve_Localities_by_desc.ipynb
	localities_retrieve_id(id + 22)	Records of localities in Algeria (The number 22 is the locality ID of Algeria.)	
	localities_list_description("volcano")	Records of localities with descriptions containing the word "volcano"	
Chemical elements' inclusion and exclusion relationships	localities_list_elems_inc(c("Dy"))	Records of localities that contain the dysprosium (Dy) element	https://github.com/quexiang/OpenMindat/blob/main/notebook/Retrieve_Localities_by_elems.ipynb
	localities_list_elems_inc_exc(c("Dy"), c("Li"))	Records of localities containing dysprosium (Dy) but no lithium (Li)	
Locality age	locality_age_list()	All records of locality age	
	locality_age(id = 60)	Records of locality age with locality ID 60	

such as Canada (Fig. 3). To achieve that, we need to perform the following steps: (1) execute the OpenMindat function "localities_list_country("Canada",expand = "∼all")" to retrieve the list of localities in Canada and the lists of geomaterials recorded in each locality. (2) Summarize the number of occurrences of each geomaterial ID and sort in descending order. (3) Check each geomaterial ID to see if it is an IMA-approved mineral and, if so, retrieve the corresponding record by using the "minerals_ima_retrieve" function. The code and results of this example are shared on GitHub: https://github.com/quexiang/OpenMindat/blob/main/notebook/Top10_IMA-Approved%20Minerals%20in%20a%20specified%20country(e.g.%20Canada).ipynb (last access: 17 July 2025).

3.4 Output the retrieved data into different formats

Users can output their retrieved data in a specified format such as CSV, JSON, TXT, JSON-LD, and TTL. The function "saveMindatDataAs" will identify the suffix of the input file name and convert the retrieved R data frame into a corresponding format. For the data conversion to the JSON-LD and TTL formats, two Microsoft Excel template files (i.e., OpenMindat_Schema_JSON-LD.xlsx and OpenMindat_Schema_TTL.xlsx) are required. Users can configure their settings in the Microsoft Excel template to customize files that meet their needs for the output. The default versions can be accessed via https://github.com/quexiang/OpenMindat/tree/main/inst/extdata (last access: 17 July 2025). Here, we take the JSON-LD template as an example to briefly introduce its basic settings (similar to template settings in the TTL format). There are two sheets in the template file; the first one is for the context settings, and the other one is for the field setting. Table 7 (i.e., the first sheet) shows the names of all schemas and how their corresponding URLs are configured. Table 8 shows the second sheet where the field named "fields" records the field names that need output corresponding to the Mindat API. The field named "ref_fields" records the output field name list of JSON-LD. The field named"context_name" records all schema names corresponding to the field. The field named "type" records the type of schema to which the field belongs. All the values of the three fields are in the form of a list, separated by commas. Besides these fields, the field named "ref_field_num" indicates which name is to be output in JSON-LD (e.g., 1 represents the name before the first comma in "ref_fields").

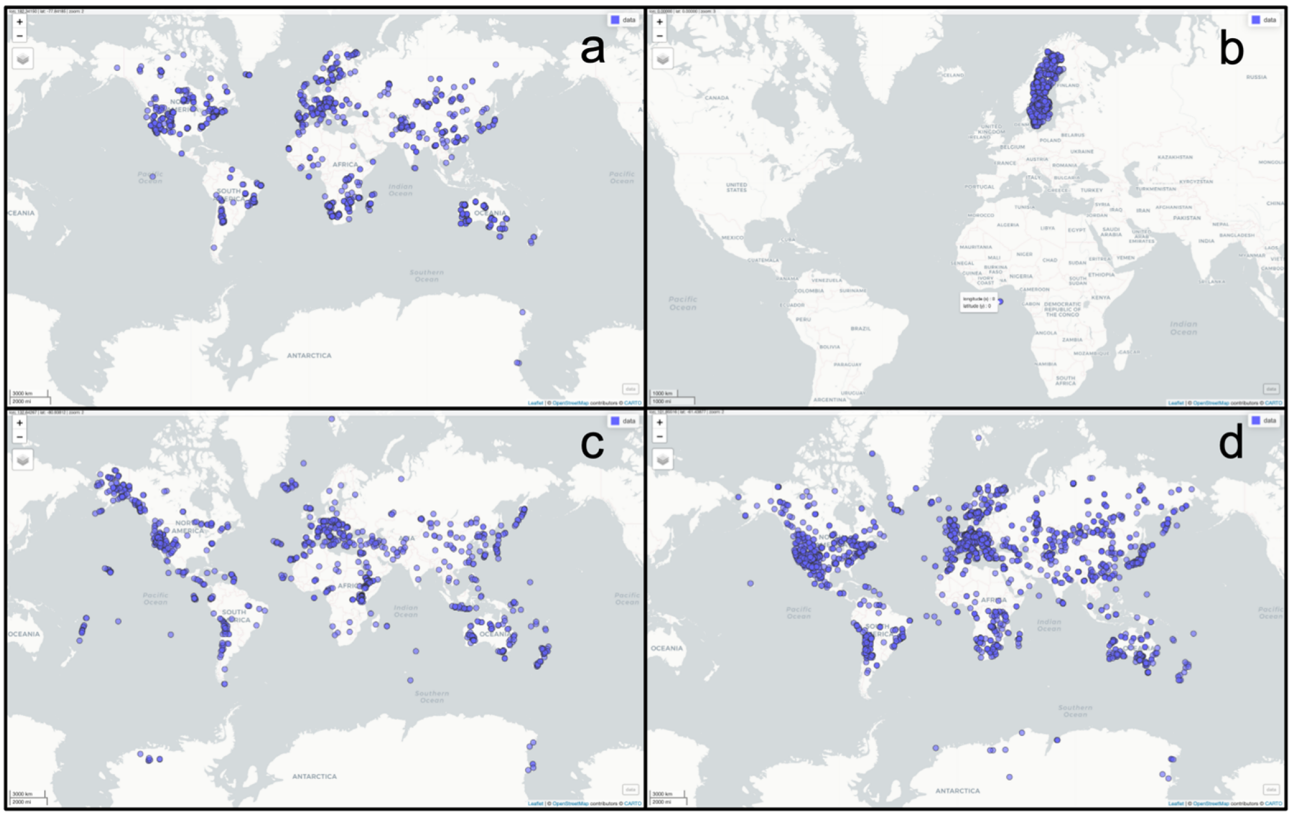

Figure 2. Mapping locality records retrieved by the OpenMindat R package: **(a)** As-containing minerals, **(b)** localities in Sweden, **(c)** locality descriptions containing "volcano", and **(d)** type localities of IMA-approved minerals. Base map © 2024 OpenStreetMap contributors . Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

Table 6. Use cases of IMA mineral list retrieval.

Function category	Function name and input	Output and description	Demo codes and results
IMA mineral list, IMA status, and ID	minerals_ima_list()	Records of all IMA-approved mineral species with detailed properties	https://github.com/quexiang/OpenMindat/blob/main/notebook/IMA_minerals.ipynb (last access: 17 July 2025)
	minerals_ima_list_ima(1)	Records of minerals that have IMA-approved status as Approved	
	minerals_ima_retrieve(2)	Records of abenakiite-(Ce) (The number "2" is the Mindat ID of abenakiite-(Ce).)	

Table 7. Context settings of the JSON-LD template. Last access for all URLs: 17 July 2025.

Context_name	Context_url
Mindat	https://mindat.org/
Schema	https://schema.org/
Gsog	https://w3id.org/gso/geology/

With the above configuration, we can obtain the exported file shown in Table 9 by executing the following code.

R > library(readxl)

R s > saveMindatDataAs(geomaterials_hardness_gt(9.8, fields = "id,longid,name ,ima_formula"),"df_geomaterials.jsonld")

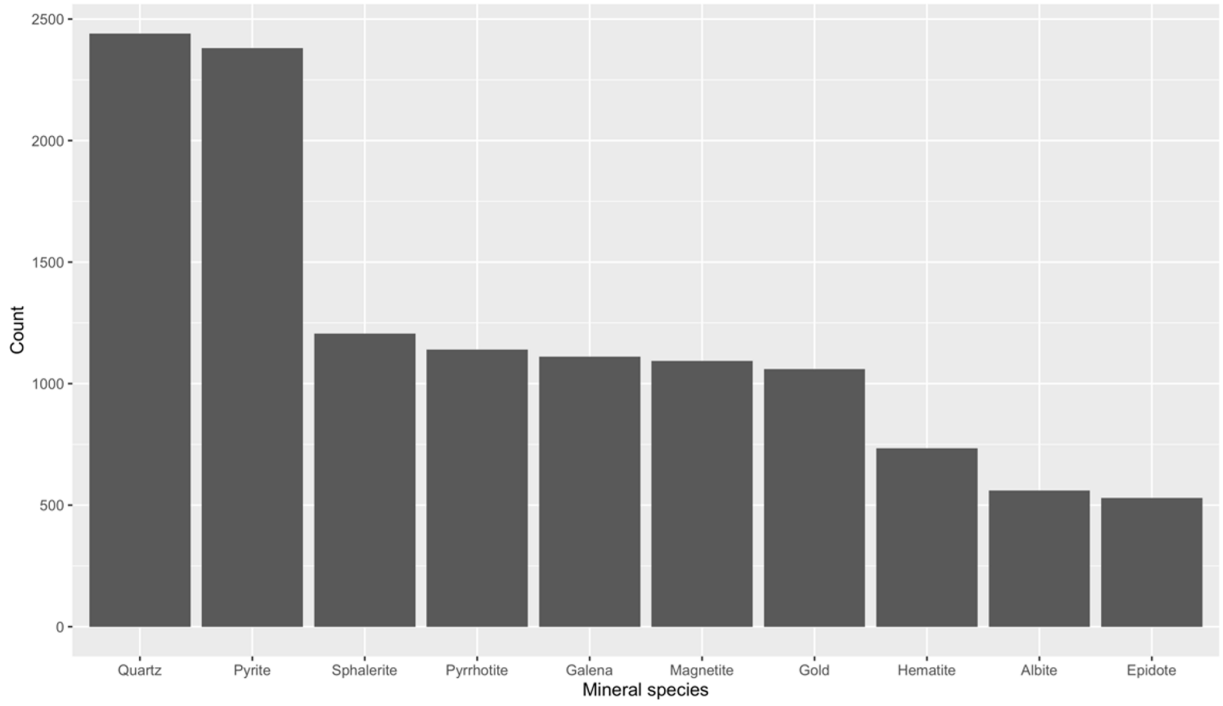

Figure 3. Top 10 IMA-approved mineral species found in Canada.

Table 8. Field settings of the JSON-LD template.

Fields	Ref_fields	Context_name	Type	Ref_field_num
id	mindat:id, ,	mindat,schema,gsog	mindat:Geomaterials,schema:Dataset,gsog:Mineral_Material	1
longid	identifier, ,	mindat,schema,gsog	mindat:Geomaterials,schema:Dataset,gsog:Mineral_Material	1
name	mindat:name, ,	mindat,schema,gsog	mindat:Geomaterials,schema:Dataset,gsog:Mineral_Material	1
ima_formula	mindat:ima_formula, ,	mindat,schema,gsog	mindat:Geomaterials,schema:Dataset,gsog:Mineral_Material	1

The full JSON-LD template is shared on GitHub: https://github.com/quexiang/OpenMindat/blob/main/inst/extdata/OpenMindat_Schema_JSON-LD.xlsx (last access: 17 July 2025).

3.5 Package releases, scientific applications, and updates

The OpenMindat R package and its source code were shared on GitHub (https://github.com/quexiang/OpenMindat, last access: 17 July 2025) together with detailed tutorials on how to install and run the package in the R environment (https://quexiang.github.io/OpenMindat, last access: 17 July 2025). The first version of this package (version 1.0.0) was also released in the comprehensive R archive network (CRAN) (Hornik, 2012) (https://cran.r-project.org/web/packages/OpenMindat, last access: 17 July 2025) on 15 February 2024. Scientists can use those functions flexibly to conduct scientifically meaningful data queries and access tasks. A big advantage of using this package is that it reduces the scientists' coding efforts; i.e., with relatively minor coding, they can retrieve a specific piece of data from the Mindat API. A list of examples and their Jupyter Notebook files (https://github.com/quexiang/OpenMindat/tree/main/notebook, last access: 17 July 2025), including those shown in the previous section, is also shared to demonstrate the functions and parameters for data query and access from the Mindat API. Readers can also refer to the "Code and data availability" section at the end of the article for a structured list of web links to all resources mentioned above.

4 Discussion

The development of the OpenMindat R package provides geoscientists with a user-friendly, efficient, and reproducible data querying tool for accessing and analyzing mineralogical data from Mindat. By wrapping the capabilities of the Mindat API into structured functions, the package overcomes barriers faced by researchers working with large-scale datasets. One of the primary advantages of the OpenMindat R package is its ability to simplify data access for geoscientists. Previously, obtaining bulk data from Mindat required manual interactions with the web page or complex API queries that

Table 9. Output file in JSON-LD format.

df_geomaterials.jsonld


```
{
"@context": {
        "mindat":"https://mindat.org/",
        "schema":"https://schema.org/",
        "gsog":"https://w3id.org/gso/geology/"}
"@graph": [{"@type": ["mindat:Geomaterials", "schema:Dataset", gsog:Mineral_Material "],
            "mindat:id":"1282",
            "identifier":"1:1:1282:5",
            "mindat:name":"Diamond",
            "mindat:ima_formula":"C"
        }
        , {"@type": ["mindat:Geomaterials", "schema:Dataset", gsog:Mineral_Material "],
            "mindat:id":"43792",
            "identifier":"1:1:43792:7",
            "mindat:name":"Qingsongite",
            "mindat:ima_formula":"BN"
        },
        {"@type": ["mindat:Geomaterials", "schema:Dataset", gsog:Mineral_Material"],
            "mindat:id":"52913",
            "identifier":"1:1:52913:0",
            "mindat:name":"Uakitite",
            "mindat:ima_formula":"VN"
        }
        ]
}
```


The code and results are shared on GitHub: https://github.com/quexiang/OpenMindat/blob/main/notebook/Output_DF2File.ipynb (last access: 17 July 2025).

demanded advanced coding skills. The package eliminates these obstacles by providing predefined functions centered on data subjects such as geomaterials and localities, enabling users to retrieve datasets with minimal effort. This accessibility is particularly beneficial for geoscientists who may not have extensive programming skills but require large datasets to drive their research. The package's ability to export data in multiple formats ensures compatibility with various analytical workflows. These formats are widely used across disciplines, allowing researchers to seamlessly integrate Mindat data into existing pipelines for visualization, statistical modeling, and geospatial analysis.

The OpenMindat R package embodies the principles of findable, accessible, interoperable, and reusable (FAIR) data. It is worth noting that data can be FAIR but not open. The "A" in FAIR stands for "accessible under well-defined conditions," meaning the data do not have to be freely accessible to everyone (Jeffery, 2021). So, it is commendable that Mindat data are open and free; users can access the data by registering with Mindat to get free tokens. By providing an intuitive interface to the Mindat API, the package ensures that mineralogical data are not only accessible but also easily integrated into diverse analytical workflows. This openness and the alignment with FAIR principles foster a culture of collaboration in the geosciences, where shared resources and tools can accelerate innovation. Reproducibility is a cornerstone of scientific research, as the concept of open science is increasingly demanded in the global geoscience community. The OpenMindat R package enhances this by allowing users to embed data retrieval processes directly into their R scripts. By automating the translation of user queries into API requests, the package ensures that data retrieval steps are transparent and replicable. This transparency not only strengthens the reliability of results but also facilitates collaboration among researchers. Shared R scripts or R Markdown documents can precisely reproduce datasets, fostering greater trust in geoscientific analyses. Accordingly, we envision the Mindat open data API and the R package as catalysts for data-driven discoveries in mineralogy and many other related geoscience disciplines. By providing a structured and efficient interface to the Mindat database, the package empowers researchers to explore complex relationships within mineralogical data. Moreover, the package's integration with R's extensive suite of analytical tools enables advanced applications such as network analysis, clustering, and predictive modeling. Researchers studying critical minerals, for instance, can use the package to analyze the geographic

and paragenetic distributions of these resources, supporting strategies for sustainable extraction and utilization.

The Mindat open data API is maintained by the Mindat technical team. They review and permit user registration requests, monitor the status of the server, and defend cyberattacks or malicious mass downloads. For individual researchers, the default API usage limit is 1000 requests per hour. Based on our experience in the past 2 years, this usage limit should be enough to meet the needs of most people. Specific users who need more frequent and larger data access can contact the Mindat technical team for permission. The Mindat technical team applied a hardware upgrade to the server in early 2025, which further stabilized the API. It is also noteworthy that the computational efficiency of the OpenMindat R package reduces the time and effort required for data retrieval and processing on the server side. By leveraging the API's pagination capabilities, the package ensures smooth handling of large datasets without overloading system memory. The caching mechanism further enhances efficiency by minimizing redundant queries, a critical feature for workflows involving iterative analyses. Scalability is another key strength. As geoscientific studies grow increasingly data-intensive, the ability to handle complex, multi-condition queries becomes necessary. The package's flexibility to combine various conditions, such as element inclusion, locality attributes, and IMA status, enables users to conduct sophisticated analyses tailored to specific research questions.

Looking into the future, we are confident about the broad variety of scientific applications enabled by the Mindat API and the OpenMindat R package. In mineral evolution studies, for example, Hazen et al. (2008, 2014), the package can facilitate analyses of temporal and spatial patterns in mineral diversity, shedding light on the coevolution of Earth's geosphere and biosphere (Hazen et al., 2014; Hazen and Morrison, 2020). In mineral ecology (Hazen et al., 2015), researchers can use the package to investigate statistical relationships between mineral species and their geological contexts, contributing to predictive models of mineral formation and distribution. The package also holds promise for cross-disciplinary collaborations. By integrating mineralogical data with environmental, economic, and social datasets, researchers can address pressing global challenges such as critical mineral supply chains and sustainable resource management. According to the discussion on mineral informatics (Prabhu et al., 2023), the work plan of the OpenMindat project (Ma et al., 2024; Que et al., 2024), and the vision of the Deep-time Data-Driven Discovery (4D) Initiative (4D Initiative, 2019), a cyberinfrastructure ecosystem is the foundation to facilitate data-driven discoveries, and the work presented in this paper is a building block for that ecosystem.

Despite its benefits and potential, the OpenMindat R package faces certain limitations and needs further extension. For instance, the current version does not support user-friendly queries involving mineral occurrences due to restrictions in the Mindat API. This limitation constrains studies that require detailed spatial analysis of mineral distributions. For example, retrieving "minerals that contain cobalt but not oxygen that are found in South Africa or Zambia" is almost impossible or may require complex commands for the current version of the package. Additionally, some users may encounter challenges in navigating the package's advanced features, underscoring the need for more detailed tutorials, examples, and user support. To address these issues, the development team is actively working on expanding the package's functionality. Planned updates include incorporating mineral occurrence records as the API evolves, enhancing the package's documentation, and developing interactive tutorials to guide users through complex queries. We are also collecting feedback from the geoscientific community to shape these improvements.

5 Conclusions

This paper introduces the OpenMindat R package, a tool designed to facilitate efficient data retrieval from Mindat, one of the world's largest databases for mineral species and their distributions. By providing a structured interface to the Mindat open data API, the package simplifies the process of accessing and utilizing mineralogical data, making these data more accessible to geoscientists who rely on the R programming environment for data analysis and visualization. The OpenMindat R package addresses a gap by enabling streamlined data retrieval for a variety of use cases. Its functionality includes querying geomaterials based on chemical and physical properties, crystal structures, and other attributes as well as accessing locality and IMA-approved mineral data. The package's support for multiple output formats ensures compatibility with a wide range of analytical workflows commonly used in geoscience research. Moreover, the availability of open and FAIR mineralogy data through this package aligns with broader efforts to enhance data-driven discoveries in the geosciences. By enabling researchers to integrate Mindat data into their workflows with greater efficiency, we hope the OpenMindat R package can provide solid support to data-intensive research and foster innovation in mineral informatics. Continued development of both the Mindat API and the R package will further expand their utility, encouraging new research directions and collaborations in the geoscience community.

Code and data availability. The OpenMindat R package v1.0.0 is free and open source. The web links for its installation guidelines, source code, tutorials, examples, and related documentation are listed in Table 10.

Author contributions. XQ: conceptualization, methodology, software, and writing – original draft, review, and editing; JZ: methodology, validation, and writing – review and editing; WC: validation

Table 10. Online resources for the OpenMindat R package.

Name	URL (last access: 17 July 2025)
CRAN OpenMindat R package v1.0.0 (Que and Ma, 2024a)	https://cran.r-project.org/web/packages/OpenMindat
Source code of the OpenMindat R Package (Que and Ma, 2025a)	https://github.com/quexiang/OpenMindat/
Tutorials of the OpenMindat R Package (Que and Ma, 2025b)	https://quexiang.github.io/OpenMindat/
Examples of the OpenMindat R Package (Que and Ma, 2025c)	https://github.com/quexiang/OpenMindat/tree/main/notebook
Reference manual of the OpenMindat R package v1.0.0 (Que and Ma, 2024b)	https://cran.r-project.org/web/packages/OpenMindat/OpenMindat.pdf
How to setup the Jupyter Notebook for R (Phuriphanvichai, 2019)	https://developers.lseg.com/en/article-catalog/article/setup-jupyter-notebook-r
How to get the Mindat API key (Zhang, 2024)	https://www.mindat.org/a/how_to_get_my_mindat_api_key
Description of geomaterial fields (Richard, 2023)	https://github.com/smrgeoinfo/How-to-Use-Mindat-API/blob/main/geomaterialfields.csv
Mindat API online documentation (Mindat, 2025)	https://api.mindat.org/schema/redoc/

and writing – review and editing; JR: data curation and writing – review and editing; XM: conceptualization, methodology, funding acquisition, validation, and writing – review and editing.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. The authors thank all those who contributed to the many fruitful discussions within the communities of the Deep-time Data-Driven Discovery (4D) Initiative and the Deep-time Digital Earth (DDE) Big Science Program of the International Union of Geological Sciences. We also thank Dominik Hezel and an anonymous reviewer for their detailed comments on an earlier version of this paper, which helped improve the readability and quality of the paper.

Financial support. This research has been supported by the National Science Foundation (grant no. 2126315).

Review statement. This paper was edited by Andy Wickert and reviewed by Dominik Hezel and one anonymous referee.

References

4D Initiative: The 4D Initiative: Deep-time Data Driven Discovery, https://4d.carnegiescience.edu/sites/default/files/4D_materials/4D_WhitePaper.pdf, last access: 15 November 2019.

Bergen, K. J., Johnson, P. A., de Hoop, M. V., and Beroza, G. C.: Machine learning for data-driven discovery in solid Earth geoscience, Science, 363, eaau0323, https://doi.org/10.1126/science.aau0323, 2019.

Broz, M. E., Cook, R. F., and Whitney, D. L.: Microhardness, toughness, and modulus of Mohs scale minerals, Am. Mineral., 91, 135–142, https://doi.org/10.2138/am.2006.1844, 2006.

Chamberlain, K. J., Lehnert, K. A., McIntosh, I. M., Morgan, D. J., and Wörner, G.: Time to change the data culture in geochemistry, Nat. Rev. Earth Environ., 2, 737–739, https://doi.org/10.1038/s43017-021-00237-w, 2021.

Chen, M., Qian, Z., Boers, N., Jakeman, A. J., Kettner, A. J., Brandt, M., Kwan, M. P., Batty, M., Li, W., Zhu, R., and Luo, W.: Iterative integration of deep learning in hybrid Earth surface system modelling, Nat. Rev. Earth Environ., 4, 568–581, https://doi.org/10.1038/s43017-023-00452-7, 2023.

Chiama, K., Gabor, M., Lupini, I., Rutledge, R., Nord, J. A., Zhang, S., Boujibar, A., Bullock, E. S., Walter, M. J., Lehnert, K., Spear, F., Morrison, S. M., and Hazen, R. M.: The secret life of garnets: a comprehensive, standardized dataset of garnet geochemical analyses integrating localities and petrogenesis, Earth Syst. Sci. Data, 15, 4235–4259, https://doi.org/10.5194/essd-15-4235-2023, 2023.

Jeffery, K. G.: FAIR, open, and free does not mean no restrictions, Patterns, 2, 100339, https://doi.org/10.1016/j.patter.2021.100339, 2021.

Hazen, R. M.: Data-driven abductive discovery in mineralogy, Am. Mineral., 99, 2165–2170, https://doi.org/10.2138/am-2014-4895, 2014.

Hazen, R. M. and Ferry, J. M.: Mineral evolution: Mineralogy in the fourth dimension, Elements, 6, 9–12, https://doi.org/10.2113/gselements.6.1.9, 2010.

Hazen, R. M. and Morrison, S. M.: An evolutionary system of mineralogy. Part I: Stellar mineralogy (> 13 to 4.6 Ga), Am. Mineral., 105, 627–651, https://doi.org/10.2138/am-2020-7173, 2020.

Hazen, R. M., Papineau, D., Bleeker, W., Downs, R. T., Ferry, J. M., McCoy, T. J., Sverjensky, D. A., and Yang, H.: Mineral evolution, Am. Mineral., 93, 1693–1720, https://doi.org/10.2138/am.2008.2955, 2008.

Hazen, R. M., Bekker, A., Bish, D. L., Bleeker, W., Downs, R. T., Farquhar, J., Ferry, J. M., Grew, E. S., Knoll, A. H., Papineau, D., and Ralph, J. P.: Needs and opportunities in mineral evolution research, Am. Mineral., 96, 953–963, https://doi.org/10.2138/am.2011.3725, 2011.

Hazen, R. M., Liu, X. M., Downs, R. T., Golden, J., Pires, A. J., Grew, E. S., Hystad, G., Estrada, C., and Sverjensky, D. A.: Mineral evolution: Episodic metallogenesis, the supercontinent cycle, and the co-evolving geosphere and biosphere, Econ. Geol. Spec. Publ., 18, 1–15, https://doi.org/10.5382/SP.18.01, 2014.

Hazen, R. M., Grew, E. S., Downs, R. T., Golden, J., and Hystad, G.: Mineral ecology: chance and necessity in the mineral diversity of terrestrial planets, Can. Mineral., 53, 295–324, https://doi.org/10.3749/canmin.1400086, 2015.

Hazen, R. M., Downs, R. T., Elesish, A., Fox, P., Gagné, O., Golden, J. J., Grew, E. S., Hummer, D. R., Hystad, G., Krivovichev, S. V., Li, C., Liu, C., Ma, X., Morrison, S. M., Pan, F., Pires, A. J., Prab-hu, A., Ralph, J., Runyon, S. E., and Zhong, H.: Data-driven discovery in mineralogy: Recent advances in data resources, analysis, and visualization, Engineering, 5, 397–405, https://doi.org/10.1016/j.eng.2019.03.006, 2019.

Hornik, K.: The comprehensive R archive network, WIREs Computation. Stat., 4, 394–398, https://doi.org/10.1002/wics.1212, 2012.

Hossain, M. A., Dwivedi, Y. K., and Rana, N. P.: State-of-the-art in open data research: Insights from existing literature and a research agenda, J. Org. Com. Elect. Com., 26, 14–40, https://doi.org/10.1080/10919392.2015.1124007, 2016.

Hystad, G., Morrison, S. M., and Hazen, R. M.: Statistical analysis of mineral evolution and mineral ecology: The current state and a vision for the future, Applied Computing and Geosciences, 1, 100005, https://doi.org/10.1016/j.acags.2019.100005, 2019.

Lehnert, K. A., Walker, D., and Sarbas, B.: EarthChem: A geochemistry data network, Geochim. Cosmochim. Ac., 71, A559, https://doi.org/10.1016/j.gca.2007.06.020, 2007.

Liu, C., Eleish, A., Hystad, G., Golden, J. J., Downs, R. T., Morrison, S. M., Hummer, D. R., Ralph, J. P., Fox, P., and Hazen, R. M.: Analysis and visualization of vanadium mineral diversity and distribution, Am. Mineral., 103, 1080–1086, https://doi.org/10.2138/am-2018-6274, 2018.

Ma, X., Ralph, J., Zhang, J., Que, X., Prabhu, A., Morrison, S. M., Hazen, R. M., Wyborn, L., and Lehnert, K.: OpenMindat: Open and FAIR mineralogy data from the Mindat database, Geosci. Data J., 11, 94–104, https://doi.org/10.1002/gdj3.204, 2024.

Mindat: Mindat API online documentation, https://api.mindat.org/schema/redoc/ (last access: 16 May 2025), 2025.

Morrison, S. M., Buongiorno, J., Downs, R. T., Eleish, A., Fox, P., Giovannelli, D., Golden, J. J., Hummer, D. R., Hystad, G., Kellogg, L. H., and Kreylos, O.: Exploring carbon mineral systems: recent advances in C mineral evolution, mineral ecology, and network analysis, Front. Earth Sci., 8, 208, https://doi.org/10.3389/feart.2020.00208, 2020.

Morrison, S. M., Prabhu, A., Eleish, A., Hazen, R. M., Golden, J. J., Downs, R. T., Perry, S., Burns, P. C., Ralph, J., and Fox, P.: Predicting new mineral occurrences and planetary analog environments via mineral association analysis, PNAS Nexus, 2, pgad110, https://doi.org/10.1093/pnasnexus/pgad110, 2023.

Phuriphanvichai, J.: How to setup Jupyter Notebook for R?, https://developers.lseg.com/en/article-catalog/article/setup-jupyter-notebook-r, last access: 8 December 2019.

Prabhu, A., Morrison, S. M., Eleish, A., Zhong, H., Huang, F., Golden, J. J., Perry, S. N., Hummer, D. R., Ralph, J., Runyon, S. E., and Fontaine, K.: Global earth mineral inventory: A data legacy, Geosci. Data J., 8, 74–89, https://doi.org/10.1002/gdj3.106, 2021.

Prabhu, A., Morrison, S. M., Fox, P., Ma, X., Wong, M. L., Williams, J. R., McGuinness, K. N., Krivovichev, S. V., Lehnert, K., Ralph, J., and Lafuente, B.: What is mineral informatics?, Am. Mineral., 108, 1242–1257, https://doi.org/10.2138/am-2022-8613, 2023.

Que, X. and Ma, X.: OpenMindat: An R package for querying and accessing open data from the Mindat API, https://cran.r-project.org/web/packages/OpenMindat (last access: 8 February 2024), 2024a.

Que, X. and Ma, X.: Reference manual of OpenMindat R package v1.0.0, https://cran.r-project.org/web/packages/OpenMindat/OpenMindat.pdf (last access: 8 February 2024), 2024b.

Que, X. and Ma, X.: Source code of the OpenMindat R Package, Github [code], https://github.com/quexiang/OpenMindat/ (last access: 16 May 2025), 2025a.

Que, X. and Ma, X.: Tutorials of the OpenMindat R Package, Github [code], https://quexiang.github.io/OpenMindat/ (last access: 16 May 2025), 2025b.

Que, X. and Ma, X.: Examples of the OpenMindat R Package, Github [code], https://github.com/quexiang/OpenMindat/tree/main/notebook (last access: 16 May 2025), 2025c.

Que, X., Huang, J., Ralph, J., Zhang, J., Prabhu, A., Morrison, S., Hazen, R., and Ma, X.: Using adjacency matrix to explore remarkable associations in big and small mineral data, Geosci. Front., 15, 101823, https://doi.org/10.1016/j.gsf.2024.101823, 2024.

Ralph, J., Martynov, P., Ma, X., Prabhu, A.: Opening mindat.org for data researchers: Announcing OpenMindat, in: The 23rd General Meeting of the International Mineralogical Association, Lyon, France, 2022.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., and Carvalhais, N.: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Richard, S.: Description of Geomaterial fields, Github [code], https://github.com/smrgeoinfo/How-to-Use-Mindat-API/blob/main/geomaterialfields.csv (last access: 29 December 2023).

Richardson, L. and Ruby, S.: RESTful Web Services, O'Reilly Media, Inc., Sebastopol, CA, 454 pp., ISBN 978-0596529260, 2008.

Walker, J. D., Lehnert, K. A., Hofmann, A. W., Sarbas, B., and Carlson, R. W.: EarthChem: international collaboration for solid earth geochemistry in geoinformatics, in: AGU Fall Meeting Ab-

stracts, Vol. 2005, IN44A-03, https://ui.adsabs.harvard.edu/abs/2005AGUFMIN44A..03W/abstract (last access: 17 July 2025), 2005.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. 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., and Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data, 3, 160018, https://doi.org/10.1038/sdata.2016.18, 2016.

Yang, H., Jenkins, R. A., Downs, R. T., Evans, S. H., and Tait, K. T.: Rruffite, $Ca_2Cu(AsO_4)_2 \cdot 2H_2O$, a new member of the roselite group, from Tierra Amarilla, Chile, Can. Mineral., 49, 877–884, https://doi.org/10.3749/canmin.49.3.877, 2011.

Zhang, J.: How to Get My Mindat API Key or Token?, https://www.mindat.org/a/how_to_get_my_mindat_api_key (last access: 12 November 2024).

Zhang, J., Que, X., Madhikarmi, B., Hazen, R. M., Ralph, J., Prabhu, A., Morrison, S. M., and Ma, X.: Using a 3D heat map to explore the diverse correlations among elements and mineral species, Appl. Comput. Geosci., 21, 100154, https://doi.org/10.1016/j.acags.2024.100154, 2024.