Identifier Service in the Mindat Database: Persistent and Structured Access to Massive Records of Minerals and Other Natural Materials

Jolyon Ralph[1], Pavel Martynov[1], Xiaogang Ma[2,3†], David Von Bargen[1], Wenjia Li[2], Jingyi Huang[2,3], Joshua Golden[3], Lucia Profeta[4], Anirudh Prabhu[3], Shaunna Morrison[3], Xiang Que[2], Jiyin Zhang[2]

[1]Hudson Institute of Mineralogy, Keswick, VA 22947, USA

[2]Department of Computer Science, University of Idaho, Moscow, ID 83844, USA

[3]Earth and Planets Laboratory, Carnegie Institution for Science, Washington, DC 20015, USA

[4]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA

Keywords: Mineralogy; Open data; Persistent identifier; FAIR principles; Semantics

Citation: Ralph J., Martynov P., Ma X.G., et al.: Identifier Service in the Mindat Database: Persistent and Structured Access to Massive Records of Minerals and Other Natural Materials. Data Intelligence, Vol. 7, Art. No.: 2025r25, pp. 692-711, 2025. DOI: https://doi.org/10.3724/2096-7004.di.2024.0021

ABSTRACT

Minerals, like many other natural materials of geological origin (i.e., geomaterials), face the challenge of name variations. This in turn hinders the data-intensive geoscience research, which often needs to integrate data from multiple sources. It is clear that mineral name is not an appropriate identifier to connect records within and amongst data sources. The Mindat database, as one of the biggest resources for open data in mineralogy, has received significant volume of feedback on the heterogeneity of mineral and rock names. To address that issue, we established a persistent identifier service on Mindat to provide persistent and meaningful access to the records of geomaterials (mineral/rock/variety), localities, mineral occurrences, references, photos, and specimens. A key development was the long-form identifier, which adds contextual information such as identifier authorities and data types into the identifier structure. Moreover, a UUID service was built along with the long-form identifier to further increase the interoperability. The identifier service has been successfully implemented to mint millions of identifiers to different types of data objects on Mindat. Several use case scenarios were developed to illustrate the utility of the identifiers in the real world. We believe the persistent identifier will help address the challenges caused by name variations, and we welcome Mindat users to test the identifiers and send feedback to us for future extensions.

[†] Correspondence author: Xiaogang Ma (E-mail: max@uidaho.edu).

1. INTRODUCTION

When working with data from different sources it is essential to have a unique key that allows these data sources to be connected properly. In research fields that need mineral data, such as mineralogy and mineral exploration, a common challenge is the cleaning, harmonization, and integration of data from multiple sources. The variations in terminology, classification structure, and data format lead to the heterogeneity in data. To address the challenge, one proven way is to apply persistent identifiers (PIDs) for data objects and use the PIDs as keys to coordinate and interconnect different data resources. The widely accepted FAIR (findable, accessible, interoperable, and reusable) data principles [1] propose PIDs as the top priority for findability of open data, especially for those shared on the Internet. Indeed, PIDs are found throughout the modern scientific world. Each PID has a resolvable landing webpage with standard metadata, which is an efficient way to maintain the findability and accessibility of the corresponding digital object and reduce ambiguity [2]. Scientific articles invariably will have a Digital Object Identifier (DOI) which uniquely represents the article being referenced. Each book has an International Standard Book Number (ISBN) to identify the specific edition, and each author can create an Open Researcher and Contributor IDentifier (ORCID) [3]. There are similar efforts on PIDs of other objects, such as the Research Organization Registry (ROR) identifier for organizations [4] and the International Generic Sample Number (IGSN) for specimens [5], just to mention a few. Nevertheless, the PIDs for mineral data objects, such as mineral species, petrological names, and other natural geomaterials (i.e., materials of geological origin), have not yet been universally adopted.

The Mindat database (mindat.org) provides an ideal platform to establish PIDs for mineral data. Mindat has been running for more than 25 years since its inception by Jolyon Ralph in the late 1990s, and it is now a leading online database for mineral species and their distributions. The crowd-sourcing style of data collection and the large number of committed volunteers on data quality checking and correction make Mindat a very popular database not only for the public but also for researchers, especially those in mineralogy and petrology [6-7]. Many researchers, groups, and other data repositories use Mindat as a reference for *de facto* standard information, such as mineral species name list and attributes. As the database usage has been significantly increasing in recent years (i.e., about 65 million page views from more than 10 million unique users in 2022), users have provided feedback [8-9] that the names of mineral species and other natural geomaterials are not appropriate identifiers for data search and retrieval. This is due to many reasons, including name change (e.g., andorite IV to quatrandorite), format variation (e.g., byströmite and bystromite), false and partial matches (e.g., the reuse of cyprine for a different meaning), language differences (e.g., quartz, cuarzo (Spanish) and 石英 (Chinese)), and more. It becomes essential that there is a consistent and reliable identifier service for representing and linking data objects of minerals and other geomaterials across Mindat and other data repositories.

The purpose of this paper is to present the recently established identifier service in the Mindat database. During this study we have compared mineral names from the Mindat database, the RRUFF Project database (rruff.info), the mineral species list (a PDF document) curated by the International Mineralogical Association (IMA) Commission on New Minerals, Nomenclature and Classification (CNMNC) (cnmnc.units.it), and the IMA mineral species database hosted by RRUFF (rruff.info/ima, which has more detailed information for each species). We have also considered the potential problems in matching datasets from beyond these sources. In the following sections, we will provide a detailed reflection of the challenges, present the developed identifier service for minerals and other data objects in Mindat, and illustrate the advantages of the PIDs with a number of real-world use cases. At the end of the paper, we will discuss the value and limitation of the developed identifier service and propose directions for future work to promote an open mineral data ecosystem.

2. CHALLENGES OF DATA HETEROGENEITY IN MINDAT

While the success of PIDs in many other areas have demonstrated the advantages, a driving force in building the Mindat identifier service was to resolve challenges of data heterogeneity within the Mindat database itself. Moreover, as Mindat is widely used as a reference for mineral species information in many other databases, the established Mindat identifier service will also help facilitate data harmonization across those databases. In this section we will list several representative issues of data heterogeneity in Mindat.

2.1 Name variation

A common assumption has been that the name of an item is its PID, indeed at the moment this is the usual way to link items together in mineral data, however names are not static and various differences prevent the use of the name as a PID (Table 1).

Table 1. Reasons for differences between names in mineral data sources.

Name Difference	Detailed Reasons
Omission	An entry in source A does not appear in source B because source B is not up-to-date.
Mistaken entry	An entry in source A does not appear in source B because it has been mistakenly added to source A or entered incorrectly.
Difference in scope	An entry in source A does not appear in source B because the two data sources have different scopes (for example, comparing petrologic and mineralogical databases)
Name changes	An entry in source A does not have the same name as the corresponding entry in source B because a name update has not been applied consistently to both data sources.
Formatting differences	An entry in source A does not match the corresponding entry in source B because there are formatting differences in the name.
International differences	An entry in source A does not match the corresponding entry in source B because the name is recorded in a different language.
False match	An entry in source A incorrectly matches an entry in source B because the same name has been used historically for two different things.
Partial match by redefinition	An entry in source A matches an entry in source B, but the name has been redefined to a subset (or superset) of the original definition.

Omission of entries may simply be due to human error in data entry or due to a delay in updating new entries as they are published. Within IMA, new mineral names are voted on once a month. Those that are approved are then for a short period of time confidential until the name is released by the IMA-CNMNC, the mechanism for publication is then through a newsletter published within the academic journals or through the IMA-CNMNC's master list (a PDF) [13]. Due to the foibles of journal publication timelines and/or the voluntary nature of the IMA-CNMNC individuals responsible for modifying the master PDF list, or even a decision by an author to release the mineral name to any of these channels, the release of a name across digital and/or physical platforms is inevitably not synchronized. Mistaken entries have been more prevalent on Mindat than in other sources due to the number of entries (over 54,000 mineral and rock names) and the crowd-sourced and community-driven nature of the content, although additions and edits of these names on Mindat is limited to a small, trusted subset of the community. Mistaken entries are usually due to unintended duplication with an incorrect spelling entry. In many cases the mistakes were due to the misspellings, alternative names, or typographical errors in the source literature of the data entries, such as such as barite for baryte, sulfur for sulphur, and fluorspar for fluorite. Matching between other sources relying on IMA-approved names only might fail if an entry has been mistakenly kept on one site but correctly removed as "discredited" on another. This does not apply to Mindat where discredited names are kept for their historical importance, such as the mineral species "mohawkite".

Difference in scope is important when comparing the Mindat list of mineral names with that of other sites which may only include IMA-approved names. An example might be trying to compare "amethyst" between Mindat and RRUFF, as amethyst is not an IMA-approved species this match would not work directly unless the "amethyst" entry is remapped to quartz. Name changes within the IMA nomenclature are frequent, despite the potential for confusion this may cause. Indeed, such changes are also one of the factors driving a need for PIDs in Mindat and other databases. In mid-2022 a series of name changes of minerals was announced by the IMA [26], for example, changing andorite IV to quatrandorite. As of mid-December 2022, this change had not yet been reflected in the RRUFF Project IMA List of Approved Mineral Species. The other issues such as formatting differences, international differences, and false and partial matches involve more complicated knowledge and technical backgrounds and will be described in the following subsections.

2.2 Format and international heterogeneity

Formatting differences can cause significant problems when trying to match names even if they appear identical on screen or when printed. For example, the letter "o" with a diaeresis in the name byströmite can be entered in several different ways in Unicode text encoding systems (see Table 2). Most importantly the variant that is used will often depend on the operating system and software used to enter the name and will not always be obvious to the person entering the data. Additionally, systems that use Hypertext Markup Language (HTML) encoded text can use multiple different HTML encoding methods for the same character. In the mineralogy community there are cases where the American Standard Code for Information Interchange (ASCII) had been used in the name (e.g., just a plain "o" in bystroemite or bystromite) rather than Latin extended ASCII or UTF-8. It was 15 years ago when the IMA standardized the use of extended characters [10].

Table 2. Different ways the ö character can be encoded in Unicode and/or HTML text encoding.

Type	Encoding	Notes
Latin extended ASCII	HEX 0xF6	Default when entered by keyboard on system with this key, or using most shortcut options on Mac, Windows and Linux.
Unicode UTF-8	HEX 0xC3B6	The UTF-8 character encoding.
Unicode UTF-16	HEX 0x00F6	The UTF-16 character encoding.
Unicode using Combining characters	HEX 0x6F0304	Unicode combining characters consisting of the standard lower-case 'o' character (0x6F) followed by the unicode combining character for a diaeresis (0x0304)

HTML encoding (decimal, hexadecimal, and HTML entity)	&#246;
	&#xF6;
	&ouml;
		These can often be found in content extracted from a web page.

International differences in names include simple language differences. For example, the English name galena would be galenit in German, 方铅矿 in Chinese, and lyijyhohde in Finnish. Also included are differences between American and British English variants, such as baryte (more commonly used in British English) and barite (more commonly used in American English).

2.3 False and partial matches

False matches are where name collisions exist for different uses of the name, usually incompatible historical and modern usage. Classic examples include siderite (i.e., the iron carbonate mineral but also used historically for an unrelated class of meteorite) and dolomite (i.e., the calcium magnesium carbonate but also widely used as the name of a magnesian limestone rock, a.k.a. dolostone). A more recent example is the reuse of the name cyprine for a copper-dominant member of the vesuvianite group compared to historic use of the name as a varietal name for a copper-bearing (but not necessarily copper dominant) vesuvianite.

Mindat handles name collisions by following the standard popularized by Hey in his book "An Index of Mineral Species and Varieties arranged Chemically" [11] by distinguishing alternative uses of names with a suffix denoting the author who first published using the name in a particular context. For example, the name apachite is used for a copper silicate mineral and is included in the IMA mineral name list, but the name was also applied independently by Osann in 1896 to represent a rock type – "a variety of peralkaline phonolite rich in sodic amphiboles and containing sodic pyroxene and aenigmatite". This variant of the name is listed as "apachite (of Osann)" in Mindat.

Partial match by redefinition is frequently an issue where something originally described as a single mineral species is redefined subsequently to be two or more slightly different phases, or distinctly different species. Ardennite, an arsenate-silicate, was first described in 1872 from Belgium. In 2007, the vanadium analogue at the T4 site of ardennite was described as a new mineral and named ardennite-(V) [12]. Ardennite and ardennite-(V) are considered two different minerals and so necessitated renaming ardennite to ardennite-(As) and making ardennite into a series name that encompasses the two end members. It also happens when complex mineral groups such as the pyroxene or amphibole groups are redefined and the name that was previously used to represent a specific composition range, now maps to a different range.

It can be seen from the above-mentioned challenges that it is not reliable to use the name as a PID for mineral species in Mindat. Alternative solutions need to be explored. Besides the purpose of identification, using a PID as the primary key for mineral related information in a database also makes the dataset much easier to update in the future as the mappings from the PID to the current name can be made automatically.

3. DESIGN AND DEVELOPMENT OF THE MINDAT IDENTIFIER SERVICE

So far there is no universally adopted unique identifier for the mineralogical world, however there are already options for moving towards one. The design and development of the identifier service on Mindat aims to address the challenges mentioned above while also being compatible with the existing database structure.

3.1 Existing mineral identifiers

For PIDs to work there needs to be either one central authority that issues them, or multiple authorities cooperating to issue compatible identifiers. In addition, these need to be resolvable, i.e., there needs to be a standard gateway where the PID can be resolved to provide information about the record and what it represents. For example, DOI records are resolved through the https://doi.org site operated by the International DOI Foundation, although individual DOI records are issued by affiliated organizations, such as Crossref (https://www.crossref.org/). Looking at what is possible for mineralogy we should first examine the existing possible codifications and identifiers that could potentially be adopted for wider use (Table 3). We also include the mineral name in the table for comparison.

The IMA, neither in the frequently-updated PDF list of mineral species (e.g., [13]) nor the RRUFF IMA List of Approved Mineral Species (rruff.info/ima), provides specific PIDs, but two items that have been suggested in the mineral species documentation are the IMA mineral symbol and the IMA proposal number (i.e., year and mineral number combination). Neither of these are appropriate for use as a PID. The short code is related to the name and would change if the name changes. The IMA proposal number is not available for all mineral species, and, in some cases, the proposal number covers multiple species (e.g., in the group/nomenclature proposals). Moreover, although the IMA's main concern is with approved mineral species, they also care about mineral groups and unnamed geomaterials which are incompletely described minerals (e.g., the chemistry is known, but crystallography may not be completely characterized). Those records (see example in Table 3) further increase the complexity.

Table 3. Characteristics of different efforts on nomenclature, codification, and identifier of mineral species.

Possible identifier	Issuing Authority	Example	Unique	Persistent	Comprehensive	Resolvable
Name	various	molybdenite	No	No (names can be changed)	Yes	No*
IMA Mineral Symbol	IMA	Sbn	Yes	No (will change when name changes)	No (only approved species)	No*
IMA Proposal No.	IMA	IMA2006-043	Yes (but note some IMA proposals are related to groups or nomenclature changes)	Yes	No (only minerals approved/ redefined since 1959)	No*
IMA Unnamed Mineral	IMA	UM1989-08-E: CaSi	Yes	Yes	No	No*
Dana Classification	Not updated since 1997	75.1.3.1	No	No	No	No*
Strunz Classification	Not updated since 2001 (unofficial updates since then)	2.EB.05a	No	No	No	No*

Table 3. *Continued.*

Possible identifier	Issuing Authority	Example	Unique	Persistent	Comprehensive	Resolvable
Hey's Chemical Index	Not updated since 1993	11.2.2	No	No	No	No*
Mindat ID	Mindat	3337	Yes	Yes	Yes	Yes, example:https://www.mindat.org/min-3337.html
Mindat Long-form Identifier (newly developed in this work)	Mindat	mindat:1:1:3337:0	Yes	Yes	Yes	Yes, example: https://www.mindat.org/1:1:3337:0
Mindat RFC4122 GUID (newly developed in this work)	Mindat	4ca61d6f-75f8-4208-8fb2-3b0eecbcd8f0	Yes	Yes	Yes	Yes, example: https://www.mindat.org/4ca61d6f-75f8-4208-8fb2-3b0eecbcd8f0

* Items not resolvable directly online but can be resolved by searching Mindat.

Other identifier systems well-known to mineralogists include the classic Dana Classifications system [14], the Strunz Classification system [15], and the Hey's Mineral Index [11, 16]. All three cover mineral species and a limited number of other historical names only.

Before the start of this work, Mindat already had an identifier (called Mindat ID) for every mineral species, natural geomaterial and name variant recorded in its database. To resolve a Mindat ID to the web page containing the information for the name usage, a Uniform Resource Locator (URL) of the format https://www.mindat.org/min-{id}.html is used, where {id} is replaced with the Mindat ID in question, so for example the page for quartz (Mindat ID 3337) would be https://www.mindat.org/min-3337.html. An issue with using the existing Mindat ID as PID is that, the number alone lacks context and resolving the ID of quartz (3337) for example requires knowledge that the specific format for the resolving URL is https://www.mindat.org/min-3337.html. In Mindat, 3337 is also a valid ID for a locality record (i.e., Mt Graham, Aravaipa Mining District, Graham County, Arizona, USA) https://www.mindat.org/loc-3337.html. This may cause confusion for end users when they are querying and integrating different types of records.

3.2 Long-form identifiers with more semantics embedded

To address the limitations of the existing identifiers, we extended a "long-form" version of the Mindat IDs (see second last row in Table 3), which consists of a namespace prefix and four other numeric components separated by the colon character to give more contextual and semantic information of the identifier. Table 4 gives more details of each component. The first component is the namespace prefix "mindat" which represents the URL https://www.mindat.org/. The second component is the authority identifier, the organization responsible for issuing new identifiers in this sequence. In most cases this will be Mindat, however where we are using and sharing external datasets such as those from RRUFF [17] and the PaleoBioDB [18] those identifiers are issued by those organizations. The third component corresponds with the data type, 1 being a mineral/rock/variety name, and 2 being a locality for example. The fourth is the identifier for the item as described previously. The fifth component is a checksum digit calculated from the previous three items; this works in the same way as the final checksum in an ISBN number.

Following the styles specified in Table 4, the long-form identifier for quartz is "mindat:1:1:3337:0". It can be resolved by mindat.org simply by prepending "https://www.mindat.org/" to replace "mindat:" and crafting a URL such as https://www.mindat.org/1:1:3337:0, which resolves to https://www.mindat.org/min-3337.html. In this way we keep the system of the original Mindat IDs and added semantics and contextual information in the long-form identifiers. Mindat also provides relationships between the PIDs. The relationships are often hierarchical, so that for example augenachat (mindat:1:1:30437:3) is mapped as a (German language) synonym of eye agate (mindat:1:1:7598:1), which is a variety of agate (mindat:1:1:51:8), which is itself a variety of chalcedony (+/- moganite) (mindat:1:1:960:9), which is in turn a variety of quartz (mindat:1:1:3337:0).

It is noteworthy that the long-form identifier is designed to cover all data objects on Mindat, not just the mineral/rock/variety name records. The third component (data type) in Table 4 gives a short list of the object types. In section 4 we will have an overview about the total numbers of long-form identifiers minted for each data type so far.

Table 4. Components of a Mindat long-form identifier.

Element	1	2	3	4	5
Description	The namespace prefix	The authority of identifier*	The data type	Unique identifier (32bit unsigned integer)	Checksum digit
Example Values	mindat	1 - Mindat 2 - GeoNames 3 - PBDB 4 - GBIF 5 - IMA CNMNC 6 - RRUFF	1 - Geomaterial (mineral/rock/variety name) 2 - Locality 3 - Locentry (i.e., geomaterial occurrence) 4 - Photo 5 - Reference 6 - Feature (i.e., non Mindat locality) 7 - PBDB record 8 - GBIF record 9 - Mindat paleo record 10 Specimen	All items have a unique identifier, e.g., 3337 for quartz.	0-9

* GeoNames: Database of geographical names [29]; PBDB: PaleoBiology Database [18]; GBIF: Global Biodiversity Information Facility [28]; IMA CNMNC: Commission on New Minerals, Nomenclature and Classification of the International Mineralogical Association [13]; RRUFF: Integrated database of Raman spectra, X-ray diffraction and chemistry data for minerals [17].

3.3 RFC 4122 UUIDs

Although the first two types of identifiers mentioned above (i.e., the Mindat ID and the long-form identifier) are more than adequate for most data science purposes, there is a desire to integrate Mindat data with other established data projects and portals. Many data repositories use Internet Engineering Task Force (IETF)'s standard RFC 4122 [27] to create Universal Unique IDentifier (UUID) or Globally Unique IDentifier (GUID) for their records. These are 36-character (hyphens included) identifiers generated randomly, there are 5.3 x 10^{36} possible combinations meaning the chance of the same identifier being issued for two items is so low that it is safe to assume the identifiers are unique.

Mindat is issuing these UUIDs for each record in addition to the long-form identifiers and these are also resolvable on Mindat simply by prepending https://www.mindat.org/ to craft the appropriate URL. The RFC 4122 UUID for quartz as issued by Mindat is 4ca61d6f-75f8-4208-8fb2-3b0eecbcd8f0. The corresponding URL https://www.mindat.org/4ca61d6f-75f8-4208-8fb2-3b0eecbcd8f0 also resolves to https://www.mindat.org/min-3337.html.

By providing three different styles of PIDs for records it should be possible to ensure easy integration of Mindat data with any other data projects. The recently established Mindat Open Data Application Programming Interface (API) [7] also allows easy and direct access to all these identifiers in data queries and downloads. It is important to note that, although the RFC 4122 style UUIDs can provide PIDs for every type of object, there is no central registry of these UUIDs, so it is important to know which authority (in our case Mindat) is responsible for issuing the identifier, otherwise it is impossible to resolve.

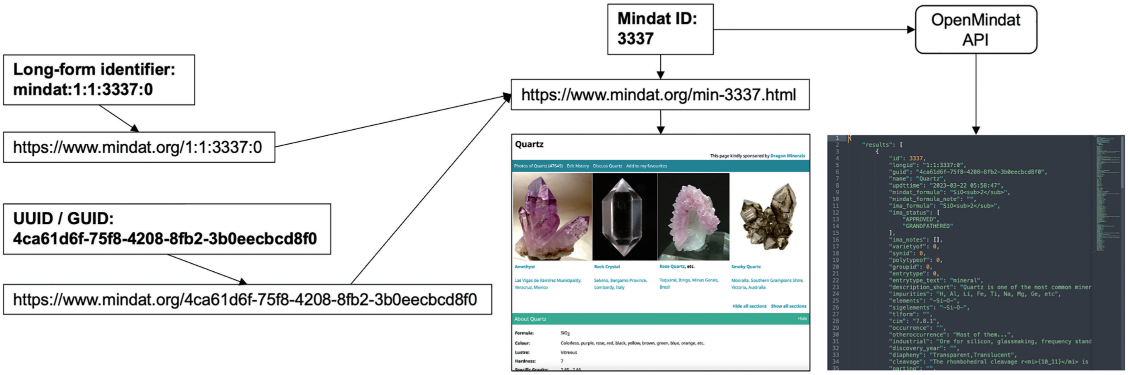
4. IMPLEMENTATION AND RESULTS

The service of long-form identifiers and UUIDs was fully implemented on Mindat, and by July 2023 a large number of data objects have been minted with those new identifiers. Arranged by some data types listed in Table 4, the results include:

1 - Geomaterial (mineral/rock/variety name): 54,640;
2 - Locality: 394,968;
3 - Mineral occurrence record: 1,499,309;
4 - Photo: 1,282,758;
5 - Reference: 15,968,107; and
10 - Specimen: 525.

The number of specimen identifiers is relatively low because each specimen is the chemical analysis of a geomaterial (mostly minerals) and right now the detailed chemical analyses are hard to add into Mindat. It requires some reasonable work on both data entry methods and data display. A few years ago, the Mindat technical team had developed the minID as a PID for specimens [19]. So far there are 1,958,647 specimens indexed with minID, but the data are just, at best, the geomaterials in the specimen and where it came from. While the service of long-form identifiers and UUIDs is already applicable to specimens, we still need more work to refine the data structure for specimens and enrich the records on Mindat, including, for example, the adaptation of the Darwin Core standard [20] and the IGSN metadata schema [21].

The implementation of the semantically-enriched identifier service to geomaterials (mineral/rock/variety name) has been impactful. Mindat is able to provide PIDs for instances of every type of natural geomaterial included in the database (Figure 1). These are divided into categories and some representative examples are given in Table 5. Beyond the IMA mineral name list, there are many other complicated situations for mineral names. For instance, many of the unnamed minerals were interpretations taken from publications in scientific journals. In the category "IMA Unnamed Mineral", the codification of minerals was not taken from the original publications but was created by the co-authors on the IMA CNMNC Sub-committee on Unnamed Minerals. The "Other" category includes a large number of geomaterials. It can include non-crystalline materials, materials that are not (yet) approved as minerals, anthropogenic materials found within the geological environment and discredited materials.


Mindat ID:
3337
OpenMindat
API
Long-form identifier:
mindat:1:1:3337:0
https://www.mindat.org/min-3337.html
https://www.mindat.org/1:1:3337:0
UUID / GUID:
4ca61d6f-75f8-4208-8fb2-3b0eecbcd8f0
https://www.mindat.org/4ca61d6f-75f8-4208-8fb2-3b0eecbcd8f0
Quartz
This page kindly sponsored by Dragon Minerals
Photos of Quartz (47649)
Edit history
Discuss Quartz
Add to my favourites
Amethyst
Las Vigas de Ramírez Municipality, Veracruz, Mexico
Rock Crystal
Selvino, Bergamo Province, Lombardy, Italy
Rose Quartz, etc.
Taquaral, Itinga, Minas Gerais, Brazil
Smoky Quartz
Mooralla, Southern Grampians Shire, Victoria, Australia
Hide all sections
Show all sections
About Quartz
Hide

Formula: SiO2
Colour: Colorless, purple, rose, red, black, yellow, brown, green, blue, orange, etc.
Lustre: Vitreous
Hardness: 7
Specific Gravity: 2.65 - 2.66

```
{
    "results": [
        {
            "id": 3337,
            "longid": "1:1:3337:0",
            "guid": "4ca61d6f-75f8-4208-8fb2-3b0eecbcd8f0",
            "name": "Quartz",
            "updttime": "2023-03-22 05:58:47",
            "mindat_formula": "SiO<sub>2</sub>",
            "mindat_formula_note": "",
            "ima_formula": "SiO<sub>2</sub>",
            "ima_status": [
                "APPROVED",
                "GRANDFATHERED"
            ],
            "ima_notes": [],
            "varietyof": 0,
            "synid": 0,
            "polytypeof": 0,
            "groupid": 0,
            "entrytype": 0,
            "entrytype_text": "mineral",
            "description_short": "Quartz is one of the most common miner
            "impurities": "H, Al, Li, Fe, Ti, Na, Mg, Ge, etc",
            "elements": "-Si-O-",
            "sigelements": "-Si-O-",
            "tlform": "",
            "cim": "7.8.1",
            "occurrence": "",
            "otheroccurrence": "Most of them...",
            "industrial": "Ore for silicon, glassmaking, frequency stand
            "discovery_year": "",
            "diapheny": "Transparent,Translucent",
            "cleavage": "The rhombohedral cleavage r<mi>{10_11}</mi> is
            "parting": "",
```

Figure 1. A simplified illustration of the established Mindat identifier service with quartz as an example. Besides the resolution of identifier URLs to web pages on the front end, the identifier can also be used to query and retrieve dataset through the Mindat Open Data API (i.e., OpenMindat).

Although the designed identifier service has been fully established in the Mindat database, the application of it by external users, such as in database curation, data integration, and inter-connection of records, will still need more time. Based on the feedback from Mindat users in the past years, we have created a list of use case scenarios to help demonstrate how the Mindat identifier service will be able to address the needs of consistent mineral and geomaterial records amongst databases and other sources. Although the actual application of those use cases depends on the end users in their databases or research activities, the description of the scenarios and the proposed solutions are straightforward, and we hope they can explain the potential usage and value of the established Mindat identifier service for the broad user community.

Consistent petrological names in geochemical data integration: Developers of a geochemical database wishing to store details of petrological information would need to ensure that the petrological names used are consistent and easy to enter. Storing such names as a free-text field can lead to inconsistencies and undetected errors. Using the Mindat Open Data API to resolve names entered into PIDs for petrological terms (e.g., granite maps to mindat:1:1:48141:4, syenite maps to mindat:1:1:48213:0, porphyritic anorthoclase syenite maps to mindat:1:1:50526:4, and larvikite maps to mindat:1:1:39944:1) will help ensure internal consistency and ease of mapping with external resources. Users of the database would still see the mapped names (e.g., granite or syenite), but they would be stored internally connected to the Mindat PIDs.

An integrated specimen catalog across museums in the world: Museums each use their own different catalog systems of their mineral specimens. Although IMA has taken initiative to build a type specimen catalog service, there are many hurdles to directly link specimen records from museums and other organizations across the world. For a global database of specimens to be able to be integrated reliably with all the different museum catalog systems, there needs to be a way to ensure the names are used consistently between the museums and the type specimen catalog service. The Mindat identifier service will ensure these links are valid and permanent. For example, in the Natural History Museum, London, the catalog entry BM.1937,51 is listed as "Andorite var. fizelyite" [22], which is corresponding to the species Fizélyite on Mindat. The long-form identifier (mindat: 1:1:1554:3) of Fizélyite can be used to establish a persistent mapping between the two. Facing the massive amounts of specimen collections, another topic of interest here is the automation of the mapping process to quickly scale up the usage of the Mindat PID.

High quality data for mineral exploration purposes: A mining company wishes to obtain mineralogical data from different agencies to combine with their own proprietary data for building a deep-learning model. Scientists in the company would need to ensure the mineral and geomaterial names used in all the different datasets are consistent. Using the Mindat PIDs, such as those in Table 5, they can easily get the formal and informal mineral name lists as well as the mapping between them. The result can be used to cleanse massive records retrieved from different sources. This will speed up the work of data integration and reduce the chance of error.

Table 5. Different types of geomaterial and exemplar identifiers on Mindat.

Type	Definition	Example(s)	Naming Authority	Number of Mindat Entries (by July 2023)
IMA Mineral	A mineral species approved by or accepted by the International Mineralogical Association	Quartz (mindat: 1:1:3337:0), Calcite (mindat: 1:1:859:4), Pharmacosiderite (mindat:1:1:3185:7)	IMA	5,943
Group	A group of two or more mineral species related by similar structure and/or chemistry	Clinopyroxene subgroup (mindat: 1:1:7630:8), Autunite group (mindat: 1:1:29274:0)	IMA	695
Variety	A named variety of an approved mineral or other variety	Amethyst (mindat: 1:1:198:0), Agate (mindat: 1:1:51:8), Ruby (mindat: 1:1:3473:5)	none	1,885
Mixture	A named mixture of 2 or more minerals	Andrewsite (mindat: 1:1:226:2)	none	193
IMA Unnamed Mineral	There is enough chemical or crystallographic evidence to suggest that there might be a new mineral, but not the evidence available to prove it	UM1989-08-E: CaSi (mindat: 1:1:51627:9)	IMA CNMNC Sub-committee on Unnamed Minerals	1,625 total; 209 in synonyms

Table 5. *Continued.*

Type	Definition	Example(s)	Naming Authority	Number of Mindat Entries (by July 2023)
Unnamed Mineral	A mineral species with known chemistry, but incomplete crystallography	Unnamed (Fe-analogue of Emplecite) (mindat: 1:1:53601:5), Unnamed (NH4-Al Fluoride) (mindat: 1:1:39605:5)	none	878 total; 648 in synonyms
Rock	A name for a type of rock	Syenite (mindat: 1:1:48213:0), Sandstone (mindat: 1:1:49438:4), Komatiite (mindat: 1:1:48568:7)	None (but rules are published by IUGS and some geological surveys)	3,085
Meteorite Classification	A name for a classification of meteorite type based on composition and alteration.	CV2 chondrite meteorite (mindat: 1:1:49578:1), IIIAB Iron meteorite (mindat: 1:1:49930:7)	The Meteoritical Society	449*
Synonym	An alternative name for another entry. This could be a historical name, a name in another language, or IMA number.	Chalybite (mindat: 1:1:6234:9) = siderite (mindat: 1:1:3647:0), Antimonglanz (mindat: 1:1:6337:7) = stibnite (mindat: 1:1:3782:8), IMA2011-044 (mindat: 1:1:41952:3) = Krasheninnikovite (mindat: 1:1:41951:4)	none	38,445
Commodity	The economic materials or elements at a mine locality	Zinc (mindat: 1:1:52522:2), Dimension Stone (mindat: 1:1:52446:3)	None (but list after USGS commodities)	128
Other	Names which do not fall conveniently into any other category.	Carbon Dioxide Ice (mindat: 1:1:25561:5), Horsfordite (mindat: 1:1:1932:9)	None (but some chemical names are covered by IUPAC rules)	2,620

* Included in rock total.

5. DISCUSSION

The identifier service on Mindat makes it possible to more efficiently curate mineral names over time and connect across databases. The functionalities based on and/or relevant to the Mindat PID have established a solid step moving Mindat towards the FAIR principles. Each Mindat PID resolves to a unique data record (i.e., findability), where the valid name, other attributes, as well as the relationships to other data records can be retrieved either from the frontend website or through the open data API (i.e., accessibility). The semantics embedded in the longer-form identifier, the application of community-level standards (e.g., IMA mineral name list and the Dana, Strunz, and Hey classifications) and the inter-connection of name variants improve the data interoperability. The appropriate citations to data sources and the inter-mappings with other databases increase the reusability of the data.

The real-world use of a PID for name variants (synonyms, spelling variants, international variants, and character encoding variations) will depend on the nature of the information being referenced. The Mindat identifier service is able to handle a variety of complicated situations. In most cases it would be more appropriate to store the PID of the valid species for which the name variants relate. This requires a gateway that can resolve such PIDs and return their relationship to any parent species. In some cases, this may be a one-to-many relationship, for example names that refer to mixtures of two or more different mineral species, or the relationship of mineral components of a specific rock type. Mindat PIDs are backed by tables of relationships between identifiers, so that historical synonyms, international name variants, varietal names or trade names can be easily mapped to the current correct species names (e.g., the example about quartz at the end of section 3.2).

Although the identifier service is initially focused on minerals and other natural geomaterials, Mindat also carries PIDs for other data objects, most notably the locality and locentry (i.e., geomaterial occurrence) records. Some of the problems with using locality names as an identifier is that they can be renamed, there can be many localities with the same name, the political hierarchy can change, and they can be merged over time. Of the 201, 000 localities considered as being mineral deposits in Mindat, 84% of them have a longitude/latitude which will be linked uniquely to the Mindat identifier service. For example, the locality identifier for the Tsumeb Mine, Namibia is 2428. This can be resolved by visiting https://www.mindat.org/loc-{ID}.html, for the Tsumeb mine example that would be https://www.mindat.org/loc-2428.html. Using the identifier service built in this work, the corresponding long-form identifier and UUID for this locality are mindat:1:2:2428:3 and 5add80d0-4d74-4781-830a-d8d7680140f5, respectively. A locentry is a locality and geomaterial pair. On Mindat, most locentries are mineral occurrences, and they all have PIDs. For example, the locentry 11255 is for the tsumebite from Tsumeb Mine, Namibia (https://www.mindat.org/locentry-11255.html). The corresponding long-form identifier and UUID are mindat: 1:3:11255:4 and 14b227d9-4110-465a-9363-8f785b23ff01, respectively. As the locentry identifier is interconnected with the mineral species and locality identifiers, from the webpage of the locentry 11255 it is very easy to navigate to the other pages. Mindat locality and locentry records are already in use within other datasets, notably the Mineral Evolution Database [23] uses the Mindat locality identifiers to support a range of studies, such as mineral ecology and mineral evolution [6, 24]. With the new identifier service developed in this study, those data integration efforts will be better supported.

The utility, maintenance, and extension of the Mindat identifier service in the global mineralogical community are still to be widely tested and evaluated by the users. One drawback to the possible adoption of Mindat PIDs for mineral names is a socio-political rather than technical issue [5, 25]. Mindat is an independent organization (part of the Hudson Institute of Mineralogy, a 501(c)3 not-for-profit organization) which has no responsibility for nomenclature in mineralogy, and it would be reasonable to assume that the IMA should have the responsibility to manage such PIDs for mineral species themselves. Other organizations, such as the International Union of Geological Sciences (IUGS), may be better to take responsibility for the PIDs of petrologic names. This might seem problematic in that different incompatible identifier systems that themselves only cover part of the entire nomenclature for minerals, rocks, and related names, could cause data confusion. However, it isn't necessarily a problem as long as there is one central authority capable of resolving all different types of identifiers, and again Mindat is able to take on this role should official identifier systems and the corresponding services be launched by other organizations. There is of course no technical need for separate identifier systems within the mineralogical nomenclature space and, as such, there is a lot of opportunities for Mindat to collaborate with associations such as IUGS and IMA to reduce duplicated work on identifiers and provide consistent information to the users.

With the established identifier service and several other open data efforts on Mindat, we are able to envision an open data ecosystem for mineralogy, petrology, and economic geology. The Mindat Open Data API was established in Spring of 2023 [7] and it now allows full programmatic access to mineral data using the PID, and most importantly offers a resolving service allowing the PIDs for names to be found and, as such, to reduce the challenges caused by formatting differences and other naming differences. These identifiers remain constant. If the name that is used to represent the fundamental properties of a geomaterial changes, the identifier remains the same, the old name is issued a new identifier as a synonym of the original identifier, and they are interconnected to enable traceability. Besides the PIDs and open data services, Mindat has also established inter-connections with other databases, such as the RRUFF Project [17]. The RRUFF Project database has an internal identifier for each mineral species, so for example the RRUFF identifier for quartz is 3763. These identifiers are different to the identifiers issued by Mindat but are interconnected through mapping. Besides the inter-connection with RRUFF, Mindat also offers mappings to a few other databases for most minerals. Overall, a stable, persistent, interconnected, and meaningful open data ecosystem needs fundamental work on the categories, identifiers, annotation, and linking of a variety of data objects [25]. Mindat has made solid progress towards such an open data ecosystem, including (1) crowd-sourced and expert-curated efforts to reduce ambiguity of mineral identifiers, nomenclature, and relationships, (2) sorted and consistent database structure within Mindat, and (3) the interconnection amongst open data resources for minerals, such as RRUFF, PaleoBioDB, and Mineral Evolution Database, as well as many other museum repositories and catalogs of mineral specimens (e.g., to locate the type specimen of mineral species). Yet, there is still a long way to go to achieve the real ecosystem. For instance, the Darwin Core standard [20] and the metadata schema of IGSN [21] can be further adapted on Mindat to improve the data structure of specimens.

6. CONCLUSIONS

The variations in mineral and geomaterial names lead to challenges in data-intensive geoscience research. With the Mindat database as the platform, we designed a semantically-enriched identifier service to address those challenges. Mindat originally already had an identifier service based on sequential numbers. The newly developed long-form identifier embeds the original system and adds more contextual information about identifier authorities and data types. To further increase the interoperability, an RFC 4122 UUID service is also deployed along with the long-form identifier. We have successfully implemented the new identifier service on Mindat to mint a large number of identifiers for different types of data objects, including geomaterial (mineral/rock/variety name), locality, mineral occurrence, reference, photo, and specimen. Several use case scenarios are documented to show the advantage of the identifiers in real-world usage. We understand that the identifier service is a mixture of socio-political and technical systems. Within the Mindat database we are confident about its utility to address the issues of name variations, while the adoption of it in the global mineralogical community may still take some time. We welcome comments and feedback from Mindat users on the identifier service, which will help our future maintenance and extension to it.

AUTHOR CONTRIBUTIONS

Jolyon Ralph: Conceptualization, methodology, software, writing original draft, writing review and editing. Pavel Martynov: methodology, software, writing original draft, writing review and editing.

Xiaogang Ma: methodology, validation, writing original draft, writing review and editing, funding acquisition.

David Von Bargen: methodology, software, writing original draft, writing review and editing.

Wenjia Li: methodology, validation, writing original draft, writing review and editing.

Jingyi Huang: validation, writing review and editing.

Joshua Golden: Conceptualization, writing review and editing.

Lucia Profeta: Conceptualization, writing review and editing.

Anirudh Prabhu: validation, writing review and editing.

Shaunna Morrison: validation, writing review and editing.

Xiang Que: validation, writing review and editing.

Jiyin Zhang: validation, writing review and editing.

FUNDING

This research was funded by the U.S. National Science Foundation (Grant No. 2126315).

DATA AVAILABILITY STATEMENT

This is a software service paper and there is no data used.

ACKNOWLEDGMENTS

We thank many colleagues in the Deep-time Data Driven Discovery (4D) initiative and Deep-time Digital Earth (DDE) program for the discussions on community needs. We also thank two anonymous reviewers for their constructive comments on an earlier version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] 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., Mons, B.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[2] Juty, N., Wimalaratne, S.M., Soiland-Reyes, S., Kunze, J., Goble, C.A., Clark, T.: Unique, persistent, resolvable: Identifiers as the foundation of FAIR. Data Intelligence 2(1-2), 30–39 (2020). https://doi.org/10.1162/dint_a_00025

[3] Haak, L.L., Fenner, M., Paglione, L., Pentz, E., Ratner, H.: ORCID: a system to uniquely identify researchers. Learned publishing 25(4), 259–264 (2012). https://doi.org/10.1087/20120404

[4] Politze, M.: Organization Information gone wild: ROR, entity IDs and the organization ontology. EPiC Series in Computing 78, 108–114 (2021). https://doi.org/10.29007/rz1j

[5] Klump, J., Fils, D., Devaraju, A., Ramdeen, S., Robertson, J., Wyborn, L., Lehnert, K.: Scaling identifiers and their metadata to gigascale: an architecture to tackle the challenges of volume and variety. Data Science Journal 22(5), 1–17 (2023). https://doi.org/10.5334/dsj-2023-005

[6] 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., Zhong, H.: Data-driven discovery in mineralogy: recent advances in data resources, analysis, and visualization. Engineering 5, 397–405 (2019). https://doi.org/10.1016/j.eng.2019.03.006

[7] Ma, X., Ralph, J., Zhang, J., Que, X., Prabhu, A., Morrison, S.M., Hazen, R.M., Wyborn, L., Lehnert, K.: OpenMindat: open and FAIR mineralogy data from the Mindat database. Geoscience Data Journal, 11(1), 94-104 (2024). https://doi.org/10.1002/gdj3.204

[8] Gavryliv, L., Ponomar, V., Putiš, M.: Classifying minerals and their related names in a relational database. Mineralogical Magazine 87(3), 480–493 (2023). https://doi.org/10.1180/mgm.2023.23

[9] Prabhu, A., Morrison, S.M., Hazen, R.M.: Mineral Informatics: origins. In: Celebrating the International Year of Mineralogy: Progress and Landmark Discoveries of the Last Decades, Springer Nature: Cham, Switzerland, pp. 39–68 (2023)

[10] Burke, E.A.: Tidying up mineral names: an IMA-CNMNC scheme for suffixes, hyphens and diacritical marks. Mineralogical Record 39(2), 131–135 (2008)

[11] Hey, N.H.: An Index of Mineral Species and Varieties Arranged Chemically. British Museum, London, UK, 728p. (1962)

[12] Barresi, A.A., Orlandi, P., Pasero, M.: History of ardennite and the new mineral ardennite-(V). European Journal of Mineralogy 19(4), 581–587 (2007). https://doi.org/10.1127/0935-1221/2007/0019-1745

[13] IMA-CNMNC: The official IMA-CNMNC list of mineral names. Available at: http://cnmnc.units.it/imalist.htm (2023). Accessed 03 July 2023

[14] Dana, J.D.: A system of mineralogy, 5th Ed. John Wiley & Son, New York, USA, 827p. (1868)

[15] Strunz, H., Nickel, E.H.: Strunz mineralogical tables, 9th Ed. E. Schweizerbart'sche Verlagsbuchhandlung, Berlin and Stuttgart, 870p. (2001)

[16] Clark, A.M.: Hey's mineral index: mineral species, varieties and synonyms, 3rd Ed. Chapman & Hall, London, 852p. (1993)

[17] Lafuente, B., Downs, R. T., Yang, H., & Stone, N.: The power of databases: The RRUFF project. Highlights in mineralogical crystallography, pp.1–30. (2015)

[18] Peters, S.E., McClennen, M.: The paleobiology database application programming interface. Paleobiology 42(1), 1–7 (2016). https://doi.org/10.1017/pab.2015.39

[19] Ralph, J.: Introducing minID: the new central mineral registration system. Rocks & Minerals 89(4), 364–369 (2014). https://doi.org/10.1080/00357529.2014.907660

[20] Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. and Vieglais, D.: Darwin core: an evolving community-developed biodiversity data standard. PloS One 7(1), e29715 (2012). https://doi.org/10.1371/journal.pone.0029715

[21] Klump, J., Huber, R.: 20 Years of persistent identifiers –which systems are here to stay? Data Science Journal 16(9), 1–7 (2017). https://doi.org/10.5334/dsj-2017-009

[22] NHM (Natural History Museum): BM.1937,51. Available at: https://data.nhm.ac.uk/object/0946ed2e-3bad-4f71-8dc7-23c6b63efbc3/1688515200000 (2023). Accessed 05 July 2023

[23] Golden, J.J., Downs, R.T., Hazen, R.M., Pires, A.J., Ralph, J.: Mineral evolution database: data-driven age assignment, how does a mineral get an age? In: GSA Annual Meeting 2019, Phoenix, Arizona, USA (2019). https://doi.org/10.1130/abs/2019AM-334056

[24] Hystad, G., Downs, R.T., Hazen, R.M.: Mineral species frequency distribution conforms to a large number of rare events model: prediction of Earth's missing minerals. Mathematical Geosciences 47(6), 647–661 (2015). https://doi.org/10.1007/s11004-015-9600-3

[25] Bazzanella, B., Bortoli, S., Bouquet, P.: Can persistent identifiers be cool? International Journal of Digital Curation 8(1), 14–28 (2013) https://doi.org/10.2218/ijdc.v8i1.246

[26] Miyawaki, R., Hatert, F., Pasero, M., Mills, S. J.: IMA Commission on New Minerals, Nomenclature and Classification (CNMNC) - Newsletter 69, European Journal of Mineralogy 34, 463–468 (2022). https://doi.org/10.5194/ejm-34-463-2022

[27] Leach, P., Mealling, M., Salz, R.: A Universally Unique IDentifier (UUID) URN Namespace (No. RFC4122). The Internet Society (2005). https://datatracker.ietf.org/doc/html/rfc4122

[28] GBIF: The global biodiversity information facility. Available at: https://www.gbif.org (2023). Accessed 05 July 2023

[29] GeoNames: GeoNames: open geographical database. Available at: https://www.geonames.org (2023). Accessed 05 July 2023

AUTHOR BIOGRAPHY

Jolyon Ralph is the founder of mindat.org. He started mindat as a personal mineral information database in 1993, which was later launched as a web site in October 2000. He also runs the gemdat.org website.

Pavel Martynov is a data manager for mindat.org.

Xiaogang Ma is an associated professor of computer science at the University of Idaho. His research focuses on deploying data science in the Semantic Web to support cross-disciplinary collaboration and scientific discovery.
ORCID: 0000-0002-9110-7369

David Von Bargen is a data manager for mindat.org.

Wenjia Li is a postdoctoral researcher at University of Idaho. Her research interests include natural language processing, text mining, semantic web, and their applications.
ORCID: 0000-0002-5728-1566

Jingyi Huang is a postdoctoral researcher at University of Idaho. Her research interests include geochemistry of deep Earth and geoinformatics applications.
ORCID: 0000-0002-5274-1068

Joshua Golden is a scientist at Carnegie Institution for Science, working on database analyses to better understand mineral systems.

Lucia Profeta is a project manager for EarthChem and data curator at the Astromaterials Data System, Columbia University.
ORCID: 0000-0001-9642-7620

Anirudh Prabhu is a research scientist at Carnegie Institution for Science. His research involves developing algorithms, visualizations, and methods in the fields of artificial intelligence, machine learning, and informatics and their applications in various fields.
ORCID: 0000-0002-9921-6084

Shaunna M. Morrison is a research scientist at Carnegie Institution for Science. She works on mineralogy and planetary science, with expertise in crystallography, crystal chemistry, and the application of data-driven techniques.
ORCID: 0000-0002-1712-8057

Xiang Que is a postdoctoral researcher at University of Idaho. His research interests include algorithms for spatio-temporal analysis, open data, and data science applications.
ORCID: 0000-0002-5687-8627

Jiyin Zhang is a PhD student at University of Idaho. His research interests include knowledge graphs, large language models, and data analytics.
ORCID: 0000-0001-7914-8953